Supplemental Data

Analysis of the peroxiredoxin family: using active site structure and sequence information for global classification and residue analysis

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

Short Title: Active site based analysis of Prx subfamilies

Supplementary Material Table of Contents

Supplemental Methods

DASP algorithm to utilize functional site profiles to search sequence databases

The steps for utilizing a functional site profile to search protein sequences are summarized in Figure S2 and are described in detail elsewhere [1]. Briefly, profile motifs are identified by traversing the functional site profile from left to right searching for continuous fragments of at least three residues in length that align. These fragments correspond to the fragments identified from the protein structures (indicated by alternating upper and lower case letters in each signature in Figure 2B). A motif is identified at *i* if the majority of the sequences in the profile have a fragment or a portion of a fragment between *i* and *j*, with *j* being the position in the profile where the fragment ends. Not all sequences contain the motif fragment; thus not all sequences may be included in every motif identified in the profile. Once all motifs in the profile have been identified, individual multiple sequence alignments (MSAs) are created for each motif. A position specific scoring matrix (PSSM) for each motif is created by iterating over the columns of the MSA and tallying the observed counts (the number of occurrences of each residue) and the pseudocounts (based on the overall frequency of the amino acid in the background database) in each column [2]. Counts for each residue in a motif are summed and normalized by the sum of the number of columns in the MSA and the pseudocount weight.

The PSSM for each motif is used to search all sequences of the database using a sliding window procedure [2]. Once a motif from the profile is matched to a position in a protein sequence, a score, S_i, for the segment of sequence *s* beginning at position *i* is obtained by summing the corresponding entries from the PSSM, as previously described [3]:

$$S_i = \sum_{j=1}^{n} m_{s(i+j-1),j}$$

where *s(x)* is the letter at position *x* in sequence *s*, $m_{a,x}$ is the score for residue *a* at position *x* in the PSSM, and *j* is the current column of the PSSM. A p-value is then obtained for the score representing the probability of finding a match as good as the observed match in a random spot of a random sequence.

The p-values for all motif matches in a given sequence are combined using QFAST[4]. Briefly, the p-value for each motif is normalized for the length of the motif, and the normalized

p-values from all the motifs are multiplied together to obtain the final product. The p-value for the product represents the statistical significance for the match of each sequence to the entire profile.

Calculation of mean and standard deviation entropy values for residue conservation

The entropy values were calculated as described previously [5] for every residue position across a sequence alignment of 204 peroxiredoxins from all subfamilies. The 204 proteins were selected from PSI-BLAST searches (cutoff score e^{-40}) using query sequences *H. sapiens* Prx5 (1hd2), *H. sapiens* Prx6 (1prx), *S. typhimurium* AhpC (1yep), *H. influenza* Tpx (1q98), and *S. pneumoniae* Tpx (1psq). Entropy values were not calculated for any position in the alignment where less than 10 of the sequences had a residue at that position. For all positions in these 204 proteins, the mean entropy value was 1.158, and the standard deviation was 0.548. We thus considered as conserved, those residues with an entropy value lower than 0.61 (the mean minus one standard deviation). Entropy values for each Prx subfamily were calculated using all of the sequences identified by DASP after removing sequences with no Prx motif or hit with a more significant p-value in another subfamily search (Table I, Total after edits).

Supplemental Results

Selection criteria for Prx key residues

Functional site signatures were created for all Prxs with structural coordinates available in the RCSB database as of Jan 2008 [6] (Table SI). The first step in creation of signatures is the identification of *key residues*—residues that are known to be important in the activity of a functional site. When the C_P was used as the only key residue to extract the functional site signature, the protein fragments identified were short and the functional site signatures did not align well. To expand the signatures, other residues were assessed based upon their conservation across known Prxs, including the essentially conserved residues Pro39, Thr43, and Arg119 (numbering for *S. typhimurium* AhpC, Figure 1). Arg119 was present in all of the signatures and inclusion of Arg119 as a key residue did not add to the functional site signatures because the surrounding sequence was not well conserved. The C_R was also not used as a key residue

because it is not present in all Prxs, is found in different locations in various subfamilies (Figure S1), and is often not located close to the C_P in the fully folded structures. Initial alignments indicated that all of the Prx signatures contained either a Trp or a Phe residue that was present in the same location across all of the structures and aligned sequences (Trp81 in *Salmonella typhimurium* AhpC), suggesting that it might be important for the mechanism of this family. Including Trp81 as a key residue in the DASP analysis provided a sequence fragment that was long enough to properly align the conserved Trp, which was determined to be important for subsequent DASP searches of the sequence database. Therefore, final functional site signatures were obtained using the residues equivalent to C_P (Cys46), Pro39, Thr43, and Trp81 as the key residues.

Selection of DASP p-value cutoff

To determine the appropriate p-value cutoffs for the GenBank(nr) search, functional site profiles were created for each Prx subfamily. DASP was used to search the RCSB PDB database because we know the correct subfamily assignment for all the structurally characterized Prxs, Table SI). At p-values more significant than 10^{-5}, all hits returned were peroxiredoxins. At p-values of 10^{-4}, other proteins that did not include the conserved PxxxTxxC motif were starting to be returned, and at 10^{-3}, the vast majority of the results could not be considered peroxiredoxins. In the case of the searches with Prx5 and Tpx, only members of the Prx5 and Tpx class, respectively were returned, even at p-values as low as 10^{-3}. Searches of the Prx6, AhpC/Prx1, or BCP/PrxQ subfamilies were able to pull up members of the other two classes at p-values between 10^{-5}-10^{-8}. With a p-value cutoff of 10^{-8}, the searches were completely specific for the appropriate subfamily and this value was used for further analysis.

Profile scores can be used to identify Prx subfamilies

A score is calculated for each profile based on the level of conservation in the profile as described previously [8]. Work by Cammer *et al* indicated that functional site profile scores ranged from 0.04 - 1.0 for 193 known protein families [8] . Higher profile scores are correlated with more similarity at the functional site. The Prx5 profile (0.25) exhibits significant scores, indicating clear relationships between these proteins. The Tpx profile score (0.14) indicates significant

diversity within the Tpx proteins of known structure, but clustering shows a clear separation of this subfamily from the others (Figure 2A). A more significant profile score was obtained for the Prx6 profile (0.31) compared to the score for a combined Prx6, AhpC, and Prx1 profile (-0.04), indicating that the Prx6 subfamily is distinct from AhpC/Prx1 based on information at the molecular functional site. This analysis suggests that AhpC and Prx1 might also be distinguished, as scores for AhpC (0.32) and Prx1 (0.16) subfamilies individually are much more significant than scores for the combined AhpC/Prx1 subfamily (0.06). The original BCP/PrxQ profile score (0.18) was low, suggesting that the structural diversity of this subfamily is insufficient to produce a robust profile. It is also possible to use profile scores to determine the family/subfamily assignment of a protein; a significant decrease in the profile score upon the addition of a signature suggests that the protein has been misassigned [8]. Addition of the AhpE or BCP functional site signatures to any of the other Prx subfamilies dramatically decreased the score for the resulting profile, indicating that neither BCP nor AhpE were sufficiently similar to be considered as a member of another subfamily.

Engineered profiles can be used to obtain a more specific profile for subfamilies lacking sufficient structural representatives: the BCP/PrxQ example.

The PSSM method utilized by DASP is limited by the diversity of the family or subfamily members used to generate the PSSM as illustrated by analysis of the BCP/PrxQ subfamily. At the time of the original analysis, only two distinct sequences were available for structurally characterized members of the BCP/PrxQ subfamily (*Aeropyrum pernix* BCP, 2a4v and *Saccharomyces cerevisiae* BCP, 2cx4), and the resulting profile was of limited diversity. Clustering of all the functional site signatures identified by the DASP search indicated that these two structures are found in two of the smaller groups identified within this subfamily and are not representative of the subfamily as a whole (Figure S3A). The largest groups (labeled groups 1 and 2 in Figure S3A) did not have a representative in the profile; however, the biochemically characterized subfamily members including *Escherichia coli* BCP [9] and *Populus tremula* x *Populus tremuloides* PrxQ [10] are members of these larger groups. Thus, an engineered profile was developed (as described in Methods) for the BCP/PrxQ subfamily using these biochemically

characterized subfamily members to better represent subfamily diversity and to improve sequence searching.

The results of searching GenBank(nr) with both profiles are shown in Figure S3D and E. The original (less diverse) profile (Figure S3B) identified 810 putative subfamily members, while the engineered (more diverse) profile (Figure S3C) identified 1130 putative subfamily members. We cannot distinguish how much of the increase in the number of putative BCP/PrxQ sequences is due to the deposition of more sequences in the GenBank(nr) database (Jan 2008 and Jan 2009 for the original and engineered profiles, respectively); however, other data also suggest that the engineered profile is more robust and diverse. First, the number of identified sequences with an extremely significant p-value ($<10^{-20}$) is lower in the original profile than the engineered (13% and 38%, respectively; Figure S3 D and E) and the distribution of the remaining scores in the engineered search is more consistent with those of other subfamily searches. Second, the number of sequences identified by more than one subfamily search are fewer in the engineered (10, 0.88%) than the original (25, 3.1%) BCP/PrxQ profile. These results show that the composition of the original, structure-based profile affects the specificity and coverage of the sequences identified by the profile. The creation of engineered profiles can therefore be used to increase the power of the sequence searching technique for subfamilies that have few structural representatives.

