BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 16 2014, pages 2386–2388
doi:10.1093/bioinformatics/btu301

Databases and ontologies Advance Access publication April 29, 2014

dbGSH: a database of *S*-glutathionylation

Yi-Ju Chen[1], Cheng-Tsung Lu[2], Tzong-Yi Lee[2,*] and Yu-Ju Chen[1,*]

[1]Institute of Chemistry, Academia Sinica, Taipei 115, Taiwan and [2]Department of Computer Science and Engineering, Yuan Ze University, Taoyuan 320, Taiwan

Associate Editor: Jonathan Wren

ABSTRACT

Summary: *S*-glutathionylation, the reversible protein posttranslational modification (PTM) that generates a mixed disulfide bond between glutathione and cysteine residue, critically regulates protein activity, stability and redox regulation. Due to its importance in regulating oxidative/nitrosative stress and balance in cellular response, a number of methods have been rapidly developed to study *S*-glutathionylation, thus expanding the dataset of experimentally determined glutathionylation sites. However, there is currently no database dedicated to the integration of all experimentally verified *S*-glutathionylation sites along with their characteristics or structural or functional information. Thus, the dbGSH database has been created to integrate all available datasets and to provide the relevant structural analysis. As of January 31, 2014, dbGSH has manually collected >2200 experimentally verified *S*-glutathionylated peptides from 169 research articles using a text-mining approach. To solve the problem of heterogeneity of the data collected from different sources, the sequence identity of the reported *S*-glutathionylated peptides is mapped to UniProtKB protein entries. To delineate the structural correlations and consensus motifs of these *S*-glutathionylation sites, the dbGSH database also provides structural and functional analyses, including the motifs of substrate sites, solvent accessibility, protein secondary and tertiary structures, protein domains and gene ontology.

Availability and implementation: dbGSH is now freely accessible at http://csb.cse.yzu.edu.tw/dbGSH/. The database content is regularly updated with new data collected by the continuous survey of research articles.

Contact: francis@saturn.yzu.edu.tw or yujuchen@gate.sinica.edu.tw

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 27, 2013; revised on April 23, 2014; accepted on April 24, 2014

1 INTRODUCTION

S-glutathionylation is a redox-dependent posttranslational modification (PTM) that involves the covalent attachment of glutathione (GSH) to the thiol group of cysteine residues, and regulates based on the physiological GSH level and the GSH/GSSG ratio through spontaneous or enzymatic reaction (Dalle-Donne *et al*., 2007, 2009;Gallogly and Mieyal,2007; Ghezzi, 2013; Pastore and Piemonte, 2012). In addition to regulating redox signaling, *S*-glutathionylation also serves to modulate cancer migration, cell death and survival, energy metabolism and glycolysis, as well as protein folding and degradation in several species, ranging from bacteria to human (Pastore and Piemonte, 2012). The various targets of *S*-glutathionylation also participate in pathogenesis of many diseases, such as neurodegenerative diseases, metabolic disorders and cancers (Dalle-Donne *et al*., 2008; Mieyal and Chock, 2012). Due to the importance of *S*-glutathionylation in clinical biological processes, numerous methods have been developed, directed toward mass spectrometry-based *S*-glutathionomics using various biological systems to investigate and identify *S*-glutathionylated targets and sites (Chiang *et al*., 2012; Lind *et al*., 2002; Newman *et al*., 2007). As the number of experimentally identified *S*-glutathionylated peptides grows, a structured database is desirable for the further investigation of the biological functions of *S*-glutathionylated proteins and the substrate specificities of *S*-glutathionylation sites. Although a number of databases have been developed for PTMs, such as dbPTM (Lee *et al*., 2006; Lu *et al*., 2013; Su *et al*., 2014), PhosphoSitePlus (Hornbeck *et al*., 2012), RegPhos (Lee *et al*., 2011a), dbOGAP (Wang *et al*., 2011) and dbSNO (Lee *et al*., 2012), limited resources have been dedicated to *S*-glutathionylated proteins and their corresponding substrate sites (Sun *et al*., 2012, 2013). Thus, a new database (dbGSH) containing the experimentally verified *S*-glutathionylated peptides from research articles is proposed to facilitate the functional analysis of *S*-glutathionylated proteins and to assist with the structural investigation of substrate sites.

