D332–D337 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 29 October 2010*
doi:10.1093/nar/gkq1060

PREX: PeroxiRedoxin classification indEX, a database of subfamily assignments across the diverse peroxiredoxin family

Laura Soito[1], Chris Williamson[1], Stacy T. Knutson[2,3], Jacquelyn S. Fetrow[2,3], Leslie B. Poole[1] and Kimberly J. Nelson[1,*]

[1]Department of Biochemistry, Wake Forest University Health Sciences, Medical Center Blvd., Winston-Salem, NC 27157, [2]Department of Physics and [3]Department of Computer Science, Wake Forest University, Winston-Salem, NC 27109, USA

Received August 18, 2010; Revised October 11, 2010; Accepted October 13, 2010

ABSTRACT

PREX (http://www.csb.wfu.edu/prex/) is a database of currently 3516 peroxiredoxin (Prx or PRDX) protein sequences unambiguously classified into one of six distinct subfamilies. Peroxiredoxins are a diverse and ubiquitous family of highly expressed, cysteine-dependent peroxidases that are important for antioxidant defense and for the regulation of cell signaling pathways in eukaryotes. Subfamily members were identified using the Deacon Active Site Profiler (DASP) bioinformatics tool to focus in on functionally relevant sequence fragments surrounding key residues required for protein activity. Searches of this database can be conducted by protein annotation, accession number, PDB ID, organism name or protein sequence. Output includes the subfamily to which each classified Prx belongs, accession and GI numbers, genus and species and the functional site signature used for classification. The query sequence is also presented aligned with a select group of Prxs for manual evaluation and interpretation by the user. A synopsis of the characteristics of members of each subfamily is also provided along with pertinent references.

INTRODUCTION

Peroxiredoxin (Prx or PRDX) proteins (EC 1.11.1.15) are a ubiquitous family of highly expressed, thioredoxin-scaffold enzymes that exhibit cysteine-dependent peroxidase activity with hydrogen peroxide and larger hydroperoxide substrates (1–3). These antioxidant enzymes are important in many organisms, including plants and bacteria as well as animals, for protecting against oxidative damage. Moreover, eukaryotic Prxs are implicated in intracellular signaling and the regulation of such processes as cell proliferation, differentiation and cell death (4). Prxs have also been shown to be induced by oxidative stress, to be aberrantly expressed in cancer and to impact the radiation sensitivity of cells (5). The importance of these proteins as potential prognostic or therapeutic targets is of widespread interest across many diseases that involve reactive oxygen species, including cancer and cardiovascular disease (6).

With the rapid increase over the last decade in the functional and structural information available about Prxs, distinctions between various Prx subfamilies have become apparent (e.g. oligomerization properties, location of mechanistically important cysteinyl residues). Prx proteins are widely distributed and members of multiple subfamilies are present in most organisms, including three subfamilies in humans. Prxs have frequently been classified based on the number of Cys residues involved in the catalytic cycle; however, the distinction between 2-Cys and 1-Cys function is not particularly useful as a global classifier because representatives of each type seem to exist within all the subfamilies (7). Bioinformatics efforts by others have helped clarify some of the features distinguishing Prxs subfamilies (3,8) and are particularly powerful when combined with structural analyses of representative Prxs (9,10). While structure-based subfamily classifications are increasingly recognized, detailed annotation to the level of Prx subfamily across the entire GenBank database remains scant and frequently confusing.

Only one other database is available where information for peroxidases has been collected and organized. PeroxiBase (http://peroxibase.toulouse.inra.fr/) (11) is a collection of sequences classified across all types of heme and non-heme peroxidases, which comprise the enzyme class EC 1.11.1.x. In contrast, PREX is focused on the detailed annotation of the Prxs, one type of non-heme

*To whom correspondence should be addressed. Tel: +1 336 716 6370; Fax: +1 336 777 3242; Email: kinelson@wfubmc.edu

peroxidase found in PeroxiBase. PeroxiBase organizes the Prxs into subfamilies based on broad biological function and overall sequence comparison (typical 1-Cys, typical 2-Cys, atypical 2-Cys and two different thioredoxin peroxidase subfamilies) (11). The information in PREX classifies the Prxs based on structural and sequence information at the reactive cysteine active site, resulting in six mechanistically based subfamilies. As of October 2010, PeroxiBase contains 826 proteins annotated as Prx, while PREX contains 3516 proteins that are each assigned to a single subfamily. Thus, the PREX information is complementary to the broader biologic functional organization provided in PeroxiBase.