DASP identifies three sites of conservation that may be important for Prx catalysis.

Other than the PXXXT/SXXC$_P$ motif [11] and Arg119 [11-13], our analysis identified only three residues that are highly conserved across all Prx functional site signatures (Figure 2B, highlighted in black). The location of both of these residues in representative Prxs are shown in Figure 4 (residues in pink). The first, the Trp noted earlier during optimization of the signatures, is Trp81 in *S. typhimurium* AhpC. This residue is replaced with a Phe in some Prxs, particularly in the BCP/PrxQ and Tpx subfamilies, where 72% and 98% of the subfamily members contain a Phe, respectively. It has previously been noted that Trp81 is conserved in the AhpC/Prx1 subfamily, and mutation of this residue has been shown to dramatically decrease the activity of some peroxiredoxins. For example, Trp81 has been mutated to Leu in a barley 2-Cys peroxiredoxin [14] and to His and Asp in *Crithidia fasciculata* tryparedoxin peroxidase [15] (both

members of the AhpC/Prx1 subfamily). In both cases, this mutation significantly decreased the activity of the protein (and stability in the case of the His and Asp mutations).

The second residue, Ser71 (AhpC, 1n8j numbering), was observed to be stringently conserved across all Prx structures and most of the signatures (Figure 3). This residue is located between the active site and the A-type interface and is part of a hydrogen bond network with other conserved residues (Figure 4, residues in orange; Figure 3, residues marked with #). Although Ser71 (Figure 4, pink) is conserved across all of the subfamilies except Prx5, the rest of the residues involved in this network differ in each subfamily. The role of this residue has not been explored experimentally.

The third residue, Glu49 (numbering from *S. typhimurium* AhpC, 1n8j), is conserved across the AhpC/Prx1 and Prx6 (Glu50 in *Homo sapiens* Prx6, 1prx) subfamilies (Figure 3B and C) and hydrogen bonds to Arg119 in some of the structures (Figure 4A and B, green). The Glu has been identified as characteristic of the "type 4" Prx subfamily which includes our AhpC/Prx1 and Prx6 subfamilies [16]. Although this Glu is not conserved in the other Prx subfamilies except AhpE, all of the subfamilies contain a residue at this position that is capable of hydrogen bonding to Arg119. In the BCP/PrxQ subfamily, this position (Glu52 in *A. pernix* BCP, 2cx4) is occupied by either a Glu (66%) or a Gln (33.5%). Members of the Tpx subfamily contain a Gln (31%), a Ser (58%) or a Glu (9.5%). In members of the Prx5 subfamily, there is a single residue insertion in this portion of the structure, described as an α-aneurism [17]; a conserved His is located at the same relative position in the *H. sapiens* Prx5 structure (1hd2, His51, Figure 4D, green) and has similar hydrogen bonding patterns as Glu49 in *S. typhimurium* AhpC (Figure 4B). This His is located one residue after the conserved Glu in sequence alignments (Figure S1) and in the signatures (Figure 2B). These observations suggest that hydrogen bonding is a key feature that this residue plays in all subfamilies and that variations in its pK_a might be important in the Prx mechanism in some subfamilies. This residue has also been identified in computational electrostatic studies as being a residue that interacts strongly with C_P [18]. Recent analysis of Prx active sites with bound substrate analogues revealed that this residue hydrogen bonds to the stringently conserved Arg and that the residue identity is at least partially responsible for determining the conformation of the conserved Arg [19].

Supplemental references

1. Huff RG, Bayram E, Tan H, Knutson ST, Knaggs MH, Richon AB, Santago P, 2nd, Fetrow JS. Chemical and structural diversity in cyclooxygenase protein active sites. Chem Biodivers 2005;2(11):1533-52.
2. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci U S A 1987;84(13):4355-8.
3. Bailey TL, Gribskov M. Methods and statistics for combining motif match scores. J Comput Biol 1998;5(2):211-21.
4. Bailey TL, Gribskov M. Combining evidence using p-values: application to sequence homology searches. Bioinformatics 1998;14(1):48-54.
5. Shenkin PS, Farid H, Fetrow JS. Prediction and evaluation of side-chain conformations for protein backbone structures. Proteins 1996;26(3):323-52.
6. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235-42.
7. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22(22):4673-80.
8. Cammer SA, Hoffman BT, Speir JA, Canady MA, Nelson MR, Knutson S, Gallina M, Baxter SM, Fetrow JS. Structure-based active site profiles for genome analysis and functional family subclassification. J Mol Biol 2003;334(3):387-401.
9. Jeong W, Cha MK, Kim IH. Thioredoxin-dependent hydroperoxide peroxidase activity of bacterioferritin comigratory protein (BCP) as a new member of the thiol-specific antioxidant protein (TSA)/Alkyl hydroperoxide peroxidase C (AhpC) family. J Biol Chem 2000;275(4):2924-30.
10. Rouhier N, Gelhaye E, Gualberto JM, Jordy MN, De Fay E, Hirasawa M, Duplessis S, Lemaire SD, Frey P, Martin F and others. Poplar peroxiredoxin Q. A thioredoxin-linked chloroplast antioxidant functional in pathogen defense. Plant Physiol 2004;134(3):1027-38.
11. Poole LB. The Catalytic Mechanism of Peroxiredoxins. In: Flohé L, Harris JR, editors. Peroxiredoxin Systems. New York: Springer; 2007. p 61-81.
12. Hofmann B, Hecht HJ, Flohe L. Peroxiredoxins. Biol Chem 2002;383(3-4):347-64.
13. Wood ZA, Schroder E, Robin Harris J, Poole LB. Structure, mechanism and regulation of peroxiredoxins. Trends Biochem Sci 2003;28(1):32-40.
14. Konig J, Lotte K, Plessow R, Brockhinke A, Baier M, Dietz KJ. Reaction mechanism of plant 2-Cys peroxiredoxin. Role of the C terminus and the quaternary structure. J Biol Chem 2003;278(27):24409-20.
15. Flohe L, Budde H, Bruns K, Castro H, Clos J, Hofmann B, Kansal-Kalavar S, Krumme D, Menge U, Plank-Schumacher K and others. Tryparedoxin peroxidase of Leishmania donovani: molecular cloning, heterologous expression, specificity, and catalytic mechanism. Arch Biochem Biophys 2002;397(2):324-35.
16. Copley SD, Novak WR, Babbitt PC. Divergence of function in the thioredoxin fold suprafamily: evidence for evolution of peroxiredoxins from a thioredoxin-like ancestor. Biochemistry 2004;43(44):13981-95.
17. Sarma GN, Nickel C, Rahlfs S, Fischer M, Becker K, Karplus PA. Crystal structure of a novel *Plasmodium falciparum* 1-Cys peroxiredoxin. J Mol Biol 2005;346(4):1021-34.

18. Salsbury FR, Jr., Knutson ST, Poole LB, Fetrow JS. Functional site profiling and electrostatic analysis of cysteines modifiable to cysteine sulfenic acid. Protein Sci 2008;17(2):299-312.
19. Hall A, Parsonage D, Poole LB, Karplus PA. Structural Evidence that Peroxiredoxin Catalytic Power Is Based on Transition-State Stabilization. J Mol Biol 2010.
20. Alphey MS, Bond CS, Tetaud E, Fairlamb AH, Hunter WN. The structure of reduced tryparedoxin peroxidase reveals a decamer and insight into reactivity of 2Cys-peroxiredoxins. J Mol Biol 2000;300(4):903-16.
21. Schroder E, Littlechild JA, Lebedev AA, Errington N, Vagin AA, Isupov MN. Crystal structure of decameric 2-Cys peroxiredoxin from human erythrocytes at 1.7 A resolution. Structure 2000;8(6):605-15.
22. Hirotsu S, Abe Y, Okada K, Nagahara N, Hori H, Nishino T, Hakoshima T. Crystal structure of a multifunctional 2-Cys peroxiredoxin heme-binding protein 23 kDa/proliferation-associated gene product. Proc Natl Acad Sci U S A 1999;96(22):12333-8.
23. Pineyro MD, Pizarro JC, Lema F, Pritsch O, Cayota A, Bentley GA, Robello C. Crystal structure of the tryparedoxin peroxidase from the human parasite *Trypanosoma cruzi*. J Struct Biol 2005;150(1):11-22.
24. Papinutto E, Windle HJ, Cendron L, Battistutta R, Kelleher D, Zanotti G. Crystal structure of alkyl hydroperoxide-reductase (AhpC) from *Helicobacter pylori*. Biochim Biophys Acta 2005;1753(2):240-6.
25. Cao Z, Roszak AW, Gourlay LJ, Lindsay JG, Isaacs NW. Bovine mitochondrial peroxiredoxin III forms a two-ring catenane. Structure 2005;13(11):1661-4.
26. Boucher IW, McMillan PJ, Gabrielsen M, Akerman SE, Brannigan JA, Schnick C, Brzozowski AM, Wilkinson AJ, Muller S. Structural and biochemical characterization of a mitochondrial peroxiredoxin from *Plasmodium falciparum*. Mol Microbiol 2006;61(4):948-59.
27. Jonsson TJ, Murray MS, Johnson LC, Poole LB, Lowther WT. Structural basis for the retroreduction of inactivated peroxiredoxins by human sulfiredoxin. Biochemistry 2005;44(24):8634-42.
28. Vedadi M, Lew J, Artz J, Amani M, Zhao Y, Dong AP, Wasney GA, Gao M, Hills T, Brokx S and others. Genome-scale protein expression and structural biology of *Plasmodium falciparum* and related *Apicomplexan* organisms. Molecular and Biochemical Parasitology 2007;151(1):100-110.
29. Parsonage D, Youngblood DS, Sarma GN, Wood ZA, Karplus PA, Poole LB. Analysis of the link between enzymatic activity and oligomeric state in AhpC, a bacterial peroxiredoxin. Biochemistry 2005;44(31):10583-92.
30. Wood ZA, Poole LB, Hantgan RR, Karplus PA. Dimers to doughnuts: redox-sensitive oligomerization of 2-cysteine peroxiredoxins. Biochemistry 2002;41(17):5493-504.
31. Wood ZA, Poole LB, Karplus PA. Peroxiredoxin evolution and the regulation of hydrogen peroxide signaling. Science 2003;300(5619):650-3.
32. Kitano K, Kita A, Hakoshima T, Niimura Y, Miki K. Crystal structure of decameric peroxiredoxin (AhpC) from *Amphibacillus xylanus*. Proteins 2005;59(3):644-7.
33. Guimaraes BG, Souchon H, Honore N, Saint-Joanis B, Brosch R, Shepard W, Cole ST, Alzari PM. Structure and mechanism of the alkyl hydroperoxidase AhpC, a key element