2 METHODS

Supplementary Figure S1 shows the exhaustive system flow of the construction of dbGSH, containing the experimentally verified data for *S*-glutathionylated proteins and *S*-glutathionylation sites. The database entries are extracted manually from research articles through a literature survey. All fields in the PubMed database are first searched using the keywords 'glutathio(ny)lation' or 'glutathio(ny)lated', followed by the downloading of full-text versions of these research articles. A text-mining system is developed to survey the full-text literature that potentially describes the site-specific identification of *S*-glutathionylation sites for the various proteomic identification experiments (Chiang *et al*., 2012; Hamnell-Pamment *et al*., 2005). Approximately 1073 original and review articles associated with protein *S*-glutathionylation are retrieved from PubMed. Next, the full-length articles are manually reviewed to ensure precise extraction of the *S*-glutathionylated peptides and the modified cysteines. To determine the exact locations of *S*-glutathionylated cysteines within a full-length protein sequence, the experimentally verified *S*-glutathionylated peptides are then mapped to UniProtKB (Consortium, 2013) protein entries based on the database identifier (ID) or sequence similarity. The *S*-glutathionylated peptides that do not align exactly to a protein sequence are discarded. Finally, each mapped *S*-glutathionylation site is attributed to the corresponding

*To whom correspondence should be addressed.

literature (PubMed ID) reference. In addition, a small number of protein *S*-glutathionylation sites in UniProtKB are integrated into dbGSH with the corresponding literature references. As of January 31, 2014, a total of 169 *S*-glutathionylation-associated articles covering 16 organisms have been retrieved from PubMed. In the dbGSH database, there are 2257 *S*-glutathionylated cysteines from 1314 *S*-glutathionylated proteins. Supplementary Table S1 shows the statistics for the *S*-glutathionylation data for each organism.

For a given protein, the biological function can be obtained from the annotation in UniProtKB. To provide more effective information about protein functional and structural annotations relevant to *S*-glutathionylated cysteines, a variety of biological databases, such as InterPro (Hunter *et al*., 2011), Gene Ontology (GO), Protein Data Bank (PDB) (Rose *et al*., 2011) and KEGG Pathway (Kanehisa *et al*., 2012), are integrated. The information regarding the molecular function, cellular components and biological process for *S*-glutathionylated proteins can be accessed by a cross-linking to the corresponding entry from GO via a UniProtKB accession number. In addition, the InterPro BioMart, which is a web service that provides the ability to build custom queries on the InterPro database, is used to extract the information about protein domains and functional sites for the *S*-glutathionylated proteins in dbGSH. Additionally, the molecular interaction and signaling network of *S*-glutathionylated proteins can be accessed by a cross-link that refers to the corresponding entries from KEGG pathway.

To study the structural characteristics of *S*-glutathionylation sites within the protein tertiary structure, all the experimentally identified *S*-glutathionylation sites are mapped to the corresponding positions in the protein entries in PDB. The solvent accessibility and secondary structure around the substrate sites are also investigated. However, due to the limited information on protein structures in PDB, only 1% of the GSH sites have corresponding tertiary structures. With reference to a previous study investigating the structural characteristics of PTMs (Chang *et al*., 2009; Huang *et al*., 2005; Lee *et al*., 2011b; Shien *et al*., 2009; Su and Lee, 2013; Wong *et al*., 2007) in proteins without known tertiary structures, two effective tools, RVP-net (Ahmad *et al*., 2003) and PSIPRED (McGuffin *et al*., 2000), are used to predict the solvent accessibility and secondary structure, respectively.