To provide global Prx subfamily assignments for the Prx field, we have recently identified 3516 members of the Prx family (12) from the January 2008 GenBank database. Based upon structural analysis by Hall *et al.* (7) and by our own bioinformatic analysis (12) utilizing the functional site profiling method (also referred to as active site profiling) (13) and the functional site profile search tool implemented in Deacon Active Site Profiler (DASP) (14) to analyze sequence conservation in the structural vicinity of the catalytic cysteine, all of these proteins have been unambiguously classified into six functionally relevant subfamilies: AhpC/Prx1, Prx6, BCP/PrxQ, Tpx, Prx5 and AhpE (Table 1). We have incorporated our newly developed 'index' of proteins into the present searchable database, called PREX (http://www.csb.wfu.edu/prex/). PREX is designed to help fill the need for accurate classification of the Prx family members and to be useful for researchers familiar with Prx function, as well as for those new to the subject.

DATABASE CONTENT

Prx classifications were made using the DASP profile search tool (14) (publically available at http://dasp.deac.wfu.edu/) that uses fine structure mapping and profiling to identify motifs of functional importance (13,15). DASP requires the selection of key residues that define a functional site within a protein family and then identifies nearby sequence fragments (Figure 1). As described in more detail in a separate research paper (12), key residues used to define the Prx active site included the three residues in the PXXX(T/S)XXC motif found in all Prxs (1,3,16) as well as a conserved Trp /Phe residue located ~6 Å from the catalytic cysteine (Trp81 in *Salmonella typhimurium* AhpC). All residues which contained an atom located within 10 Å of the center of geometry of at least one of these key residues were extracted and the sequence fragments containing these residues were placed in order from N- to C-terminus to form the 'functional site signature' from all 29 non-redundant Prxs of known structure in the RCSB PDB database as of January 2008 (Figure 1).

Each Prx signature was assigned to a single subfamily (AhpC/Prx1, Prx6, BCP/PrxQ, Prx5, Tpx or AhpE) based on previously published structural characterizations (10) which agreed well with hierarchical clustering of the aligned functional site signatures. Signatures for multiple representatives in each subfamily were combined to generate a subfamily-specific 'functional site profile' that was used to identify subfamily members from GenBank(nr), according to the method described elsewhere (12,14). Each returned sequence is associated

Table 1. Summary of the Prx subfamilies present in PREX

Subfamily	Number of database members	Canonical subfamily members	Phylogenetic distribution	Typical location of C_R when present[a]
AhpC/Prx1[b]	1059	*Salmonella typhimurium* AhpC, *Homo sapiens* Prx1 through Prx4	Archea, Bacteria, Plants, Unicellular and Multicellular Eukaryotes	C-terminus (>96%)[c]
BCP/PrxQ	1115	*Escherichia coli* bacterioferritin comigratory protein, plant PrxQ	Bacteria, Plants	Helix α2 (~50%) or α3 (~7%)[d]
Prx5[e]	517	*H. sapiens* Prx5	Bacteria, Eukaryotes	Helix α5 (~17%)[d]
Prx6[f]	493	*H. sapiens* Prx6	Archea, Bacteria, Plants, Unicellular and Multicellular Eukaryotes	No C_R
Tpx[g]	307	*E. coli* Tpx	Bacteria	Helix α3 (>96%)[d]
AhpE	25	*Mycobacterium tuberculosis* AhpE	Bacteria	Unknown[h]

