January 2007

Supporting Online Material for "High-Throughput Identification of Catalytic Redox-Active Cysteine Residues"

Dmitri E. Fomenko
University of Nebraska - Lincoln, dfomenko2@unl.edu

Weibing Xing

Blakely M. Adair

David J. Thomas

Vadim Gladyshev
University of Nebraska - Lincoln, vgladyshev1@unl.edu

www.sciencemag.org/cgi/content/full/315/5810/387/DC1

Supporting Online Material for

High-Throughput Identification of Catalytic Redox-Active Cysteine Residues

Dmitri E. Fomenko, Weibing Xing, Blakely M. Adair, David J. Thomas, Vadim N. Gladyshev*

*To whom correspondence should be addressed. E-mail: vgladyshev1@unl.edu

Published 19 January 2007, *Science* **315**, 387 (2007)
DOI: 10.1126/science.1133114

This PDF file includes:

Supporting On-line Materials

This manuscript has been reviewed in accordance with policy of the National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency, and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

Methods

Sequence databases

The following NCBI sequence databases were used: non-redundant protein database, Sargasso Sea environmental protein database, Sargasso Sea environmental nucleotide sequence database, non-redundant nucleotide sequence database, shotgun sequence database, EST database (all above as of Jan 10, 2005), database of conserved domains (Oct 29, 2005), taxonomy data base and environmental sequence databases with accession numbers AACY00000000 (1) (Dec 23, 2004), AAFX01000000 (Feb 19, 2005), AAFY01000000 (Feb 23, 2005), AAFZ01000000 (Feb 23, 2005) (2), AADL01000000 (May 05, 2004), DU731018-DU796676 and DU800850-DU800864 (Jan 27, 2006) (3). The searches utilized the PrairieFire Beowulf cluster from Research Computing Facility, University of Nebraska – Lincoln.

Identification of redox-active Cys by homology to selenoproteins

To identify Cys-containing homologs of selenoproteins, PSI-blast (4) was used to analyze various collections of protein sequences with the following parameters: expectation value, 10; expectation value for multipass model, 0.1; and number of iterations, 3. This procedure was also used to verify Cys/Sec pairs in homologous sequences (see below). The set of selenoproteins used in this search was compiled manually based on our prior experience with selenoproteome analyses (5-13).

Identification of Cys/Sec pairs

As selenoproteins are incorrectly annotated in sequence databases (mostly due to misannotation of Sec-encoding UGA codons), nucleotide sequence databases were used as the source of potential selenoprotein sequences for identification of Cys/Sec pairs. Environmental, non-redundant nucleotide, shotgun, EST and genomic sequences were translated in 6 frames and analyzed with tblastn against the set of proteins that contained Cys residues. The tblastn output was automatically analyzed using in-house Perl scripts to identify proteins, in which Cys aligned with candidate Sec. Redundancy was eliminated by filtering protein gi-numbers across the taxonomy database. Sequences from the same organism, which exhibited more then 98% identity, were considered identical. Using this procedure, we filtered out multiple DNA sequences, which corresponded to the same protein sequence in the same organism. This procedure also allowed us to divide the proteins into major phyletic groups represented in the NCBI taxonomy database. RPS-Blast (4) against the NCBI database of conserved domains was then used to classify proteins into protein families.

All hits were clustered in pairwise alignments and tested for presence of eukaryotic, bacterial and archaeal SECIS elements using SECISearch tools (12). To exclude sequence errors, we only selected protein families, which were represented by at least three Cys/Sec pairs corresponding to the same location in the alignment. Families with 1 or 2 Cys/Sec pairs were selected only if a high-scoring SECIS element could be identified in the candidate selenoprotein sequence. For environmental sequences, identification of bacterial, archaeal and eukaryotic SECIS elements also allowed us to classify selenoprotein sequences into bacterial, archaeal and eukaryotic sequences. For each selected group of aligned Cys/Sec pairs, the corresponding nucleotide sequences were analyzed in the remaining five open reading frames to exclude the possibility that the correct ORF was in a different reading frame. Sequence alignments were prepared using ClustalW and T-Coffee. Conserved residues were highlighted with BoxShade v3.21.

Statistical analysis of redox-active Cys neighborhoods

Twenty random chosen representative sequences from 27 protein families shown in Table S1 were extracted. The analysis was limited to families of proteins with more then ten amino acids on each side of the redox active Cys/Sec. The 21-amino acid sequences were aligned based on the location of Sec or Cys in the middle and frequencies of amino acids were calculated as the number of times a particular amino acid was observed in each position divided by the total number of proteins in the set.

The secondary structure context of redox-active Cys was estimated for all selected proteins and separately for a subset of non-thioredoxin fold proteins. Secondary structures were first predicted with PSI Pred (14) and then the 21-amino acid sequences with Cys in the middle were extracted. Frequencies of secondary structures were calculated as the number of times alpha helixes, beta strands and loops were present in each position divided by the total number of proteins in set.