of the *Mycobacterium tuberculosis* defense system against oxidative stress. J Biol Chem 2005;280(27):25735-42.
34. Choi HJ, Kang SW, Yang CH, Rhee SG, Ryu SE. Crystal structure of a novel human peroxidase enzyme at 2.0 A resolution. Nat Struct Biol 1998;5(5):400-6.
35. Nakamura T, Yamamoto T, Inoue T, Matsumura H, Kobayashi A, Hagihara Y, Uegaki K, Ataka M, Kai Y, Ishikawa K. Crystal structure of thioredoxin peroxidase from aerobic hyperthermophilic archaeon *Aeropyrum pernix K1*. Proteins 2006;62(3):822-6.
36. Mizohata E, Sakai H, Fusatomi E, Terada T, Murayama K, Shirouzu M, Yokoyama S. Crystal structure of an archaeal peroxiredoxin from the aerobic hyperthermophilic crenarchaeon *Aeropyrum pernix K1*. J Mol Biol 2005;354(2):317-29.
37. Declercq JP, Evrard C, Clippe A, Stricht DV, Bernard A, Knoops B. Crystal structure of human peroxiredoxin 5, a novel type of mammalian peroxiredoxin at 1.5 A resolution. J Mol Biol 2001;311(4):751-9.
38. Evrard C, Smeets A, Knoops B, Declercq JP. Crystal structure of the C47S mutant of human peroxiredoxin 5. Journal of Chemical Crystallography 2004;34(8):553-558.
39. Echalier A, Trivelli X, Corbier C, Rouhier N, Walker O, Tsan P, Jacquot JP, Aubry A, Krimm I, Lancelin JM. Crystal structure and solution NMR dynamics of a D (type II) peroxiredoxin glutaredoxin and thioredoxin dependent: a new insight into the peroxiredoxin oligomerism. Biochemistry 2005;44(6):1755-67.
40. Kim SJ, Woo JR, Hwang YS, Jeong DG, Shin DH, Kim K, Ryu SE. The tetrameric structure of *Haemophilus influenza* hybrid Prx5 reveals interactions between electron donor and acceptor proteins. J Biol Chem 2003;278(12):10790-8.
41. Choi J, Choi S, Cha MK, Kim IH, Shin W. 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 2003;278(49):49478-86.
42. Rho BS, Hung LW, Holton JM, Vigil D, Kim SI, Park MS, Terwilliger TC, Pedelacq JD. Functional and structural characterization of a thiol peroxidase from *Mycobacterium tuberculosis*. J Mol Biol 2006;361(5):850-63.
43. Stehr M, Hecht HJ, Jager T, Flohe L, Singh M. Structure of the inactive variant C60S of *Mycobacterium tuberculosis* thiol peroxidase. Acta Crystallogr D Biol Crystallogr 2006;62(Pt 5):563-7.
44. Choi J, Choi S, Chon JK, Cha MK, Kim IH, Shin W. Crystal structure of the C107S/C112S mutant of yeast nuclear 2-Cys peroxiredoxin. Proteins 2005;61(4):1146-9.
45. Li S, Peterson NA, Kim MY, Kim CY, Hung LW, Yu M, Lekin T, Segelke BW, Lott JS, Baker EN. Crystal Structure of AhpE from *Mycobacterium tuberculosis*, a 1-Cys peroxiredoxin. J Mol Biol 2005;346(4):1035-46.

Supplemental Figure Legends

Figure S1. Multiple sequence alignment of representative Prxs shows alignment of some key residues, and the inconsistent location of C_R. Key residues used to create the functional site profiles are starred and the location of residues found in the resulting functional site profiles are labeled with the blue rectangles. The subfamily assignments for each Prx are in parentheses after the protein name. The location of the resolving cysteine (C_R) for a subfamily is highlighted in green for typical 2-Cys (*Salmonella typhimurium* AhpC, *Trypanosoma cruzi* tryparedoxin peroxidase), atypical 2-Cys (*Homo sapiens* Prx5, *Mycobacterium tuberculosis* Tpx, *Aeropyrum pernix* BCP), and 1-Cys (*H. sapiens* Prx6, *M. tuberculosis* AhpE) Prxs. Residues conserved across the entire alignment are highlighted in red and residues identified as conserved in this study in each subfamily are highlighted in yellow. Sequences were aligned using T-coffee [80] and the figure was created using ESPript [49].

Figure S2. **Summary of the DASP process for searching sequence databases.** Functional site profiles (also called active site profiles) are generated as described in detail elsewhere [7], and all motifs (of at least 3 residues in length) for the profile are identified. For each motif, a multiple sequence alignment is created and a position specific scoring matrix (PSSM) is calculated [2]. For each sequence in the database, the PSSM for each motif is used to find the best match, and a p-value score is calculated for each PSSM (representing the probability of finding a similar match in a random sequence) [3]. The p-values from each motif are normalized and then combined using QFAST [4] to give a final p-value, which represents the overall profile to sequence score (more details of this process are described in Supplemental Methods and published elsewhere [1]).

Figure S3. The first two structurally characterized BCP/PrxQ subfamily members are not representative of the entire subfamily and engineered signatures can be used to create a more robust and specific BCP/PrxQ profile. (A) The functional site signatures obtained from the Genbank(nr) search for the BCP/PrxQ subfamily members using the engineered profile were clustered in Matlab, a cluster cutoff was identified (blue line in the dendrogram) and the

subfamily was subdivided into eight groups. Structural representatives and biochemically characterized BCP/PrxQ proteins are listed to the right of the group to which they belong. The GenBank(nr) database was searched using **(B)** the original functional site profile developed for *Aeropyrum pernix* (2cx4, 2cx3) and *Saccharomyces cerevisiae* (2a4v) BCP or **(C)** the engineered profile reflecting the functional site sequences of *A. pernix* BCP, *S. cerevisiae* BCP, *Escherichia coli* BCP, *Helicobacter pylori* BCP, *and Poplar denticolas* PrxQ using a p-value cutoff of 10^{-8} as described in Material and Methods. The results from the **(D)** original BCP/PrxQ search and the **(E)** engineered BCP/Prx search were analyzed to determine whether sequences were specific for that subfamily (dark gray bars), found in the AhpC/Prx1 subfamily (white bars) or Prx6 subfamily (hashed bars) with a more significant p-value, or contained no Prx motif (black bars). The p-value distribution for sequences returned from GenBank(nr) using the engineered profile (shown in E) is more representative of the results from other Prx subfamilies than the p-value distribution from a search using the original BCP/PrxQ functional site profile (shown in D).