3 UTILITY

To facilitate the use of the dbGSH resource, a web interface has been developed for users to browse and efficiently search for their *S*-glutathionylation proteins of interest. Supplementary Figure S2 shows the content of a typical dbGSH query, including basic information, graphical visualization of *S*-glutathionylation sites, table of *S*-glutathionylation sites with the associated literature and visualization of tertiary structures by the Jmol program. According to the GO annotations, the distributions of the biological processes, molecular functions and cellular components of *S*-glutathionylated proteins are presented in Supplementary Table S2. The investigation of protein domains shows that the most abundant domain in *S*-glutathionylated proteins is the RNA recognition motif (Supplementary Table S3), and ~70% of the reported *S*-glutathionylation sites are located within functional domains (Supplementary Table S4). Additionally, Supplementary Table S5 presents the distribution of KEGG pathways for *S*-glutathionylated proteins.

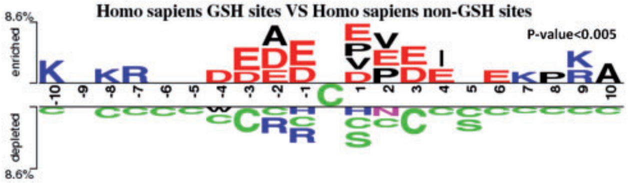

Fig. 1. The compositional biases of amino acids around *S*-glutathionylation sites compared with the non-*S*-glutathionylation sites. The amino acids that are significantly enriched or depleted ($P < 0.005$) around *S*-glutathionylation sites are presented

Figure 1shows that the *S*-glutathionylation sites (upper panel) have a higher preference for containing negatively charged residues (D and E) than non-glutathionylation sites (lower panel) do. To investigate the substrate specificity of the *S*-glutathionylation sites, MDDLogo (Lee *et al*., 2011c), a maximal dependence decomposition (MDD) analysis that clusters the aligned signal sequences into subgroups containing statistically significant motifs, was performed to identify the substrate motifs for the *S*-glutathionylation data in humans, with a sufficient dataset. Supplementary Table S6 shows that a total of nine substrate motifs were identified from 1735 *S*-glutathionylation sites with a 21-mer window length: six groups contain a conserved motif of negatively charged amino acids (D and E) and three groups contain a conserved motif of positively charged amino acids (K, R and H) surrounding the *S*-glutathionylation sites. The substrate sites of *S*-glutathionylation might favor the cysteines flanked by acidic and basic amino acids, suggesting that the acid/basic environment in a close three-dimensional proximity may contribute to the specificity of reactive cysteine toward *S*-glutathionylation.

4 CONCLUSIONS

The dbGSH database is the first public resource to allow efficient access to experimentally validated *S*-glutathionylation sites and comprehensive bioinformatics analyses, including functional annotations, structural characteristics and substrate motifs for *S*-glutathionylation sites. The database not only provides a benchmark dataset for the development of computational prediction tools and their comparison but also supplies the potential glutathionylated targets for verification and other biological applications.

Funding: This work was supported by the National Science Council in Taiwan (NSC 101-2628-E-155-002-MY2 to T.Y.L. and NSC-100-2922-I-016-002 to Y.J.C.).

Conflict of Interest: none declared.

REFERENCES

Ahmad,S. *et al*. (2003) RVP-net: online prediction of real valued accessible surface area of proteins from single sequences. *Bioinformatics*, **19**, 1849–1851.

Chang,W.C. *et al*. (2009) Incorporating support vector machine for identifying protein tyrosine sulfation sites. *J. Comput. Chem*., **30**, 2526–2537.

Chiang,B.Y. *et al*. (2012) *In vivo* tagging and characterization of S-glutathionylated proteins by a chemoenzymatic method. *Angew. Chem. Int. Ed. Engl*., **51**, 5871–5875.

Consortium,U. (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res*., **41**, D43–D47.

Dalle-Donne,I. *et al*. (2007) S-glutathionylation in protein redox regulation. *Free Radic. Biol. Med*., **43**, 883–898.

Dalle-Donne,I. *et al.* (2008) Molecular mechanisms and potential clinical significance of S-glutathionylation. *Antioxid. Redox Signal.*, **10**, 445–473.

