Classification of Peroxiredoxin Subfamilies Using Regular Expressions

Jae Kyung Chon[1,2], Jongkeun Choi[1], Sang Soo Kim[3]* and Whanchul Shin[1,2]*

[1]Department of Chemistry, [2]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-742, Korea, [3]Department of Bioinformatics, Soongsil University, Seoul 156-743, Korea.

Abstract

Peroxiredoxins (Prx's) are a superfamily of peroxidases that are ubiquitous in all super-kingdoms. Previous biochemical and structural studies have suggested that Prx's could be divided into five subfamilies (1-Cys, Typical 2-Cys, Atypical 2-Cys C-, L- and R- types). In this work, we have developed a set of regular expression patterns describing subfamily-specific spatial constraints of the key catalytic residues. Using these patterns, 1,016 Prx's available in public databases were classified into the five subfamilies. Our method performed well for most of the types except for Atypical 2 Cys R type.

Keywords: peroxiredoxin, regular expression, subfamily classification

Introduction

Reactive oxygen species (ROS), such as the superoxide anion radical ($O_2^{\bullet -}$), hydrogen peroxide (H_2O_2) and the hydroxyl radical ($HO^{\bullet}$), are intermediately formed during the univalent reduction of oxygen to water, and are also produced during β-oxidation of fatty acids and upon exposure to radiation, metals, and redox drugs (Nordberg and Arnér, 2001). H_2O_2 itself is not very reactive, but can be further reduced to the extremely reactive $HO^{\bullet}$. ROS can cause molecular damage to various cellular components such as lipids, proteins, and nucleic acids, leading to cell death, and were thus considered to be highly toxic byproducts of oxygen metabolism and of no useful biological significance. To minimize the damaging effects of ROS, aerobic organisms evolved both non-enzymatic and enzymatic antioxidant defenses (Ahmad, 1995). Enzymatic defenses include superoxide dismutases which convert $O_2^{\bullet -}$ to H_2O_2, and catalases, gluthathione peroxidases and peroxiredoxins which convert H_2O_2 to H_2O. Recent findings show that, in addition to its deleterious effects, H_2O_2 is produced transiently in response to the activation of many cell surface receptors and serves as a ubiquitous intracellular messenger at sub-toxic concentrations (Rhee *et al.*, 2005).

Peroxiredoxins (Prx's) are a large family of antioxidant enzymes which are abundant, in several isoforms, in yeast, plant, and animal cells and in most eubacteria and archaea (Chae *et al.*, 1994; Rhee *et al.*, 2001; Hofmann *et al.*, 2002). Prx's reduce deleterious H_2O_2 or alkyl hydroperoxides utilizing the thiol group of a cysteine residue (Wood *et al.*, 2003b), and in some cases are involved in the decomposition of highly toxic peroxynitrite (Bryk *et al.*, 2000). Some Prx's also play significant roles in receptor signaling as, after completion of its mission as an intracellular messenger, the timely elimination of H_2O_2 is critical (Wood *et al.*, 2003a; Rhee *et al.*, 2005).

All Prx's, belonging to the thioredoxin-fold superfamily, contain a conserved 'peroxidatic' Cys (C_P) in the N-terminal portion and share the same peroxidatic active -site structure (Wood *et al.*, 2003b). The C_P residue is oxidized by peroxides to a cysteine sulfenic acid (C_P-SOH) intermediate. Prx's are classified into either the 2-Cys or 1-Cys types, based on the occurrence of the 'resolving' Cys (C_R) (Rhee *et al.*, 2001; Hofmann *et al.*, 2002). In 2-Cys Prx's, the C_P-SOH and C_R-SH react and form a disulfide (C_P-S-S-C_R). The stable disulfide form is then reduced by one of several cell-specific disulfide oxidoreductases (e.g., thioredoxin (Trx), tryparedoxin, AhpD, or AhpF), completing the catalytic cycle. In 1-Cys Prx's, the sulfenic acid is directly recycled *via* oxidoreductases such as Trx and glutaredoxin. The 2-Cys Prx's have been subdivided into either 'typical' or 'atypical' types depending on the location of the C_R residue. In typical 2-Cys Prx's, the C_P-SOH reacts with the C_R residue located in the C-terminal arm from the other subunit. In contrast, the C_R residue in atypical 2-Cys Prx's resides within the same subunit. As the atypical 2-Cys Prx's have been further subdivided into 'L', 'C', or 'R' type subfamilies, also depending on the