[a]Structural designations as in (10). If no C_R is present, resolving thiol must come from another protein or small molecule.
[b]The AhpC/Prx1 subfamily is also known as the 'typical 2-Cys' Prxs and includes tryparedoxin peroxidases, *Arabidopsis thaliana* 2-Cys Prx, barley Bas1 and *Saccharomyces cerevisiae* TSA1 and TSA2.
[c]The C_R is near the C-terminus of the partner subunit within the homodimer; upon oxidation, intersubunit disulfide forms between the C_P and the C_R of the two chains.
[d]Intrasubunit disulfide formed in oxidized protein (so-called 'atypical' 2-Cys Prxs).
[e]The Prx5 subfamily includes *Populus trichocarpa* PrxD, the plant type II Prxs, mammalian Prx5 and a group of bacterial glutaredoxin-Prx5 fusion proteins.
[f]The Prx6 subfamily (frequently referred to as the '1-Cys' group) also includes the bacterial Prx6 proteins, *A. thaliana* 1-Cys Prx and *S. cerevisiae* mitochondrial Prx1.
[g]The Tpx subfamily includes bacterial proteins (*e.g.* from *Streptococcus pneumoniae* and *Helicobacter pylori)* named thiol peroxidase, p20 and scavengase.
[h]Canonical member contains no C_R, but >50% of sequences include a potential C_R in α2, similar to *E. coli* BCP.