Dalle-Donne,I. *et al.* (2009) Protein S-glutathionylation: a regulatory device from bacteria to humans. *Trends Biochem. Sci.*, **34**, 85–96.

Gallogly,M.M. and Mieyal,J.J. (2007) Mechanisms of reversible protein glutathionylation in redox signaling and oxidative stress. *Curr. Opin. Pharmacol.*, **7**, 381–391.

Ghezzi,P. (2013) Protein glutathionylation in health and disease. *Biochim. Biophys. Acta*, **1830**, 3165–3172.

Hamnell-Pamment,Y. *et al.* (2005) Determination of site-specificity of S-glutathionylated cellular proteins. *Biochem. Biophys. Res. Commun.*, **332**, 362–369.

Hornbeck,P.V. *et al.* (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.*, **40**, D261–D270.

Huang,H.D. *et al.* (2005) KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Res.*, **33**, W226–W229.

Hunter,S. *et al.* (2011) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.

Kanehisa,M. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.

Lee,T.Y. *et al.* (2006) dbPTM: an information repository of protein post-translational modification. *Nucleic Acids Res.*, **34**, D622–D627.

Lee,T.Y. *et al.* (2011a) RegPhos: a system to explore the protein kinase-substrate phosphorylation network in humans. *Nucleic Acids Res.*, **39**, D777–D787.

Lee,T.Y. *et al.* (2011b) SNOSite: exploiting maximal dependence decomposition to identify cysteine S-nitrosylation with substrate site specificity. *PLoS One*, **6**, e21849.

Lee,T.Y. *et al.* (2011c) Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences. *Bioinformatics*, **27**, 1780–1787.

Lee,T.Y. *et al.* (2012) dbSNO: a database of cysteine S-nitrosylation. *Bioinformatics*, **28**, 2293–2295.

Lind,C. *et al.* (2002) Identification of S-glutathionylated cellular proteins during oxidative stress and constitutive metabolism by affinity purification and proteomic analysis. *Arch. Biochem. Biophys.*, **406**, 229–240.

Lu,C.T. *et al.* (2013) DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. *Nucleic Acids Res.*, **41**, D295–D305.

McGuffin,L.J. *et al.* (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.

Mieyal,J.J. and Chock,P.B. (2012) Posttranslational modification of cysteine in redox signaling and oxidative stress: Focus on S-glutathionylation. *Antioxid. Redox Signal.*, **16**, 471–475.

Newman,S.F. *et al.* (2007) An increase in S-glutathionylated proteins in the Alzheimer's disease inferior parietal lobule, a proteomics approach. *J. Neurosci. Res.*, **85**, 1506–1514.

Pastore,A. and Piemonte,F. (2012) S-Glutathionylation signaling in cell biology: progress and prospects. *Eur. J. Pharm. Sci.*, **46**, 279–292.

Rose,P.W. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res*, **39**, D392–D401.

Shien,D.M. *et al.* (2009) Incorporating structural characteristics for identification of protein methylation sites. *J. Comput. Chem.*, **30**, 1532–1543.

Su,M.G. and Lee,T.Y. (2013) Incorporating substrate sequence motifs and spatial amino acid composition to identify kinase-specific phosphorylation sites on protein three-dimensional structures. *BMC Bioinformatics*, **14** (**Suppl. 16**), S2.

Sun,C. *et al.* (2013) Prediction of S-Glutathionylation sites based on protein sequences. *PLoS One*, **8**, e55512.

Sun,M.A. *et al.* (2012) RedoxDB—a curated database for experimentally verified protein oxidative modification. *Bioinformatics*, **28**, 2551–2552.

Su,M.G. *et al.* (2014) topPTM: a new module of dbPTM for identifying functional post-translational modifications in transmembrane proteins. *Nucleic Acids Res.*, **42**, D537–D545.

Wang,J. *et al.* (2011) dbOGAP—an integrated bioinformatics resource for protein O-GlcNAcylation. *BMC Bioinformatics*, **12**, 91.

Wong,Y.H. *et al.* (2007) KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res.*, **35**, W588–W594.