*Corresponding authors: E-mail sskimb@ssu.ac.kr.,
Tel +82-2-820-0457, Fax +82-2-824-4383,
E-mail nswcshin@snu.ac.kr,
Tel +82-2-880-6656, Fax +82-2-889-5304

Accepted 10 May 2005

Table 1. Protein family and domain databases used in he retrieval of Prx's

Database	Access Number	Annotation	Entry number
Pfam (Releases 17)	PF00578	TSA/AhpC families	1023
PROSITE (Releases 19.2)	PS01265	Tpx Family (S-x-D-L-P-F-[AS]-x(2)-[KRQ]-[FWI]-C)	38
BLOCKS (Version 14.1)	IPB 00866A	AhpC-TSA	834
InterPro (Releases 10)	IPR00866	Alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen	1075
	Total		2970
	Non-redundant entries		1085

spatial location of the C_R residue (Choi *et al.*, 2003), there are five unique Prx's subfamilies in total. Sub-classification of all known Prx's at this fine scale may provide further insights into the biochemical mechanism and protein evolution of this important protein family.

Biochemical functions of a protein are typically predicted from its sequence similarity to those proteins whose functions or 3D structures have been experimentally determined. The first line of sequence search method is using the pair-wise sequence similarity algorithms such as BLAST (Altschul *et al.,* 1997) and FASTA (Pearson, 1994). Such methods are appropriate for detecting sequences of relatively high homology, but are not efficient in detecting distant members of a divergent protein family such as Prx's. The sensitivity of sequence search methods has been improved by profile - or motif-based analysis, which uses information derived from multiple sequence alignments to construct and search for sequence patterns. Unlike pairwise sequence similarity methods, the latter can exploit additional information, such as the position and identity of residues that are conserved throughout the family, as well as variable insertion and deletion probabilities. The hidden Markov model is one appropriate method to express a profile or motif because it provides a solid statistical foundation to model information from a multiple sequence alignment (Eddy, 1998). While profile and hidden Markov model methods perform appropriately in detecting homologous proteins, these generally afford only a very high level of functional classification. Most of the protein family databases based on profile, motif or domain, such as Pfam (Bateman *et al.*, 2000), PROSITE (Hofmann *et al.,* 1999), BLOCKS (Henikoff *et al.,* 1999), PRINTS (Attwood *et al.,* 2000) and InterPRO (Apweiler *et al.,* 2000), have superfamily of the Prx's. But these databases do not include the classification of Prx's subfamily. We have classified Prx's to distinguish between subfamilies within a structurally and functionally diverse superfamily. This work represents the classification tool and method focused on a specific protein superfamily, Prx's.

Methods

Collecting the sequences of peroxiredoxin family from domain and motif databases

Various protein superfamily databases are publicly available in the internet. We queried the following databases with a keyword 'peroxiredoxin': Pfam, PROSITE, InterPro and BLOCKS. The resulting superfamily had annotations such as TSA (thiol-specific antioxidant), AhpC (alkyl hydroperoxide reductase), Tpx (thioredoxin peroxidase) and peroxiredoxin. The protein entries belonging to these families were downloaded and the redundant entries were removed. Summary of the collected sequences used in this analysis is provided in Table 1.

Compilation of the relevant PDB entries

We retrieved entries from PDB (Berman *et al.*, 2000, release May 31, 2005) by querying a keyword "peroxiredoxin", resulting 25 hits. From the manual inspection of the description field, 21 structures were identified as the true members of Prx's family (Table 2). These entries were classified into 4 subfamilies of Prx's, except Atypical 2 Cys C Type for which no crystal structure has been reported yet.