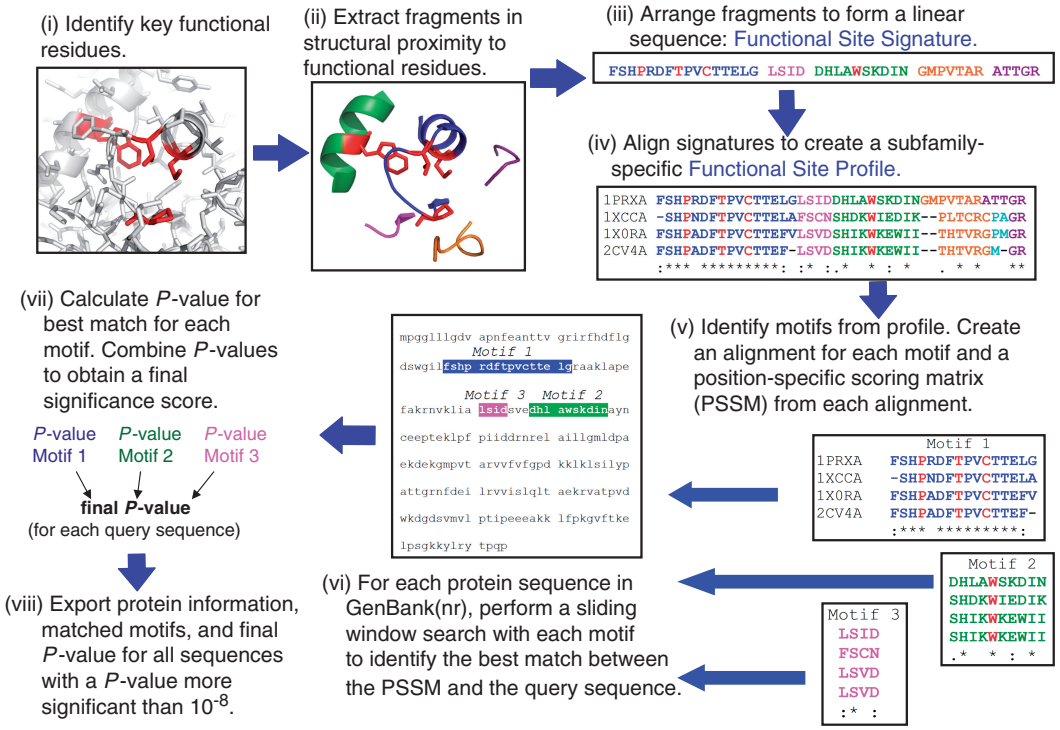

Figure 1. Identification of Prx sequences using the DASP tool. (i) The active site of human Prx6 (PDB identifier 1prx) is shown with the four key residues highlighted in red. (ii) Structural segments located within 10 Å of the center of geometry of the key catalytic residues are identified (each segment shown in a different color) and extracted from the global structure. (iii) The sequence fragments are then combined to form a functional site signature [residue colors correspond to the color of structure segments in (ii); key residues are highlighted in red]. (iv) Functional-site signatures for structurally characterized members of the Prx6 subfamily are aligned using ClustalW (22,24) to create a functional site profile. (v) Motifs are identified within any fragments that contain at least three residues and position specific scoring matrices (PSSM) (25) are created for each motif. (vi) For each sequence in a user-selected sequence database, the PSSM for each motif is used to find and score the segment within a query sequence which best matches a motif. (vii) Each time a motif is matched to a position in the protein sequence, a *P*-value is calculated that represents the probability of finding a match as good as the observed match within a random sequence. The *P*-values for all motifs in a single sequence are then combined using QFAST to obtain the final statistical significance score (final *P*-value) (26). (viii) The protein information (including accession numbers, annotations and species), final *P*-value and sequence fragments matched to each queried motif are exported for all sequences with a final *P*-value more significant than a user-selected *P*-value. See (13–15) for a more detailed description of DASP utilities and architecture.

with a score (*P*-value) based on the probability of finding as good of a random match in a random sequence as the observed match. A *P*-value cutoff of 1×10^{-8} was selected to define 'true' DASP hits for each subfamily, as described elsewhere (12). Due to the limited number of structures available for the BCP/PrxQ and AhpE subfamilies, 'engineered' profiles (including sequence fragments and manually generated 'signatures' from structurally uncharacterized proteins) were created in order to increase the robustness of the profiles. The results for each subfamily were hand curated to remove any sequences that did not contain the PXXX(T/S)XXC Prx motif or that were identified with a more significant *P*-value in another Prx subfamily (62 sequences were removed out 3578 sequences originally identified for all subfamilies). Complete details are described in Nelson *et al.* (12).

DATABASE ARCHITECTURE

This subfamily-specific list of Prxs was used to create PREX, a relational database built on MySQL and accessed from a PHP web-based interface (http://www.csb.wfu.edu/prex/). The basic relationship diagram is provided in Supplementary Figure S1. The PREX database contains 3516 unique signatures representing 8658 GenBank identification numbers classified into only one of six Prx subfamilies. Each signature in the PREX database is associated with its GenBank annotations, organism of origin, the DASP subfamily assignment, and the full sequence of the protein from which it comes. To construct the local sequence database for BLAST searching, the full protein sequence associated with each signature was downloaded from GenBank and incorporated into a local BLAST+2.2.23 database (17,18).

DATABASE INTERFACE ORGANIZATION

PREX has been designed to aid in the characterization of Prx proteins by providing a user friendly tool to identify the subfamily assignment for a Prx protein of interest and to quickly and easily find distinguishing features and typical characteristics for each Prx subfamily. Access and interface to the PREX database is provided by the publically available website, at http://www.csb.wfu.edu/prex/. This website is divided into three sections: (i) Peroxiredoxin Information, (ii) Search Tools and (iii) References.

Peroxiredoxin information

A general introduction to the Prx family is provided on the Home Page along with the list of subfamilies present in this database. Additional information for each subfamily can be accessed from the Prx Subfamily tab and provides structural and functional characteristics of members of each subfamily, PDB identifiers for structurally characterized subfamily members, and some pertinent references. In addition, a representative multiple sequence alignment is provided containing the functional site signatures for 4–5 members of the selected protein subfamily and one representative from each of the other subfamilies.

Search tools

In order to identify the subfamily assignment of a particular Prx protein, searches of this database can be conducted by accession number, PDB ID, protein annotation, organism name or protein sequence (Figure 2). It should be noted that the BLAST search algorithm utilized for the sequence search relies on conservation across the entire protein sequence while subfamily assignments in this database are based exclusively on sequence conservation around the functional site using the DASP signatures. We therefore stress that caution must be exercised with any query sequence that is not an identical match with a pre-classified member of the PREX database. A stringent *E*-value of $1 \times e^{-40}$ has been selected as the default cutoff based upon trial runs with both true Prx proteins and closely related decoy proteins.