AdoMet-dependent methyltransferase

Mouse AS3MT structure was modeled with Modeler8v2 (15) based on the structure of mRNA cap (Guanine-n7) methyltransferase (PDB 1ri1). Mouse AS3MT cDNA was prepared from total mouse RNA using random primers and reverse transcription and subsequent amplification with the following primers: 5'-agatcgtgacatatggctgcttcccgagacgctg-3' and 5'-gcgctggccctcgagctagcagtttttcctcttgccacagcag-3'. The product was ligated into the NdeI/XhoI sites of pET15b. The following primers were used for site-directed mutagenesis: Cys157Ser, 5'-atgatattgtcatatccaactctgttatcaaccttgttcct-3' and 5'-aggaacaaggttgataacagagttggatatgacaatatcat-3'; and Cys207Ser, 5'-agttttatgggggggaatccctgggaggcgctctg-3' and 5'-cagagcgcctcccagggattccccccataaaactttg-3'. Wild type AS3MT protein and Cys to Ser mutants were purified by His-tag affinity chromatography using Talon resin (Clontech) according to the manufacturers' protocol. Recombinant proteins were more than 85% pure based on SDS PAGE analysis.

Activity of AS3MT and its Cys-to-Ser mutants

Reaction mixtures (80 μL final volume) that contained 5 μg of wild type AS3MT or Cys157Ser or Cys207Ser mutant AS3MT, 1 mM AdoMet, 3 μM [^{73}As]-iAsIII, 0.5 μM *E. coli* thioredoxin, 0.26 μM rat liver thioredoxin reductase, 300 μM NADPH in 100 mM Tris/100 mM sodium phosphate buffer, pH 7.4, were incubated at 37 °C for 60 min. Reactions were stopped by addition of 16 μL of 30% H_2O_2 to oxidize and release arsenicals. Aliquots of oxidized reaction mixtures were chromatographed on a PRP-100 anion exchange column to separate iAs, MAs, and DMAs (16). Radiolabeled arsenicals were eluted with 7.5 mM phosphate mobile phase, pH 5.8, at a flow rate of 1.5 mL per minute. A Packard flow scintillation analyzer with a scintillant flow rate of 4.5 mL per minute detected [^{73}As]-iAs^{V}, [^{73}As]- MAs^{V}, and [^{73}As]- DMAs^{V}. Authentic [^{73}As]-iAs^{V}, [^{14}C]-MAs^{V}, and [^{14}C]-DMAs^{V} were used as standards for calibration and quantitation.

Additional predicted thiol-based oxidoreductases not discussed in the main text of the article

Hypothetical protein 1. This protein family includes proteins of unknown function and the Cys/Sec pair was formed with proteins detected in environmental genome sequences (Fig. S16). Hypothetical protein 1 is represented by only two Cys-containing sequences and dominantly exists in the Sec-containing form. Hypothetical protein 1 shows no similarity to proteins with known function. A structural alignment shows low similarity to 3-dehydroquinate dehydratase (PDB 1GQO), which catalyzes dehydration of 3-dehydroquinate to 3-dehydroshikimate in the third step of the shikimate pathway; however, the low level of similarity and the absence of Cys in appropriate dehydroquinate dehydratase positions preclude further functional assignment. Hypothetical protein 1 appears to be abundant in marine bacteria.

Hypothetical protein 2. Phosphodiesterase homologs catalyze the hydrolysis of ribonucleotides, deoxyribonucleotides, and UDP sugars to nucleosides which are then transported into cells. A major function of this periplasmic protein is to salvage nucleotides which can be used as energy or carbon sources. We detected an N-terminal extension in a subset of bacterial proteins of this family (Fig. S17). The N-terminal region contained a CxC motif in which the second Cys was replaced with Sec in 28 environmental sequences. The presence of the N-terminal domain suggests a new function associated with a redox reaction involving an unidentified nucleotide/nucleoside. Regardless of the fact that the identified protein exhibits high sequence similarity to phosphodiesterase, its natural function is likely to be unrelated to polynucleotide hydrolysis; however, it might be linked to a new type of nucleotide modification.

Hypothetical protein 3. This protein shows sequence and structural similarity with a periplasmic vitamin B12 binding protein which functions in a complex with transporter BtuCD and to a periplasmic Fe^{3+} transport component FhuD of ferric enterobactin transport system which delivers ferrichrome from outer membrane FhuA complex to the cytoplasmic membrane FhuB transport system. A small group of bacterial proteins in this family contains an N-terminal CxxC motif (Fig. S18) which is replaced with the CxxU motif in two environmental bacterial sequences. We hypothesize that the CxxC/U motif-containing proteins evolved from metal transporter proteins and could function in transporting an unidentified compound or metal by acting as an oxidoreductase. One possible function is the reduction of Fe^{3+} to Fe^{2+}. A C-terminal

region of the protein may be involved in recognition of the FhuA complex, in iron binding, or in its delivery to a cytoplasmic membrane transport complex. The N-terminal domain may be involved in the reduction of FhuD substrate.