Defining regular expression

A regular expression (abbreviated as regexp, regex or regxp) is a string that describes or matches a set of strings, according to certain syntax rules. Regular expressions are used by many text editors and utilities to search and manipulate bodies of text based on certain patterns (Stubbletine, 2003; Fridl, 1997). The protein motifs can be represented as regular expressions. And the structural properties of motif can be formulated by regular expression. The advantages of using regular expressions are 1) fairly concise and easy to

Table 2. PDB entries of Prx's

No	PDB ID	Redox state	SWISS-PROTAC Number
		Atypical 2 Cys L type	
1	1QXH	S-S	P37901
2	1PSQ	SH	P0A4M5
3	1Q98	S-S	Q57549
4	1XVQ	S-S	P66952
		Atypical 2 Cys R type	
5	1H4O	SH	P30044
6	1HD2	SH	
7	1OC3	S-S	
8	1URM	SOH	
		Typical 2 Cys	
9	1QMV	SO$_2^-$	P32119
10	1QQ2	S-S	Q63716
11	1KYG	S-S	P0A251
12	1N8J	SH	
13	1E2Y	SH	Q9TZX2
14	1UUL	SH	O96763
		1 Cys	
15	1PRX	SOH	P30041
16	1NM3	SH	P44758
17	1XCC	SH	Q7RGR1
18	1XIY	SO$_3^{2-}$	Q8IBG7
19	1XVW	SOH	P65688
20	1XXU	SH	P65688
21	1VGS	SH	Q9Y9L0

understand, 2) well known algorithms for matching, 3) fairly easy to display and 4) accepting insertion or deletion.

There are several key amino acid residues that have been recognized as playing critical roles in the peroxiredoxin activity. For example, a Pro (P) is located several residues upstream of the peroxidatic Cys (C$_P$), while an Arg (R) in C-terminal neutralizes the negative charge of the intermediate. Their importance is also supported by the strong conservation of sequences and structures (Choi *et al.*, 2003). Prx's family can be classified by the existence and location of resolving Cys (C$_R$) that form S-S bond with the peroxidatic Cys (C$_P$). Sequence patterns of C$_P$ and C$_R$ were learned from the 21 PDB entries. We superposed the protein structures of Prx's family with Deep-Viewer (Guex and Peitsch, 1997) and the subsequent visual inspection located the critical P, C$_P$, C$_P$ and R on sequences. See Fig. 1 for schematic diagram of key residues. The minimum and maximum number of intervening amino acids residues were calculated from these sequences and converted into regex forms. These forms were further modified by allowing potential insertions and deletions. For example, atypical 2 Cys C type of peroxiredoxin could be defined by the pattern “P .{6} C$_P$.{4} C$_R$.{60,90} R”. This regular expression is translated as: (Pro) - (6 any residues) (Cys) (4 any residues) (Cys) (between 60 and 90 of any residues) (Arg). Details of pattern used in classification are provided in Table 3. A short program for matching the regular expressions with the amino acid sequences was written in Perl Verseion 5.8.4 (the source code available upon request).

Results

Classification using regular expression

A set of Prx's family was converted into FASTA format file (a total of 1,085 sequences). Some sequences were excluded from the set, based on following criteria: (1) The sequences whose amino acid length was too short (100 AA) or too long (300 AA) were removed, as Prx's molecular weight typically ranges between 20 ~ 30kD. (2) The sequences where the pattern “P.{6}C” was located within 25 AA from the N-terminus was removed,