Table SI. Prx structures and residues used to create functional site signatures

PDB	Name	Species	Chain	Key Residues	redox state	Ref
AhpC/Prx1 (customarily Typical 2-cys Prxs with both A & B interfaces)						
1e2y[1]	tryparedoxin peroxidase	*Crithidia fasciculata*	B	P45, T49, C52, W87	SH	[20]
1qmv[1]	Prx2	*Homo sapiens*	A	P44, T48, C51, W86	SO	[21]
1qq2[1]	Prx1	*Rattus norvegicus*	A	P45, T49, C52, W87	SS	[22]
1uul[1]	tryparedoxin peroxidase	*Trypanosoma cruzi*	A	P45, T49, C52, W87	SH	[23]
1zof[1]	AhpC	*Helicobacter pylori*	A	P42, T46, C49, W84	S-S	[24]
1zye[1]	Prx 3	*Bos taurus*	A	P40, T44, C47, W82	SH	[25]
2h01[1,2]	thiol peroxidase 1	*Plasmodium yoelli*	A	P43, T47, C50, W85	SH	
2c0d[1]	Mitochondrial 2-Cys Prx	*Plasmodium falciparum*	A	P60, T64, C67, W102	S-S	[26]
2pn8[1,2]	Prx4	*Homo sapiens*	A	P124, T121, C117, W159	SH	
2rii[1]	Prx1 (Complex with Srx)	*Homo sapiens*	A	P45, T49, C52, W87	SH	[27]
2h66[1]	2-Cys	*Plasmodium vivax*	B	P43, T47, C50, W85	SH	[28]
2i81[1,2]	Prx5	*Plasmodium vivax*	C	P43, T47, C50, W85	SH	
1yep[3]	AhpC	*Salmonella typhimurium*	A	P39, T43, C46, W81	S-S	[29,30]
1n8j[3]	AhpC	*Salmonella typhimurium*	A	P39, T43, C46S, W81	Ser	[31]
1we0[3]	AhpC	*Amphibacillus xylanus*	A	P40, T44, C47, W82	S-S	[32]
2bmx[3]	AhpC	*Mycobacterium tuberculosis*	A	P54, T58, C61, W96	S-S	[33]
Prx6 (Customarily 1-Cys Prxs with a B-type interface)						
1prx	Prx6	*Homo sapiens*	A	P40, T44, C47, W82	SOH	[34]
1xcc	1-Cys Prx	*Plasmodium yoelli*	A	P40, T44, C47, W82	SH	[28]
1x0r	thioredoxin peroxidase	*Aeropyrum pernix*	A	P43, T47, C50, W85	SH & S-S	[35]
2cv4	thioredoxin peroxidase	*Aeropyrum pernix*	A	P43, T47, C50, W85	SO_3	[36]
Prx5 (Includes both 1-Cys and atypical 2-Cys Prxs that have an A-type interface)						
1hd2	Prx5	*Homo sapiens*	A	P40, T44, C47, W84	SH	[37]
1urm	Prx5	*Homo sapiens*	A	P40, T44, C47S, W84	Ser	[38]
1tp9	PrxD	*Populus tremula*	A	P44, T48, C51, W88	SH	[39]
1nm3	PrxV	*Haemophilus influenzae*	A	P42, T46, C49, W86	SH	[40]
1xiy	pfAOP	*Plasmodium falciparum*	A	P52, T56, C59, W97	SO_3	[17]
Tpx (customarily atypical 2-Cys Prxs with an A-type interface)						
1psq[2]	Thiol peroxidase	*Streptococcus pneumoniae*	A	P51, T55, C58, W91	SH	
1q98[2]	thiol peroxidase	*Haemophilus influenzae*	A	P52, T56, C59, W92	S-S	
1qxh	thiol peroxidase	*Escherichia coli*	A	P54, T58, C61, W94	S-S	[41]

1xvq	Tpx	*Mycobacterium tuberculosis*	A	P53, T57, C60, W92	S-S	[42]
1y25	Tpx	*Mycobacterium tuberculosis*	A	P53, T57, C60S, W92	Ser	[43]
2yzh[2]	Thiol peroxidase	*Aquifex aeolicus*	A	P54, T58, C61, W94	SH	
BCP/PrxQ (Includes both atypical 2-Cys and 1-Cys Prxs that are monomeric)						
2a4v	BCP	*Saccharomyces cerevisiae*	A	P100, T104, C107S, W141	Ser	[44]
2cx4[2]	BCP	*Aeropyrum pernix*	D	P42, T46, C49, W84	SH & S-S	
AhpE						
1xvw	AhpE	*Mycobacterium tuberculosis*	A	P38, T42, C45, W80	SOH	[45]
1xxu	AhpE	*Mycobacterium tuberculosis*	A	P38, T42, C45, W80	SH	[45]

[1]Used to create Prx1 profile
[2]Not published. Coordinates available in RCSB Protein Database
[3]Used to create AhpC profile

Table SIIA. Test Prx Proteins (Default PSI-BLAST Parameters)

