Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw301
Advance Access Publication Date: 13 May 2016
Applications Note

OXFORD

Structural bioinformatics

PRESS: PRotEin S-Sulfenylation server

Marianna Sakka[1,†], Grigorios Tzortzis[2,†], Michalis D. Mantzaris[3,†], Nick Bekas[1], Tahsin F. Kellici[1], Aristidis Likas[2], Dimitrios Galaris[3], Ioannis P. Gerothanassis[1] and Andreas G. Tzakos[1,*]

[1]Department of Chemistry, [2]Department of Computer Science and Engineering and [3]Laboratory of Biological Chemistry Faculty of Medicine, University of Ioannina, Ioannina, Greece

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.
Associate Editor: Alfonso Valencia

Received on October 4, 2015; revised on April 14, 2016; accepted on May 6, 2016

Abstract

Motivation: Transient S-sulfenylation of cysteine thiols mediated by reactive oxygen species plays a critical role in pathology, physiology and cell signaling. Therefore, discovery of new S-sulfenylated sites in proteins is of great importance towards understanding how protein function is regulated upon redox conditions.
Results: We developed PRESS (PRotEin S-Sulfenylation) web server, a server which can effectively predict the cysteine thiols of a protein that could undergo S-sulfenylation under redox conditions. We envisage that this server will boost and facilitate the discovery of new and currently unknown functions of proteins triggered upon redox conditions, signal regulation and transduction, thus uncovering the role of S-sulfenylation in human health and disease.
Availability and implementation: The PRESS web server is freely available at http://press-sulfenylation.cse.uoi.gr/
Contacts: agtzakos@gmail.com or gtzortzi@cs.uoi.gr
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Protein S-sulfenylation is a transient post-translational modification through which cysteine (CYS) thiols of proteins are reversibly oxidized to cysteine sulfenic acids (CSO), an event that modulates their function and activity. It has been shown that reactive oxygen species (ROS) mediated S-sulfenylation of cysteine thiols is a crucial intermediate state in a variety of cysteine modifications and acts as a regulatory mechanism in pathology, physiology and cell signaling (Paulsen and Carroll, 2013; Paulsen *et al.*, 2012; Yang *et al.*, 2014). The discovery, therefore, of new S-sulfenylated sites in proteins is of great importance towards understanding how protein function is regulated upon redox conditions.

Experimental discovery and identification of new CSO in proteins is an expensive, intensive and time-consuming process. Although numerous servers exist for the prediction of cysteine modifications such as S-Nitrosylation, S-Glutathionylation, etc., structure-based prediction of transient S-sulfenylation is still elusive. Herein, we developed PRotEin S-Sulfenylation (PRESS), a server that predicts, rapidly and efficiently, CSO sites in proteins. This is a structure-based methodology that relies on the observation of a high conservation in the spatial environment of CSO sites. (Supplementary Section 1, Tables S1, S2 and Figs S1–S5). To build PRESS, we took advantage of the under-utilized information of high resolution X-ray structures bearing CSO sites.

2 Server description

The PRESS server is a useful tool for exploring redox regulated pathways. It provides a user friendly interface, where a user can upload Protein Data Bank (PDB)-formatted protein structures of interest, derived either from experimental data or homology modeling, and have access to: (i) Information regarding the surrounding environment of cysteines from three different perspectives (Supplementary Section 2). First, the neighboring amino acids of each cysteine in the

protein sequence (lying within a window size of $W = 11$, centered at the cysteine) are extracted. Second, the spatial amino acid neighbors surrounding each cysteine in 3D space, lying within a user defined radius, are extracted. The distance between a cysteine and a neighboring amino acid corresponds to the shortest distance between any pair of their atoms, based on their atom coordinates. Third, water molecules and/or metal ions surrounding each cysteine in 3D space (lying within a user defined radius) are also extracted. Moreover, complementary statistical information, concerning all the available cysteines in the uploaded protein structures, is provided in each of the above three cases in the form of sequence logos or bar charts (Supplementary Figs S6–S11). (ii) Automatic prediction of cysteines in proteins that are prone to undergo transient S-sulfenylation via an SVM-based predictive algorithm (Supplementary Section 3). Usage instructions for PRESS can be found in Supplementary Section 5.