Search output

Output for the text searches includes: identifying numbers (including Genbank accession numbers, PDB identifier and Swiss-Prot entry names), any associated annotations, genus and species, the subfamily to which the query sequence belongs and the functional site signature generated by DASP (Figure 2). This information is also accessible from the BLAST output screen for each PREX database protein identified as similar to a query sequence (by selecting the GI number). As described above, additional information for the relevant subfamily can be

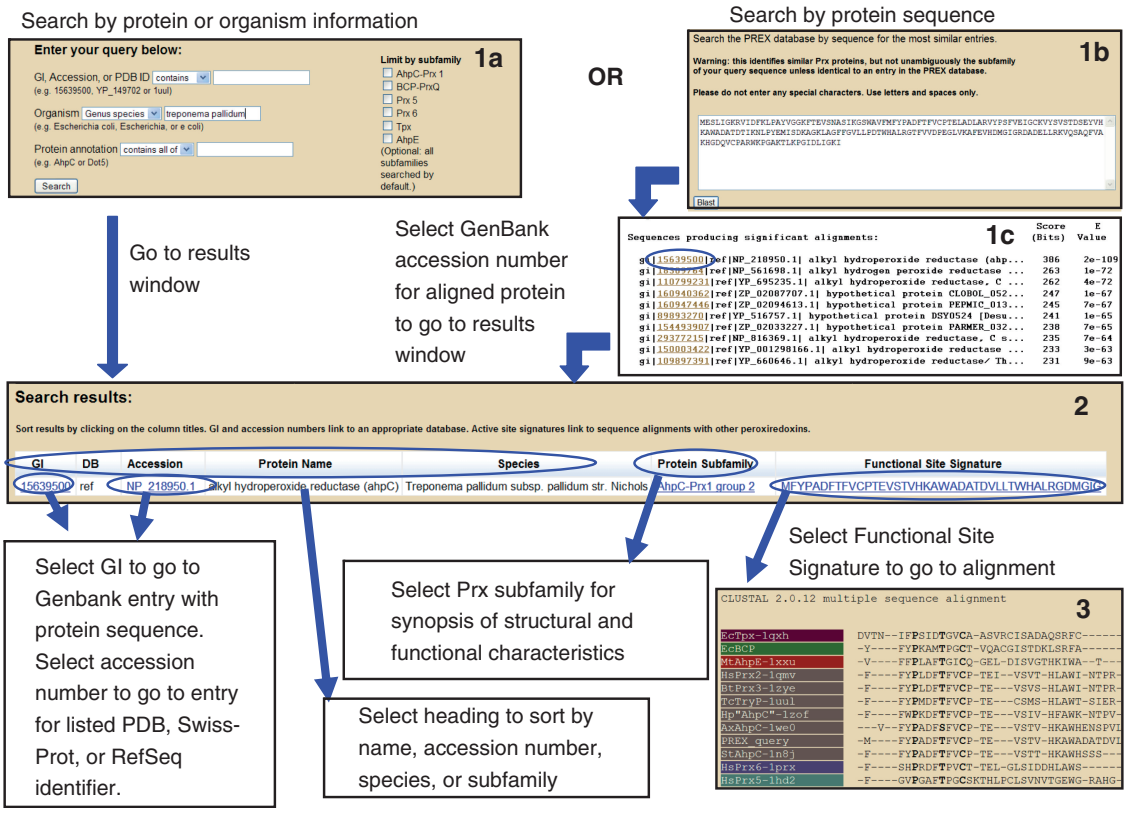

Figure 2. Examples of queries and results from PREX. The screenshot shown in Box 1a represents a taxonomy search for database members from *Treponema pallidum*. Text searches of the PREX database can also be conducted by GI, accession numbers, PDB ID or protein annotation. If a matching protein is identified, the user is taken directly to the results window; the screenshot shown in Box 2 represents the single Prx found in *T. pallidum*. Protein sequence searches of PREX (Box 1b) utilize BLAST to identify PREX database members with high sequence similarity to the query sequence. Shown in Box 1c is the BLAST output obtained after searching the full sequence of *T. pallidum* AhpC. Selecting the GI number of one of the identified proteins will direct the user to the results window for that PREX database member (Box 2). Selecting the functional site signature generates a multiple sequence alignment (Box 3) containing the functional site signatures for the selected PREX database protein (labeled as PREX_query), 4–5 selected members of the same subfamily and one representative from each of the other subfamilies. If accessed through a BLAST search, the multiple sequence alignment also includes the full sequence of the original sequence query (labeled as BLAST_query). Colors in Box 3 identify the subfamily assignment for each signature. In bold is the PXXX(T/S)XXC sequence motif that is invariant at the active site of Prx proteins (16).