GI#	literature subfamily	test protein name	species	Reference (pubmed id)	DASP p-value				PSI-BLAST e-value											
					Prx5	BCP	Tpx	Prx6	AhpC/Prx1	Prx5_1xiy	Prx5_1hd2	BCP_2cx3	BCP_2a4v	Tpx_1psq	Tpx_1xvq	Prx6_1xcc	Prx6_2cv4	A/Prx1_1qmv	A/Prx1_1yep
24379223	AhpC/Prx1	AhpC	Steptococcus mutans						**2.8834E-14**	3E-16	5E-21	7E-36	9E-22	6E-25	3E-27	3E-32	1E-37	**4E-33**	**5E-46**
5802974	AhpC/Prx1	Prx3	Homo sapiens	12033427					**0**	3E-20	3E-30	3E-47	3E-34	5E-39	6E-41	2E-43	3E-53	**4E-53**	**4E-49**
13265490	AhpC/Prx1	2-cys PrxB	Arabidopsis thaliana	15890615					**0**	2E-17	6E-32	2E-51	1E-41	4E-37	1E-38	9E-47	6E-51	**2E-51**	**9E-53**
11558242	AhpC/Prx1	0	Phaseolus vulgaris	12033427					**0**	4E-17	1E-32	6E-53	2E-42	3E-38	7E-39	8E-47	2E-51	**1E-51**	**1E-52**
6320661	AhpC/Prx1	TSA2	Saccharomyces cerevisiae	10681558					**2.3363E-15**	2E-10	1E-24	3E-40	2E-28	1E-33	7E-35	3E-36	4E-44	**3E-48**	**7E-42**
6323613	AhpC/Prx1	TSA1	Saccharomyces cerevisiae	10681558					**5.3871E-17**	3E-11	5E-27	9E-43	1E-30	7E-36	9E-38	2E-37	1E-45	**3E-49**	**2E-43**
67464646	AhpC/Prx1	AhpC	Entamoeba histolytica	9378375					**4.1914E-14**	4E-14	3E-28	5E-43	7E-33	5E-32	6E-34	2E-38	8E-46	**9E-46**	**5E-45**
15229806	AhpC/Prx1	2-cys PrxA	Arabidopsis thaliana	15890615					**0**	2E-16	6E-31	4E-51	5E-41	5E-37	5E-38	5E-46	9E-51	**4E-51**	**3E-52**
81301118	AhpC/Prx1	typ 2-Cys	Synechococcus elongates PCC 7	16214169					**0**	2E-15	5E-31	2E-50	1E-40	5E-39	3E-40	1E-44	6E-52	**4E-52**	**3E-52**
5163492	AhpC/Prx1	Prx2	Schistosoma mansoni	15075328					**7.1735E-19**	1E-12	4E-29	5E-45	3E-32						
17157991	AhpC/Prx1	jfrac1 (prx1)	Drosophila melanogaster	12033427					**0**	3E-15	1E-27	1E-46	1E-35	2E-34	5E-37	1E-40	5E-51	**4E-53**	**1E-46**
42540580	AhpC/Prx1	2-Cys	Taiwanofungus camphoratus	17031636					**0**	1E-14	6E-27	1E-45	2E-34	1E-36	1E-33	1E-29	4E-38	**1E-36**	**3E-38**
131774	AhpC/Prx1	Cp20	Clostridium pasteurianum	11827546					**0**	3E-14	6E-27	2E-45	7E-37	3E-35	1E-33	1E-33	2E-37	**3E-36**	**5E-36**
7963723	AhpC/Prx1	nkef1	Oncorhynchus mykiss	12033427					**0**	5E-14	2E-28	2E-45	1E-35	9E-36	5E-37	9E-44	3E-54	**3E-55**	**5E-49**
15899339	BCP/PrxQ	BCP-4	Sulfolobus solfataricus P2	18355320		**3E-11**				1E-21	1E-23	**8E-47**	**7E-39**	2E-31	8E-34	9E-22	2E-26	1E-17	2E-21
6322180	BCP/PrxQ	yPrxQ DOT5	Saccharomyces cerevisiae	14640681		**0**				9E-12	1E-13	**9E-29**	**2E-50**	3E-17	2E-16	8E-10	6E-18	1E-11	2E-13
21674812	BCP/PrxQ	group1	Chlorobium tepidum	15518547		**7E-10**				1E-12	1E-21	**5E-42**	**3E-40**	4E-26	5E-28	2E-19	5E-26	8E-19	1E-23
14324635	BCP/PrxQ	group1	Thermoplasma Volcanium	15518547		**2E-12**				6E-15	1E-23	**3E-48**	**9E-53**	2E-34	1E-35	1E-22	5E-31	1E-24	1E-29
15899027	BCP/PrxQ	BCP-3	Sulfolobus solfataricus P2	18355320		**2E-11**				1E-12	7E-20	**2E-40**	**6E-40**	4E-29	1E-28	2E-23	5E-31	1E-24	3E-29
15803003	BCP/PrxQ	BCP	Escherichia coli	16214169		**0**				4E-12	1E-24	**1E-42**	**1E-51**	5E-28	3E-29	1E-23	2E-27	4E-20	6E-22
15230982	BCP/PrxQ	PrxQ	Arabidopsis thaliana	15890615		**5E-14**				5E-14	2E-25	**5E-47**	**3E-48**	2E-32	6E-31	1E-23	1E-32	5E-22	1E-30
17230675	BCP/PrxQ	PrxQ-A	Nostoc sp. PCC 7120 (Anabaena	17210455		**1E-11**				6E-12	3E-24	**6E-46**	**4E-50**	3E-32	2E-32	7E-26	2E-32	4E-22	2E-29
75336180	BCP/PrxQ	PrxQ	Sedum lineare	10998352		**2E-15**				4E-11	6E-23	**4E-45**	**7E-46**	7E-30	1E-29	6E-22	2E-30	1E-19	6E-29
15898858	BCP/PrxQ	BCP-1	Sulfolobus solfataricus P2	18355320		**5E-14**				2E-11	1E-24	**8E-45**	**5E-49**	6E-30	2E-30	1E-23	6E-31	2E-19	1E-26
15613511	BCP/PrxQ	group1	Ba. Halodurans	15518547		**5E-14**				4E-09	6E-23	**7E-50**	**5E-54**	5E-35	2E-35	4E-26	2E-31	4E-20	3E-27
15641362	Prx5	type II	Vibrio cholerae O1 biovar eltor	16214169	**5E-20**					**3E-43**	**9E-33**						2E-14		
15218877	Prx5	PrxIIB	Arabidopsis thaliana	15890615	**3E-22**					**5E-40**	**2E-31**								
15218876	Prx5	PrxIIC	Arabidopsis thaliana	15890615	**3E-22**					**3E-40**	**7E-32**								
21230504	Prx5	group 3	Xanthomonas campestris, gi21	15518547	**0**					**1E-41**	**2E-31**					2E-11	5E-14		
77464478	Prx5	type II PrxT	Rhodobacter sphaeroides	17644813	**6E-19**					**2E-38**	**1E-31**		1E-10			5E-10	2E-12		
17229033	Prx5	group3	Nostoc sp. PCC 7120	15518547	**0**					**4E-50**	**1E-32**					7E-11	1E-14		9E-14
161869858	Prx5	type2 prx-grx fusion	Neisseria meningitidis	14596930	**0**					**1E-47**	**1E-31**					5E-12	2E-15		9E-13
18397457	Prx5	PrxIIF	Arabidopsis thaliana	15890615	**9E-11**					**1E-35**	**1E-25**								
6323138	Prx5	Ahp1p (D)	Saccharomyces cerevisiae	14640681						**1E-28**	**1E-13**								
118721272	Prx5	PrxII F	Pisum sativum	17881238	**7E-09**					**2E-15**									
28393058	Prx6	1-Cys Prx	Arabidopsis thaliana	15890615				**0**		4E-13	4E-28	3E-40	2E-23	3E-26	9E-25	**3E-44**	**2E-46**	4E-32	2E-31
14591037	Prx6	archeal	Pyrococcus horikoshii OT3	16214169				**0**		5E-16	1E-29	8E-41	1E-28	6E-29	9E-28	**2E-37**	**2E-53**	4E-31	8E-31
15898905	Prx6	BCP-2	Sulfolobus solfataricus P2	16441659				**0**		3E-19	3E-29	9E-42	2E-27	1E-30	7E-30	**8E-36**	**2E-55**	9E-35	1E-37
15643570	Prx6	archeal	Thermotoga maritima MSB8	16214169				**0**		4E-19	7E-29	3E-45	9E-31	2E-35	2E-34	**2E-42**	**5E-62**	8E-34	2E-36
20806791	Prx6	archeal	Thermoanaerobacter tengcong	16214169				**0**		2E-19	1E-29	9E-42	2E-28	8E-32	4E-32	**3E-35**	**3E-50**	3E-30	9E-33
45358737	Prx6	archeal	Methanococcus maripaludis S2	16214169				**0**		2E-15	1E-27	5E-42	9E-27	2E-28	3E-28	**2E-37**	**1E-53**	6E-29	4E-32
3334372	Prx6	archeal	Sulfolobus metallicus	16214169				**0**		4E-14	2E-24	1E-33	9E-25	1E-23	1E-20	**2E-31**	**3E-42**	8E-30	1E-30
81301258	Prx6	1-Cys	Synechococcus elongates PCC 7	16214169				**0**		2E-12	2E-29	6E-46	1E-25	4E-30	7E-33	**3E-49**	**1E-53**	3E-36	2E-34
68348727	Prx6	Prx6	Arenicola marina	18359859				**0**		9E-09	2E-29	2E-40	6E-20	5E-26	8E-26	**2E-49**	**9E-49**	3E-33	7E-31
28210605	Prx6	archeal	Clostridium tetani E88	16214169				**0**		5E-14	6E-28	8E-46	4E-32	8E-35	6E-35	**8E-42**	**1E-54**	2E-39	2E-41
16081592	Prx6	archeal	Thermoplasma acidophilum DS	16214169				**0**		4E-15	9E-27	2E-39	9E-24	4E-31	1E-29	**2E-30**	**3E-51**	1E-29	7E-33
62005080	Prx6	1-Cys	Taiwanofungus camphoratus	17103164				**0**		6E-11	1E-27	3E-43	2E-24	3E-29	3E-29	**5E-45**	**3E-48**	1E-34	4E-32
124512718	Prx6	1-Cys	plasmodium falciparum	17890140				**2.7E-20**		1E-08	2E-25	1E-34	7E-19	5E-23	4E-22	**5E-52**	**8E-40**	2E-29	4E-29
6319407	Prx6	1-cys Prx1 (ymTpx)	Saccharomyces cerevisiae	14640681				**0**		1E-08	4E-25	4E-40	3E-22	1E-26	2E-27	**7E-45**	**2E-47**	8E-36	4E-32
42525530	Prx6	archeal	Treponema denticola ATCC 354	16214169				**0**		1E-11	9E-24	4E-38	2E-23	9E-29	7E-30	**1E-33**	**2E-51**	2E-30	2E-29
161485727	Prx6	archeal	Chlorobium tepidum TLS	16214169				**0**		7E-12	6E-23	2E-37	4E-22	2E-28	7E-27	**7E-35**	**3E-52**	3E-29	7E-29
15011539	Prx6	1-Cys	Toxoplasma gondii	16214169				**1.2E-15**		1E-10	8E-25	1E-35	6E-18	2E-21	6E-23	**3E-45**	**8E-43**	3E-29	2E-27
16080001	Tpx	Tpx	Bacillus subtilis	18588855			**2.7E-15**			2E-15	2E-14	2E-32	1E-17	**2E-44**	**3E-41**	4E-12	1E-15	1E-13	3E-15
29346739	Tpx	group2	Bacteroides thetaiotaomicron V	15518547			**1.5E-21**			1E-13	4E-14	3E-33	3E-16	**7E-40**	**1E-42**	1E-11	3E-16	0.000000003	2E-12
23014942	Tpx	group2	Magnetospirillum magnetotact	15518547			**5.2E-23**			2E-15	7E-14	4E-29	3E-15	**5E-38**	**2E-39**		5E-13		
21283385	Tpx	group2	Staphylococcus aureus	15518547			**1.3E-14**			4E-14	5E-13	1E-27	3E-15	**3E-39**	**4E-37**			9E-10	2E-11
29377396	Tpx	Tpx	Enterococcus faecalis	17971082			**5.2E-09**			8E-11	2E-12	3E-32	3E-20	**1E-35**	**2E-32**	1E-08	3E-15	3E-10	6E-14
15792117	Tpx	Tpx	Campylobacter jejuni	18515414			**8.5E-14**			2E-13		1E-23	3E-11	**5E-29**	**1E-29**				

Table SIIB. Test Prx Proteins (Stringent PSI-BLAST Parameters)