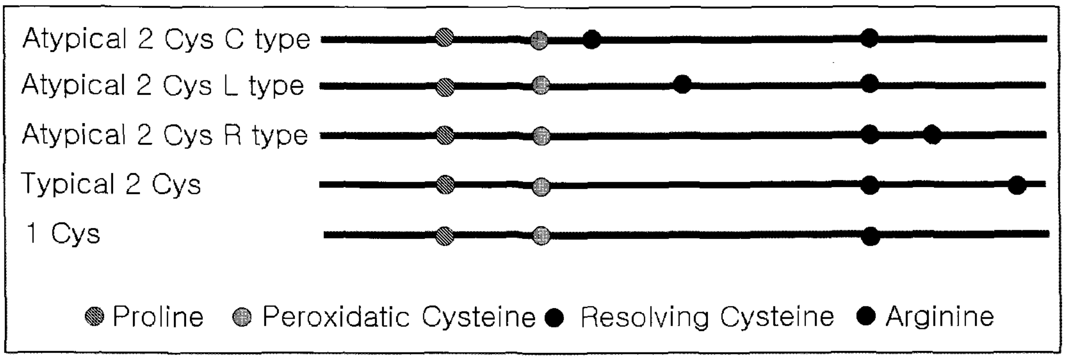

Fig. 1. Schematic diagram of key catalytic residues and their relative positions in the protein sequences. Note that the resolving Cys is located toward the N-terminal side of Arg for C and L types of Atypical 2 Cys families, while it is found in the C-terminal side of Arg for Atypical 2 Cys R type and Typical 2 Cys families.

Table 3. List of regular expression patterns

Subfamily	Pattern (regular expression) (C_P, C_R , R)
Atypical 2 Cys C type	P .{6} C_P .{4} C_R .{60,90} R
Atypical 2 Cys L type	P .{6} C_P .{28,38} C_R .{60,90} R
Atypical 2 Cys R type	P .{6} C_P .{60,90} R .{20,25} C_R
Typical 2 Cys	P .{6} C_P .{60,90} R .{40,50} C_R
1 Cys	P .[^C] .{4} C_P .{60,90} R

as they lost some important secondary structure elements characteristic to Prx's. The final set of Prx's includes 1,016 entries, with which we ran the search for classification of Prx's by using the defined regular expressions. Since the regex for Atypical 2 Cys L type is inclusive of and less restrictive than that of Typical 2 Cys, we gave higher priority to Typical 2 Cys. Classification of 1 Cys is difficult due to the lack of resolving Cys (C_R). Accordingly, we assigned those without any other matches than the basic Prx's pattern to 1 Cys. See Fig. 2 for the flowchart of the analysis. Summary of classification is provided in Table 4. The full list of proteins for each subfamily is available at the supplementary web site (http://xray2.snu.ac.kr/Prx/).

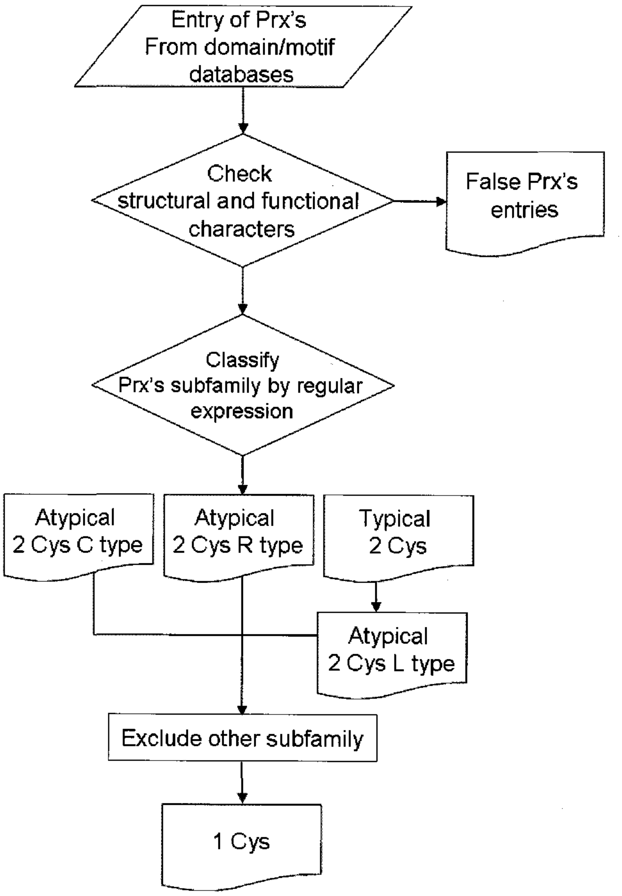

Fig. 2. Flowchart of the classification based on regex patterns matching.