accessed through a link associated with the subfamily name. The GenBank entry containing the protein sequence is accessed through a link associated with the GI number and associated entries in UniProt (19), EMBL (20) and the RCSB Protein databank (21) may be accessed through links in the column labeled Accession. The output is in a format than can be easily sorted by the user.

In addition, a sequence alignment containing the functional site signature for each PREX entry aligned with signatures from 4 to 5 members of the same subfamily and a representative member from each of the other subfamilies can be accessed by selecting the functional site signature on the results page (Figure 2). Alignments are created with a local copy of ClustalW, version 2.0.12 (22) (parameters for Gap Open Penalty and Gap Extension Penalty are set to 5 and 0.5, respectively; all other parameters are set to default values). When a BLAST search has been performed, the alignment includes the full sequence of the protein query in addition to the chosen PREX database signature. Protein names are colored according to their subfamily assignments for all but the BLAST query sequence. This allows the user to verify the presence of the PXXXT/SXXC motif (this motif is defined as mandatory for a functional Prx) and compare the aligned query sequence with the functional site signatures of the selected PREX database member and other Prxs from within and outside that subfamily.

References

Resources pertaining to the Prx family are listed in the reference section, providing key links and literature references to help direct researchers new to the Prx field.

FURTHER DEVELOPMENTS AND DATABASE MAINTENANCE

In order to provide more flexible search options, functionality will be added to PREX by the implementation of a taxonomy browser that will allow the user to identify all Prxs within a user-selected taxonomic group. We also plan to develop a search algorithm based on DASP that will utilize subfamily profiles to evaluate subfamily assignment for user entered sequences. Because this method would rely on sequence conservation in the region of the functional site, it would allow users to obtain reliable subfamily assignments for proteins not currently found in the PREX database. Areas of future development are indicated by the blue ovals in Supplementary Figure S1.

Database maintenance and updating is crucial. As structures of additional Prx proteins become available, their signatures can be used to improve our subfamily profiles and, thus, our ability to identify and classify Prxs from the GenBank (nr) database (23). New searches of GenBank (nr) will provide annotations of newly deposited sequences. Because the current DASP search method is manual and somewhat laborious, we have established a collaboration which aims to develop an automated pipeline for doing this search. Once this pipeline is developed, we plan to identify new Prx structures from the RCSB protein databank (21) and to utilize the pipeline to produce profiles and search GenBank(nr) to generate an updated database regularly. References and subfamily descriptions will be updated yearly to include new key references and to describe current advances in the Prx field.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Special thanks to Stacy Reeves, Chananat Klomsiri, Derek Parsonage and Edward Pryor for advice and for database testing.

FUNDING

National Science Foundation (MCB0517343 to J.S.F.); National Institutes of Health (F32 GM074537 to K.J.N. and RO1 GM050389 to L.B.P.). Calculations to identify and classify subfamily members for the database were performed on Wake Forest University's DEAC cluster, http://www.deac.wfu.edu, supported by a Shared University Research award from IBM, Inc. for storage hardware and by the Wake Forest Information Systems department. Funding for open access charge: MCB0517343 and F32 GM074537.

Conflict of interest statement. None declared.