2.1 SVM training

To train the predictive algorithm, we obtained a protein dataset consisting of 204 PDB files from the PDB repository (Berman *et al.*, 2000). The advanced search utility was used to search for entries containing the word 'CSO'. To avoid structure redundancy, results were filtered to retrieve single representatives in cases of 90% sequence identity. Residues annotated as 'missing' were removed. To avoid overestimating the potential of our method, duplicates were removed (see the description of the training dataset in Supplementary Section 1). A total of 243 CSO and 777 CYS residues were retained and used for SVM training (see Supplementary Section 3). The developed, SVM-based, predictive algorithm operates by utilizing a combination of four cysteine-related structural features: (i) neighboring amino acids in the protein sequence, (ii) neighboring amino acids in the 3D space (lying within a radius of 5 Angstroms), (iii) the absolute solvent accessibility (ASA) and (iv) the secondary structure. This selection emerged after performing a series of different combinations, as described in the supplementary information, Section 3 – Table S3, and was found to be optimum for capturing the information related to the surrounding environment of CSO and CYS residues, contributing in their effective discrimination. Using 10-fold cross-validation, PRESS can predict CSO sites with a sensitivity of 79.9%, specificity of 73.6% and an overall balanced accuracy of 76.8%. To further test the performance of the predictive algorithm, CSO bearing structures were filtered, as above, to retrieve single representatives in structures with sequence identity as low as 30% (50 and 70% are also illustrated in the Supplementary Section 3 – Table S4). Interestingly, even in this case, PRESS can achieve sensitivity of 75.3%, specificity of 72.4% and an overall balanced accuracy of 73.8% to predict cysteines sensitive to S-sulfenylation. This provides solid evidence that the four chosen structural features incorporated in the SVM successfully capture the information required for efficient discrimination between CSO and CYS residues and that our method can achieve a high prediction accuracy.

2.2 SVM validation

The SVM model obtained from training was employed to make predictions for different experimentally verified datasets containing S-sulfenylated proteins (Supplementary Section 1). Endogenous and oxidant stimulated sets (i.e. H_2O_2, EGF) verified by dimedone-based labeling (Gould *et al.*, 2015; Yang *et al.*, 2014) were included for testing. By utilizing experimental and homology-modeled PDB files, PRESS could selectively predict CSO sites in: (i) H_2O_2 dataset (sensitivity 73.50%, specificity 66.60%, accuracy 70.10%), (ii) EGF dataset (sensitivity 63.70%, specificity 67.10%, accuracy 65.50%), (iii) endogenous dataset (sensitivity 60.40%, specificity 68.90%, accuracy 64.70%) and (iv) X-ray dataset (sensitivity 83.90%, specificity 82.90%, accuracy 83.40%) (Supplementary Sections S1, S3 and Table S5). Interestingly, oxidant stimulated sets returned higher prediction scores compared to the endogenous sets. Differences in sensitivity could be explained by the fact that oxidant sets represent quantitative data and thus are 'more accurate' for validating the predictive algorithm than the endogenous. Endogenous sets represent qualitative data and thus are more likely to include false-positives due to non-specific CSO labeling. On the other hand, lower scores in EGF data could be attributed 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.

3 Discussion

The crystallographic PDB repository contains a large number of high resolution X-ray structures of proteins that bear S-sulfenylation sites. This information that could be used to decode the principles governing cysteine S-sulfenylation, has remained largely unexplored. Using a database of CSO-containing proteins, we identified several features contributing to the potential of a cysteine to undergo S-sulfenylation and devised an SVM-based algorithm, which is available via the PRESS server, to predict CSO sites. This is of judicial importance given the intrinsic character of this post translational modification and its implication in cell signaling, pathology and physiology. The PRESS server has an overall balanced accuracy of 76.8% on the, currently available, X-ray protein structural data to predict S-sulfenylation sites in proteins and, thus, it is of great importance towards understanding how protein function is regulated upon redox conditions. As an indicative example of the effectiveness of PRESS, we provide the prediction result on the EGF receptor, which is experimentally verified to be S-sulfenylated on CYS 797 (Paulsen *et al.*, 2012) (Supplementary Section 5 and Fig. S12-A). Another example is the protein monoacylglycerol lipase, containing four cysteines from which CYS 201 and CYS 208 were, very recently, experimentally verified to be S-sulfenylated (Dotsey *et al.*, 2015) (Supplementary Section 5 and Fig. S12-B). Applications of this server could be oriented towards the discovery of new and currently unknown functions of proteins triggered upon redox conditions, signal regulation and transduction as also their implication in disease.

Funding

This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALES. Investing in knowledge society through the European Social Fund. (Acronym: FeRedox; ID: 1356)

Conflict of Interest: none declared.

References

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Dotsey,E.Y. *et al.* (2015) Peroxide-dependent MGL sulfenylation regulates 2-AG-mediated endocannabinoid signaling in brain neurons. *Chem. Biol.*, **22**, 619–628.

Gould,N.S. *et al.* (2015) Site-specific proteomic mapping identifies selectively modified regulatory cysteine residues in functionally distinct protein networks. *Chem. Biol.*, **22**, 965–975.

Paulsen,C.E. and Carroll,K.S. (2013) Cysteine-mediated redox signaling: chemistry, biology, and tools for discovery. *Chem. Rev.*, **113**, 4633–4679.

Paulsen,C.E. *et al.* (2012) Peroxide-dependent sulfenylation of the EGFR catalytic site enhances kinase activity. *Nat. Chem. Biol.*, **8**, 57–64.

Yang,J. *et al.* (2014) Site-specific mapping and quantification of protein S-sulphenylation in cells. *Nat. Commun.*, **5**, 4776.