Validation of the classification

In order to validate our classification result, we performed multiple sequence alignments (MSA) of each resulting cluster. For this, we used the ClustalW program with default settings (Higgins *et al.*, 1994). The key catalytic residues (P, C_P, C_P and R) and all the secondary structure elements were aligned extremely well. Compared to the alignment provided by Pfam that included all potential Prx's in a single MSA output, our subfamily-specific MSA provided much improved and well defined alignments (see the supplementary web site for details), demonstrating the validity of our classification scheme. Some of the sequences matched more than one regex patterns. For example, Uniprot accession O96763, known as a member of Typical 2 Cys on the structural basis (PDB ID 1UUL), matched the regex pattern of either Atypical 2 Cys C type or Typical 2 Cys. For these ambiguous sequences, we confirmed their true membership by manual inspection of the subfamily-specific MSA output. For example, O96763 was well absorbed in the MSA of Typical 2 Cys sequences, while it became an outlier in the MSA of Atypical 2 Cys L type sequences. In this way, we were able to unambiguously confirm the true membership of all the previously ambiguous sequences (Table 5). The error rates of the regex-based classification were then 7/140, 1/142, 7/51, and 1/414 for Atypical 2 Cys C, L, R and Typical 2 Cys types, respectively. Our simple method performed extremely well for most of the types except for Atypical 2 Cys R type. It appears that the regex for the latter is subtle and prone to include false positives. We may need more key residues yet to be discovered based on the reaction mechanism of the enzyme. It should also be noted that the multiple sequence alignments hinted the potential of even finer sub-classification.

Table 4. Results of the classification based on regex pattern matching

Subfamily	Regular expression	Number of hits	Number of ambiguous sequences
Atypical 2 Cys C type (C)	P .{6} C_P.{4} C_R .{60,90} R	140	L type : 2 R type : 3 T type : 8
Atypical 2 Cys L type (L)	P .{6} C_P .{28,38} C_R .{60,90} R	146	C type : 2 R type : 4
Atypical 2 Cys R type (R)	P .{6} C_P .{60,90} R .{20,25} C_R	51	C type : 3 L type : 4 T type : 8
Typical 2 Cys (T)	P .{6} C_P.{60,90} R .{40,50} C_R	417	C type : 8 L type : 108 R type : 8
1 Cys	Exclusion of the others	293	

Table 5. Results of the validation based on multiple sequence alignments (MSA)

Subfamily	Number of ambiguous sequences	True members	False positives
Atypical 2 Cys C type	2 (L type) 3 (R type) 8 (T type)	2	7 (Typical 2 Cys)
Atypical 2 Cys L type	2 (C type) 4 (R type)	3	1 (Atypical 2 Cys C type)
Atypical 2 Cys R type	3 (C type) 4 (L type) 8 (T type)	2	2 (Atypical 2 Cys C type) 5 (Typical 2 Cys)
Typical 2 Cys	8 (C type) 108 (L type) 8 (R type)	8	1 (Atypical 2 Cyc C type)

Discussion

Peroxiredoxins form such a divergent family with over 1,000 sequences in the public sequence databases, occurring in from prokaryotes to mammals. Current protein domain or family databases list them as a single large family and thus are not much informative in understanding detailed molecular mechanism of this biologically important class of proteins. It should also be noted that these databases except for PROSITE contain contaminations, that is, classifying some non-Prx's as Prx's. For example, Q8RB02, belonging to thioredoxins, do not share the key catalytic residues but it shares a noticeable global sequence similarity with some Prx's. This is not at all surprising considering that Prx's belong to the Trx structural fold, demonstrating the risk associated with profile-based methods, albeit significant advantages. We employed PROSITE-like sequence patterns, which may be too simple to sub-classify such a large and diverse family. We successfully augmented its weakness by meticulous choice of patterns based on the Prx's reaction mechanism as well as their structural characteristics and by hierarchical classification of subfamilies. However, our method showed some small but non-negligible mis-classifications. Future improvement may require even more key catalytic residues to be included in the regex patterns. Alternatively, judicious use of multiple sequence alignments combined with profile-based models may be applied to our pattern-based subfamilies. Besides improving the accuracy of sub-classification, identifying even finer sub-classes should be also important as it may shed light on understanding the detailed biochemical mechanism. Accurate classification of a protein family is also critical in constructing molecular phylogeny and thus understanding evolution of Prx's.