GI#	literature subfamily	test protein name	species	Reference (pubmed id)	DASP p-value					PSI-BLAST e-value									
					Prx5	BCP	Tpx	Prx6	AhpC/Prx1	Prx5_1xiy	Prx5_1hd2	BCP_2cx3	BCP_2a4v	Tpx_1psq	Tpx_1xvq	Prx6_1xcc	Prx6_2cv4	A/Prx1_1qmv	A/Prx1_1yep
24379223	AhpC/Prx1	AhpC	Steptococcus mutans	10656297					**2.883E-14**		0.004	3E-10		0.21		6E-16	1E-27	**2E-65**	**1E-84**
5802974	AhpC/Prx1	Prx3	Homo sapiens	12033427					**0**		0.017	1E-18		1E-15	1E-13	2E-32	1E-58	**8E-96**	**1E-91**
13265490	AhpC/Prx1	2-cys PrxB	Arabidopsis thaliana	15890615					**0**		0.19	1E-21	0.00001	6E-14	3E-12	5E-38	4E-60	**5E-90**	**3E-89**
11558242	AhpC/Prx1		Phaseolus vulgaris	12033427					**0**		0.42	6E-21	0.002	3E-14	2E-11	9E-39	4E-61	**5E-90**	**2E-89**
6320661	AhpC/Prx1	TSA2	Saccharomyces cerevisiae	10681558					**2.336E-15**		0.56	3E-13		2E-10	4E-08	4E-25	2E-44	**2E-82**	**5E-76**
6323613	AhpC/Prx1	TSA1	Saccharomyces cerevisiae	10681558					**5.387E-17**		0.91	7E-15		2E-11	3E-09	9E-27	6E-48	**9E-86**	**9E-80**
67464646	AhpC/Prx1	AhpC	Entamoeba histolytica	9378375					**4.191E-14**		1.3	3E-18	0.003	2E-11	5E-09	1E-35	2E-50	**4E-77**	**3E-71**
15229806	AhpC/Prx1	2-cys PrxA	Arabidopsis thaliana	15890615					**0**		1.8	2E-22	0.00003	1E-14	3E-12	3E-38	1E-59	**1E-88**	**1E-88**
81301118	AhpC/Prx1	typ 2-Cys	Synechococcus elongates PCC	16214169					**0**			5E-22	0.0004	9E-11	4E-10	2E-37	3E-60	**1E-89**	**5E-89**
5163492	AhpC/Prx1	Prx2	Schistosoma mansoni	15075328					**7.174E-19**										
17157991	AhpC/Prx1	jfrac1 (prx1)	Drosophila melanogaster	12033427					**0**			8E-14		1E-08	0.00003	1E-38	4E-68	**6E-94**	**5E-87**
42540580	AhpC/Prx1	2-Cys	Taiwanofungus camphoratus	17031636					**0**			1E-16		3E-10	6E-09	8E-27	3E-48	**2E-73**	**2E-75**
131774	AhpC/Prx1	Cp20	Clostridium pasteurianum	11827546					**0**			3E-16	0.072	3E-13	6E-11	3E-36	5E-52	**9E-69**	**4E-69**
7963723	AhpC/Prx1	nkef1	Oncorhynchus mykiss	12033427					**0**	5E-14	2E-28	8E-17	0.58	0.000001	0.0002	5E-38	6E-68	**2E-97**	**5E-89**
15899339	BCP/PrxQ	BCP-4	Sulfolobus solfataricus P2	18355320		**3E-11**					0.0002	**1E-59**	**5E-13**	7E-22	2E-22	6E-14	7E-19	1E-21	4E-23
6322180	BCP/PrxQ	yPrxQ DOT5	Saccharomyces cerevisiae	14640681		**0**					0.001	**1E-15**	**6E-69**	2	5.1	0.0002	0.00006	0.00001	
21674812	BCP/PrxQ	group1	Chlorobium tepidum	15518547		**6.6E-10**					0.008	**6E-27**	**8E-14**	2E-11	3E-11	7E-12	8E-16	1E-19	3E-18
14324635	BCP/PrxQ	group1	Thermoplasma Volcanium	15518547		**2.1E-12**					0.018	**4E-29**	**5E-19**	1E-16	4E-15	1E-15	5E-21	3E-24	1E-24
15899027	BCP/PrxQ	BCP-3	Sulfolobus solfataricus P2	18355320		**2.2E-11**					0.021	**2E-20**	**3E-10**	7E-10	6E-09	9E-24	4E-31	1E-28	3E-28
15803003	BCP/PrxQ	BCP	Escherichia coli	16214169		**0**					0.023	**6E-23**	**3E-19**	1E-10	1E-10	0.000006	5E-08	1E-12	4E-13
15230982	BCP/PrxQ	PrxQ	Arabidopsis thaliana	15890615		**4.8E-14**					0.031	**1E-24**	**8E-19**	0.000003	4E-09	7E-09	9E-17	5E-21	2E-22
17230675	BCP/PrxQ	PrxQ-A	Nostoc sp. PCC 7120 (Anabaen	17210455		**1.2E-11**					0.031	**2E-21**	**6E-18**	6E-10	2E-09	1E-13	3E-17	4E-19	5E-19
75336180	BCP/PrxQ	PrxQ	Sedum lineare	10998352		**1.5E-15**					0.14	**2E-23**	**4E-16**	0.0003	0.000003	4E-09	2E-16	8E-19	4E-20
15898858	BCP/PrxQ	BCP-1	Sulfolobus solfataricus P2	18355320		**4.8E-14**					0.18	**1E-20**	**2E-22**	2E-09	2E-08	3E-14	4E-20	3E-19	2E-19
15613511	BCP/PrxQ	group1	Ba. Halodurans	15518547		**5.2E-14**						**4E-22**	**4E-15**	9E-09	5E-08	1E-14	1E-19	2E-21	6E-22
15641362	Prx5	type II	Vibrio cholerae O1 biovar eltor	16214169	**5.44E-20**					**1E-19**	**1E-56**		0.041	0.019	0.001	0.00007	0.000004		
15218877	Prx5	PrxIIB	Arabidopsis thaliana	15890615	**3.17E-22**					**1E-14**	**5E-48**		0.03	0.099	0.25	0.004	0.000005		
15218876	Prx5	PrxIIC	Arabidopsis thaliana	15890615	**3.23E-22**					**2E-14**	**6E-48**		0.048	0.069	0.15	0.007	0.00002		
21230504	Prx5	group 3	Xanthomonas campestris, gi21	15518547	**0**					**1E-17**	**5E-47**	0.00002	0.004	2.3	6.4	0.0005	0.009	0.00001	
77464478	Prx5	type II PrxT	Rhodobacter sphaeroides	17644813	**6.36E-19**					**6E-09**	**8E-45**	0.000001	0.001	0.002	0.00003	0.0003	0.002		0.0002
17229033	Prx5	group3	Nostoc sp. PCC 7120	15518547	**0**					**6E-28**	**2E-44**				1.6		0.004		0.00004
161869858	Prx5	type2 prx-grx fusion	Neisseria meningitidis	14596930	**0**					**5E-28**	**2E-40**			0.008	0.005	0.26	0.005	0.00001	0.000008
18397457	Prx5	PrxIIF	Arabidopsis thaliana	15890615	**9.11E-11**					**3E-12**	**9E-38**			0.006	0.003				
6323138	Prx5	Ahp1p (D)	Saccharomyces cerevisiae	14640681						**4E-11**	**2E-24**								
118721272	Prx5	PrxII F	Pisum sativum	17881238	**7.22E-09**					**1E-08**	**4E-18**								
28393058	Prx6	1-Cys Prx	Arabidopsis thaliana	15890615				**0**		0.026	0.0003	5E-09	0.00004			**3E-87**	**1E-81**	5E-35	2E-26
14591037	Prx6	archeal	Pyrococcus horikoshii OT3	16214169				**0**			0.0006	1E-12	0.00004	0.003	0.0003	**4E-66**	**2E-89**	1E-44	1E-30
15898905	Prx6	BCP-2	Sulfolobus solfataricus P2	16441659				**0**			0.0007	2E-14	0.000003			**3E-71**	**4E-92**	3E-50	3E-37
15643570	Prx6	archeal	Thermotoga maritima MSB8	16214169				**0**		3.3	0.002	2E-15	0.0001	0.00001	0.0008	**7E-73**	**5E-100**	7E-53	2E-40
20806791	Prx6	archeal	Thermoanaerobacter tengcong	16214169				**0**			0.003	7E-13	0.009	0.0005		**6E-71**	**6E-91**	2E-48	3E-34
45358737	Prx6	archeal	Methanococcus maripaludis S2	16214169				**0**		5.7	0.004	7E-15	0.000002		1.4	**2E-67**	**2E-90**	1E-45	1E-31
3334372	Prx6	archeal	Sulfolobus metallicus	16214169				**0**			0.032	2E-09	3E-07			**7E-61**	**9E-75**	2E-45	5E-36
81301258	Prx6	1-Cys	Synechococcus elongates PCC	16214169				**0**			0.069	1E-11	0.00006			**1E-97**	**3E-96**	7E-40	4E-30
68348727	Prx6	Prx6	Arenicola marina	18359859				**0**			0.59	0.000003	0.024			**2E-97**	**5E-87**	4E-36	1E-26
28210605	Prx6	archeal	Clostridium tetani E88	16214169				**0**		0.49	0.66	5E-16	0.005	0.0002	0.028	**3E-71**	**6E-87**	4E-58	8E-41
16081592	Prx6	archeal	Thermoplasma acidophilum DS	16214169				**0**			0.82	8E-11	0.0005			**8E-65**	**1E-88**	5E-47	1E-33
62005080	Prx6	1-Cys	Taiwanofungus camphoratus	17103164				**0**		6.8		2E-11				**3E-91**	**2E-88**	5E-39	2E-28
124512718	Prx6	1-Cys	plasmodium falciparum	17890140				**2.68E-20**								**3E-88**	**1E-73**	1E-38	2E-30
6319407	Prx6	1-cys Prx1 (ymTpx)	Saccharomyces cerevisiae	14640681				**0**				6E-10	1.6	0.52		**5E-91**	**2E-89**	3E-43	7E-32
42525530	Prx6	archeal	Treponema denticola ATCC 354	16214169				**0**				8E-11	0.01	0.054		**3E-71**	**2E-92**	3E-42	1E-29
161485727	Prx6	archeal	Chlorobium tepidum TLS	16214169				**0**				1E-10	0.022	0.004	0.14	**1E-71**	**1E-93**	3E-42	5E-28
15011539	Prx6	1-Cys	Toxoplasma gondii	16214169				**1.18E-15**					0.23			**2E-85**	**2E-73**	2E-35	2E-26
16080001	Tpx	Tpx	Bacillus subtilis	18588855			**2.65E-15**				1.6	2E-13	0.011	**4E-78**	**2E-70**			0.000000008	0.0000001
29346739	Tpx	group2	Bacteroides thetaiotaomicron	15518547			**1.53E-21**					1E-13	0.0007	**1E-70**	**6E-74**	0.38		0.0000001	0.00001
23014942	Tpx	group2	Magnetospirillum magnetotact	15518547			**5.18E-23**					2E-08	0.001	**4E-68**	**5E-71**				
21283385	Tpx	group2	Staphylococcus aureus	15518547			**1.32E-14**					1E-10	0.0003	**2E-73**	**5E-70**			0.00001	0.00004
29377396	Tpx	Tpx	Enterococcus faecalis	17971082			**5.21E-09**					2E-10	0.0008	**6E-58**	**2E-45**		0.00009	0.000000003	2E-11
15792117	Tpx	Tpx	Campylobacter jejuni	18515414			**8.5E-14**					4E-09		**1E-52**	**3E-55**				