REFERENCES

1. Poole,L.B. (2007) The catalytic mechanism of peroxiredoxins. In Flohé,L. and Harris,J.R. (eds), *Peroxiredoxin Systems*. Springer, NY, pp. 61–81.
2. Wood,Z.A., Schröder,E., Harris,J.R. and Poole,L.B. (2003) Structure, mechanism and regulation of peroxiredoxins. *Trends Biochem. Sci.*, **28**, 32–40.
3. Hofmann,B., Hecht,H.J. and Flohe,L. (2002) Peroxiredoxins. *Biol. Chem.*, **383**, 347–364.
4. Veal,E.A., Day,A.M. and Morgan,B.A. (2007) Hydrogen peroxide sensing and signaling. *Mol. Cell*, **26**, 1–14.
5. Zhang,B., Wang,Y. and Su,Y. (2009) Peroxiredoxins, a novel target in cancer radiotherapy. *Cancer Lett.*, **286**, 154–160.
6. Kang,S.W., Rhee,S.G., Chang,T.S., Jeong,W. and Choi,M.H. (2005) 2-Cys peroxiredoxin function in intracellular signal transduction: therapeutic implications. *Trends Mol. Med.*, **11**, 571–578.
7. Hall,A., Nelson,K., Poole,L.B. and Karplus,P.A. (2010) Structure-based insights into the catalytic power and conformational dexterity of peroxiredoxins. *Antioxid. Redox Signal*, doi:10.1089/ars.2010.3624.
8. Knoops,B., Loumaye,E. and Van Der Eecken,V. (2007) Evolution of the peroxiredoxins. In Flohé,L. and Harris,J.R. (eds), *Peroxiredoxin Systems*. Springer, NY, pp. 27–40.
9. Copley,S.D., Novak,W.R. and Babbitt,P.C. (2004) Divergence of function in the thioredoxin fold suprafamily: evidence for evolution of peroxiredoxins from a thioredoxin-like ancestor. *Biochemistry*, **43**, 13981–13995.
10. Karplus,P.A. and Hall,A. (2007) Structural survey of the peroxiredoxins. In Flohé,L. and Harris,J.R. (eds), *Peroxiredoxin Systems*. Springer, NY, pp. 41–60.
11. Koua,D., Cerutti,L., Falquet,L., Sigrist,C.J., Theiler,G., Hulo,N. and Dunand,C. (2009) PeroxiBase: a database with new tools for

peroxidase family classification. *Nucleic Acids Res.*, **37**, D261–D266.
12. Nelson,K.J., Knutson,S.T., Soito,L., Klomsiri,C., Poole,L.B. and Fetrow,J.S. (2010) Analysis of the peroxiredoxin family: using active site structure and sequence information for global classification and residue analysis. *Proteins* (in press).
13. Cammer,S.A., Hoffman,B.T., Speir,J.A., Canady,M.A., Nelson,M.R., Knutson,S., Gallina,M., Baxter,S.M. and Fetrow,J.S. (2003) Structure-based active site profiles for genome analysis and functional family subclassification. *J. Mol. Biol.*, **334**, 387–401.
14. Huff,R.G., Bayram,E., Tan,H., Knutson,S.T., Knaggs,M.H., Richon,A.B., Santago,P. 2nd and Fetrow,J.S. (2005) Chemical and structural diversity in cyclooxygenase protein active sites. *Chem. Biodivers.*, **2**, 1533–1552.
15. Fetrow,J.S. (2006) Active site profiling to identify protein functional sites in sequences and structures using the Deacon Active Site Profiler (DASP). *Curr. Protoc. Bioinformatics*, Chapter 8, Unit 8 10.
16. Fomenko,D.E. and Gladyshev,V.N. (2003) Identity and functions of CxxC-derived motifs. *Biochemistry*, **42**, 11214–11225.
17. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
18. Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
19. UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
20. Leinonen,R., Akhtar,R., Birney,E., Bonfield,J., Bower,L., Corbett,M., Cheng,Y., Demiralp,F., Faruque,N., Goodgame,N. *et al.* (2010) Improvements to services at the European Nucleotide Archive. *Nucleic Acids Res.*, **38**, D39–D45.
21. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
22. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
23. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2008) GenBank. *Nucleic Acids Res.*, **36**, D25–D30.
24. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
25. Bailey,T.L. and Gribskov,M. (1998) Methods and statistics for combining motif match scores. *J. Comput. Biol.*, **5**, 211–221.
26. Bailey,T.L. and Gribskov,M. (1998) Combining evidence using *P*-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.