Acknowledgements

This work was supported by a grant (CBM2-B512-001-1-0-0) from the Center for Biological Modulators of the 21st Century Frontier R&D Program, the Ministry of Science and Technology, Korea. J.C. is a recipient of the BK21 fellowship.

References

Ahmad, S. (1995). Oxidative stress and antioxidant defenses in biology (New York: Chapman and Hall).

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.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T.M., Pagni, M., Servant, F., Sigrist, C.J., and Zdobnov, E.M. (2000). InterPro - An integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* 16, 1145 1150.

Attwood, T.K., Croning, M.D., Flower, D.R., Lewis, A.P., Mabey, J.E., Scordis, P., Selley, J.N., and Wright, W. (2000). PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.* 28, 225 227.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., and Sonnhammer, E.L. (2000). The Pfam protein families database. *Nucleic Acids Res.* 28, 263 266.

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.

Bryk, R., Griffin, P., and Nathan, C. (2000). Peroxynitrite reductase activity of bacterial peroxiredoxins. *Nature* 407, 211 215.

Chae, H.Z., Robinson, K., Poole, L.B., Church, G., Storz, G., and Rhee, S.G. (1994). Dimerization of thiol-specific antioxidant and the essential role of cysteine 47. *Proc. Natl. Acad. Sci. USA* 91, 7017 7021.

Choi, J., Choi, S., Choi, J., Cha, M.K., Kim, I.H., and Shin, W. (2003). Crystal structure of Escherichia coli thiol peroxidase in the oxidized state: insights into intramolecular disulfide formation and substrate binding in atypical 2-Cys peroxiredoxins. *J. Biol. Chem.* 278, 49478 49486.

Eddy, S.R. (1998). Profile hidden Markov models. *Bioinformatics* 14, 755 763.

Fridl, J.E.F. (1997). Mastering regular expressions (Sebastopol, CA: O'Reilly Media).

Guex, N. and Peitsch, M.C. (1997). SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* 18, 2714-2723.

Henikoff, S., Henikoff, J.G., and Pietrokovski, S. (1999). Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* 15, 471 479.

Higgins, D., Thompson, J., Gibson, T., 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.

Hofmann, B., Hecht, H.J., and Flohé, L. (2002). Peroxiredoxins. *Biol. Chem.* 383, 347 364.

Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. (1999). The PROSITE database, its status in 1999. *Nucleic Acids Res.* 27, 215 219.

Nordberg, J. and Arnér, E.S. (2001). Reactive oxygen species, antioxidants, and the mammalian thioredoxin system. Free. *Radic. Biol. Med.* 31, 1287-1312.

Pearson, W.R. (1994). Using the FASTA program to search protein and DNA sequence databases. *Meth. Mol. Biol.* 25, 365 389.

Rhee, S.G., Kang, S.W., Jeong, W., Chang, T.-S, Yang, K.-S., and Woo, H.A. (2005). Intracellular messenger function of hydrogen peroxide and its regulation by peroxiredoxin. *Curr. Opin. Cell Biol.* 17, 183 189.

Rhee, S.G., Kang, S.W., Chang, T.-S., Jeong, W., and Kim, K. (2001). Peroxiredoxin, a novel family of peroxidases. *IUBMB Life* 52, 35 41.

Stubbletine, T. (2003). Regular expression pocket reference (Sebastopol, CA: O'Reilly Media).

Wood, Z.A., Poole, L.B., and Karplus, P.A. (2003a). Peroxiredoxin evolution and the regulation of hydrogen peroxide signaling. *Science* 300, 650 653.

Wood, Z.A., Schröder, E., Harris, J.R., and Poole, L.B. (2003b). Structure, mechanism and regulation of peroxiredoxins. *Trends Biochem. Sci.* 28, 32 40.