Table SIII. Hits with no Prx motif

Accession number	Name	Species	p-value	Missing Residues
BCP/PrxQ (engineered)				
125973977	Redoxin	*Clostridium thermocellum ATCC 27405*	2.24E-13	P replaced with other residue
196253450	Alkyl hydroperoxide reductase/ Thiol specific antioxidant	*Clostridium thermocellum DSM 4150*	1.90E-13	P replaced with other residue
82702925	Alkyl hydroperoxide reductase/ Thiol specific antioxidant	*Nitrosospira multiformis ATCC 25196*	1.42E-14	P replaced with other residue
167755398	hypothetical protein CLORAM_00912	*Clostridium ramosum DSM 1402*	0.00E+00	P replaced with other residue
AhpC/Prx1				
46204795	Peroxiredoxin	*Magnetospirillum magnetotacticum MS-1*	7.97E-22	T replaced with other residue
47193078	unnamed protein product	*Tetraodon nigroviridis*	1.73E-13	P fragment missing
118094466	similar to natural killer cell enhancing factor isoform 2	*Gallus gallus*	3.23E-11	P fragment missing
12718511	peroxiredoxin	*Platichthys flesus*	9.14E-15	Cp fragment missing
158593205	Thioredoxin peroxidase 1, putative	*Brugia malayi*	1.09E-12	P replaced with other residue
116500579	hypothetical protein CC1G_04730	*Coprinopsis cinerea okayama7#130*	5.38E-10	Cp fragment missing
110602165	Alkyl hydroperoxide reductase/ Thiol specific antioxidant	*Geobacter sp. FRC-32*	2.15E-11	Cp fragment missing
114688004	thioredoxin peroxidase	*Pan troglodytes*	1.20E-15	P replaced with other residue
148697774	mCG116719	*Mus musculus*	2.32E-16	P replaced with other residue
73946795	PREDICTED: similar to Peroxiredoxin 2	*Canis familiaris*	7.97E-11	P replaced with other residue
147919347	putative 2-cysteine peroxiredoxin	*uncultured methanogenic archaeon RC-I*	7.09E-09	P replaced with other residue
Prx6				
1710079	REHY_TORRU Probable 1-Cys peroxiredoxin (Rehydrin)	*Syntrichia ruralis*	6.99E-12	T replaced with other residue
119871684	alkyl hydroperoxide reductase/ Thiol specific antioxidant	*Pyrobaculum islandicum DSM 4184*	9.17E-15	T replaced with other residue
163718158	alkyl hydroperoxide reductase/ Thiol specific antioxidant	*Thermoproteus neutrophilus V24Sta*	4.19E-14	T replaced with other residue
119617928	hCG2041492	*Homo sapiens*	1.93E-16	T replaced with other residue
Prx5				
2462742	Unknown protein	*Arabidopsis thaliana*	3.70E-10	Cp P T replaced with other residues
149391021	peroxiredoxin 5	*Oryza sativa (indica cultivar-group)*	2.33E-10	truncation up to T
56182370	putative thioredoxin peroxidase 1	*Saccharum officinarum*	1.28E-11	Cp fragment missing
114638297	similar to antioxidant enzyme B166 isoform 4	*Pan troglodytes*	8.10E-13	Cp fragment missing
115745775	similar to peroxiredoxin V protein	*Strongylocentrotus purpuratus*	4.55E-11	Cp fragment missing
Tpx				
1103833	thiol peroxidase	*Escherichia coli*	0.00E+00	Cp replaced with W

Table SIV. Proteins Assigned by DASP to two Prx subfamilies

GenBank Accession Number	p-value Prx6	p-value AhpE	p-value AhpC/Prx1	p-value BCP/PrxQ
110799231		8.20E-09	8.47E-16	
19357674		2.77E-09	0	
125979671		8.88E-09	0	
149179118		1.94E-09	0	
16331338		1.54E-09	0	
17864676		8.40E-09	0	
18309764		7.52E-09	4.01E-16	
68551025		7.06E-10	0	
91090021		8.81E-09	3.66E-18	
91203633		7.33E-09	5.71E-15	
149278593		2.92E-09		1.07E-09
149918375		2.93E-09		1.31E-09
150020653		8.53E-09		8.53E-09
123437746	3.84E-09		5.99E-21	
123449270	9.77E-10		2.73E-23	
123459140	1.83E-09		7.16E-24	
123974738	1.84E-09		4.95E-23	
146304289	0		8.30E-09	
150400760	0		3.71E-09	
156934873	5.00E-09		0	
163781576	3.83E-09		0	
50083688	9.68E-10		0	
50085223	2.76E-09		0	
78223919	1.37E-10		0	
90416750	3.67E-09		0	
3024730	0			3.61E-10
13472213	0			2.01E-09
14286173	0			9.37E-09
15678187	0			2.71E-09
10281259			0	3.85E-10
13186337			6.91E-09	3.44E-09
163789074			0	5.04E-09
193627310			0	4.10E-09

Scores highlighted in red were the least significant for a given protein. Scores shown in black indicate the subfamily assignment for a given protein.
Results from the final BCP/PrxQ profile are shown; the original BCP/PrxQ search identified the same 11 sequences as crosshits with scores from 10^{-10} - 10^{-9}.

Table SV. Prx subfamily members

Table SV can be downloaded separately as an excel spreadsheet

Table SVI. Conserved residues in each Prx subfamily[a]

	Conserved residues in total sequence	Conserved residues in functional site profile	% total sequence conserved	% conserved residues in functional site profile
BCP/PrxQ	23	12	9%	52%
AhpC/Prx1	33	22	12%	67%
Prx6	42	20	15%	48%
Prx5	46	16	12%	35%
Tpx	39	16	20%	41%

[a]The full sequences of all proteins in each subfamily were aligned using ClustalW and the entropy values were calculated for every position that contained at least 20 sequences. Residues were considered conserved with an entropy value less than the mean minus one standard deviation (0.61).

Figure S1

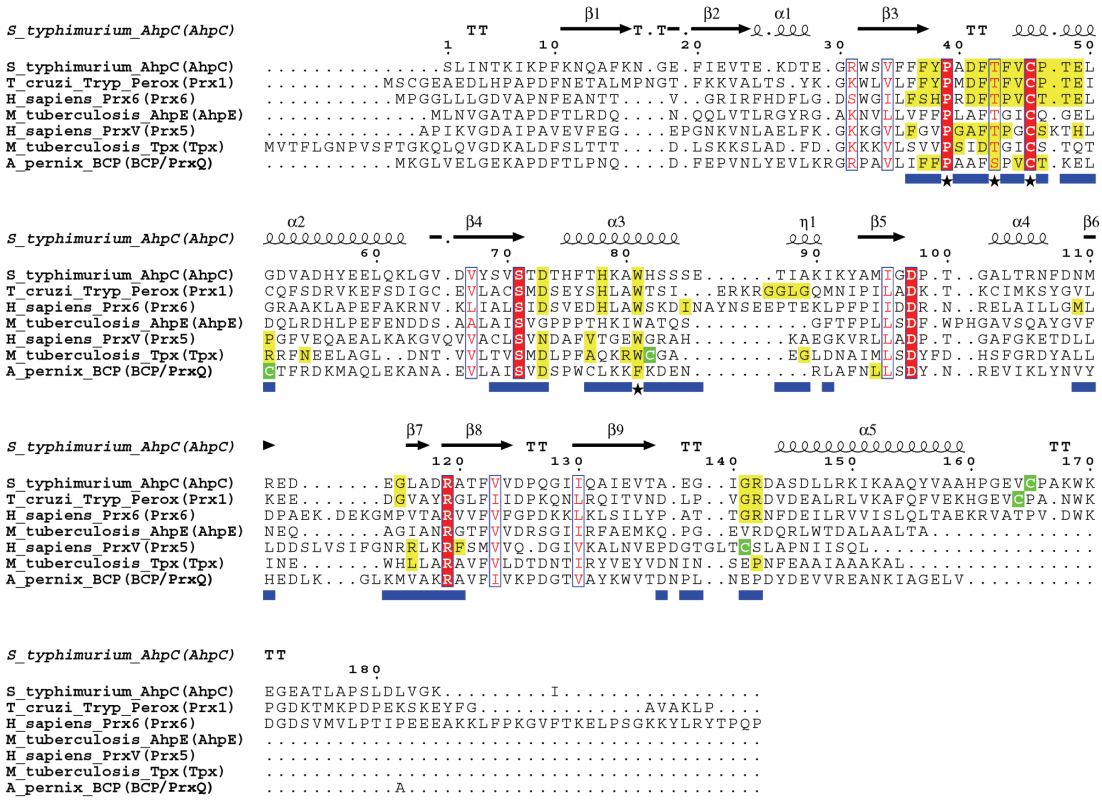

S_typhimurium_AhpC(AhpC)
S_typhimurium_AhpC(AhpC)
T_cruzi_Tryp_Perox(Prx1)
H_sapiens_Prx6(Prx6)
M_tuberculosis_AhpE(AhpE)
H_sapiens_PrxV(Prx5)
M_tuberculosis_Tpx(Tpx)
A_pernix_BCP(BCP/PrxQ)

Figure S2

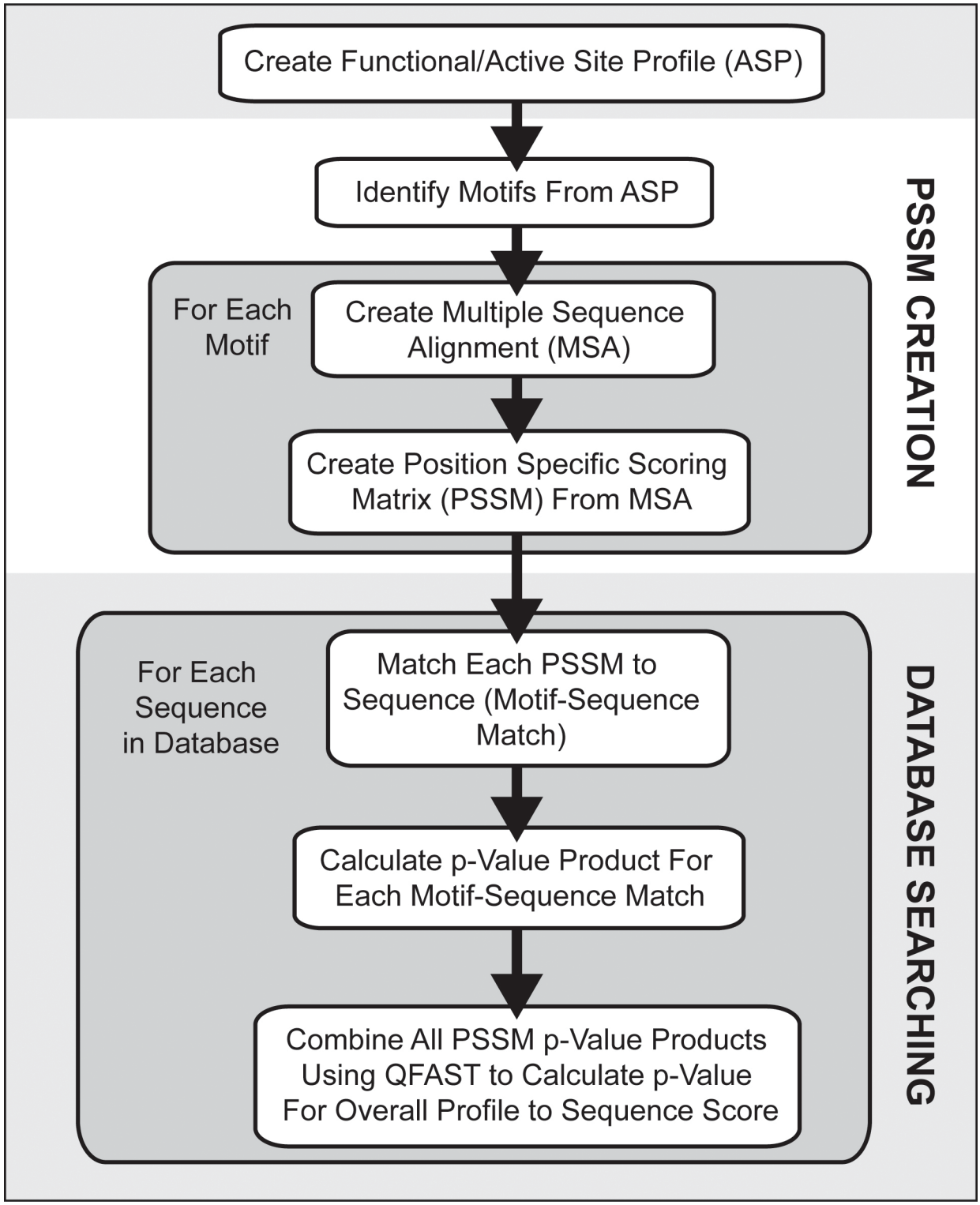
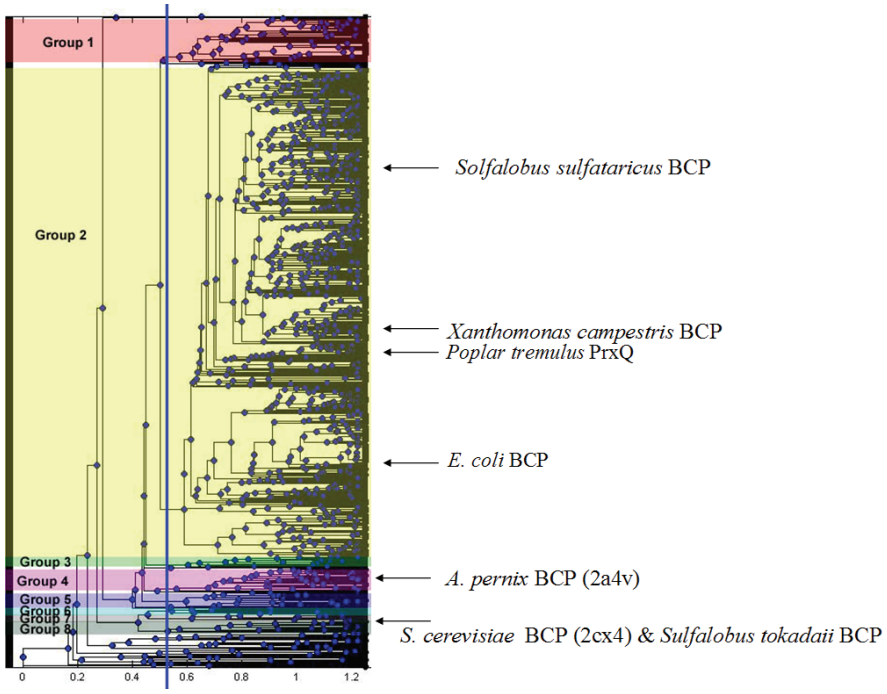

Create Functional/Active Site Profile (ASP)
Identify Motifs From ASP
For Each Motif
Create Multiple Sequence Alignment (MSA)
Create Position Specific Scoring Matrix (PSSM) From MSA
PSSM CREATION
For Each Sequence in Database
Match Each PSSM to Sequence (Motif-Sequence Match)
Calculate p-Value Product For Each Motif-Sequence Match
Combine All PSSM p-Value Products Using QFAST to Calculate p-Value For Overall Profile to Sequence Score
DATABASE SEARCHING

Figure S3

A Dendrogram for BCP subfamily search



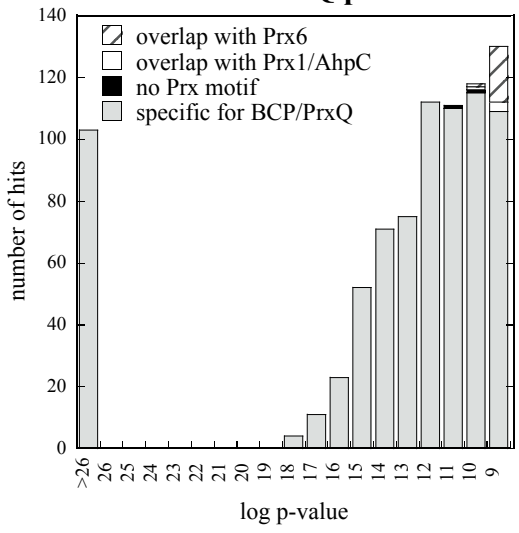
B Active Site Profile for Original BCP/PrxQ subfamily search

```
PDB2A4VA      Efvyprastpgstrqas----GLSADsqkkfqskqnlYl-IaKgsirs-P---
PDB2CX3A      -QffpaafspvctkelctfDmAISVDclkkfk-denLfL-hLlVAKRAnplne
PDB2CX4D      QiffpaafspvctkelC-----ISVDclkkfk-denrlLvHlLvakra-PLe-
                .:* * :* .*:: .     :*.*. ***:.:: * :    : *: *
```

C Active Site Profile for Engineered BCP/PrxQ subfamily search

```
PDB2A4VA      Efvyprastpgctrqac----GLSADsqkkfqskqnlYl-IaKgsirs-P---
E.c. BCP      -QfypkamtpgctvqacglDmGISTDklsrfa-ekeLfLvKmTgihriFKTSN
PDB2CX3A      -QffpaafspvctkelctfDmAISVDclkkfk-denLfL-hLlVAKRAnplne
H.p. BCP      --YFPKDNTPGCTLEAKDFaFavspdSHQKFI-SQClVlAgYeGIIRSvkakg
PDB2CX4D      QiffpaafspvctkelC-----ISVDclkkfk-denrlLvHlLvakra-PLe-
P.j.PRXQ      -QfypadetpgctkqacafDyGISGDshkafa-kkyLfL-aFl-PGRQfqpek
                :*   :* ** :       :* *  . *  .:   *       *
```

D Results from original BCP/PrxQ profile



E Results from "engineered" BCP/PrxQ profile