Supporting references

1. S. G. Tringe *et al.*, *Science* **308**, 554 (2005).
2. J. C. Venter *et al.*, *Science* **304**, 66 (2004).
3. E. F. DeLong *et al*., *Science* **311**, 496 (2006).
4. S. F. Altschul *et al*., *Nucleic Acids Res.* **25**, 3389 (1997).
5. A. V. Lobanov, S. Gromer, G. Salinas, V. N. Gladyshev, *Nucleic Acids Res*, **34**, 4012 (2006).
6. A. V. Lobanov *et al.*, *Nucleic Acids Res.*, **34**, 496 (2006).
7. S. Castellano *et al*., *Proc Natl Acad Sci*, **102**, 16188 (2005).
8. K. Taskov *et al*., *Nucleic Acids Research*, **33**, 2227 (2005).
9. Y. Zhang, D. E. Fomenko, V. N. Gladyshev, *Genome Biology*, **6**, R37 (2005).
10. Y. Zhang, V. N. Gladyshev, *Bioinformatics*, **21**, 2580 (2005).
11. G. V. Kryukov, V. N. Gladyshev, *EMBO Rep.*, **5**, 538 (2004).
12. G. V. Kryukov *et al*., *Science*, **300**, 1439 (2003).
13. G. V. Kryukov, V. N. Gladyshev, *Methods Enzymol.*, **347**, 84 (2002).
14. K. Bryson *et al*., *Nucl. Acids Res.* **33**, W36 (2005).
15. M. A. Marti-Renom *et al*., *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291 (2000).
16. X. C. Le, M. Ma, *Anal. Chem.* **70**, 1926 (1998).

Legends to supporting figures

Fig. S1. Dependence of the number of identified proteins containing redox-active Cys residues on the proteome size (number of protein-coding genes) in representative organisms. The data are plotted for indicated eukaryotes (shown in red), bacteria (black) and archaea (blue)**.**

Fig. S2. a) Distribution of Sec-containing proteins in the three domains of life. The total number of Sec-containing proteins/number of corresponding protein families in each domain is indicated. **b) Distribution of identified redox-active Cys-containing proteins in the three domains of life.** The total number of Cys-containing proteins/number of corresponding protein families in each domain is indicated. Diameters of the circles are proportional to the number of corresponding Sec- or Cys-containing protein families.

Fig. S3-S10. Multiple sequence alignments of thioredoxins and thioredoxin-like proteins (S3), glutaredoxins and glutaredoxin-like proteins (S4), peroxiredoxins (S5), glutathione peroxidases (S6), arsenate reductases (S7), HesB-like proteins (S8) and DsrE proteins (S9). The alignments only show the active sites of the enzymes and their flanking regions. Accession numbers (GI-numbers) of the sequences and their origins are shown on the left. Predicted selenocysteines are shown in red and the corresponding cysteines in blue. If predicted, resolving cysteines are shown in green. Conserved residues are highlighted using BoxShade program v3.21.

Fig. S11. Multiple sequence alignment of rhodanese-like proteins. The alignment is limited to the active site of the enzymes and its flanking regions. Predicted Sec are shown in red and the corresponding Cys in blue. This figure shows four different families of rhodanese-like proteins that evolved into selenoproteins, including CD01448 (shown in red), CD00158 (blue), CD01444 (green) and CD01524 (purple). There are additional 24 rhodanese families, for which no evidence of redox function could be obtained.

Fig. S12. Multiple sequence alignment of MoeB proteins and their distant homologs. The alignment shows the active site of the enzyme and its flanking regions. Predicted Sec are shown in red and the corresponding Cys in blue. The alignment includes MoeB-like proteins (CD30111, shown in green), thiamine biosynthesis ThiF proteins (CD10349, in red), and E1-like proteins (CD30117, in blue). Conserved residues are highlighted by BoxShade program v3.21.

Fig. S13-S17. Multiple sequence alignments of heterodisulfide reductases, Hypothetical protein 1 (S14), Hypothetical protein 2 (S15), Hypothetical protein 3 (S16) and OsmC proteins (S17). Predicted Sec are shown in red and the corresponding Cys in blue. Conserved residues are highlighted by BoxShade program v3.21.

Fig. S18-S19. Multiple sequence alignments of methionine-S-sulfoxide reductases (S18) and methionine-R-sulfoxide reductases (S19). Predicted Sec are shown in red and the corresponding Cys in blue. Conserved residues are highlighted by BoxShade program v3.21.

Fig. S20. Distribution of amino acids around redox-active Cys. Frequencies of each of the 20 amino acids in ten positions upstream and ten positions downstream of the predicted redox Cys were determined for 20 representative proteins from each of the protein families identified in the searches. A second Cys was found to often occur in the position to generate CxxC motifs with redox Cys being the first or second Cys in the CxxC motif. Some proteins contained a CxC motif. Glutamic and aspartic acids were not found in positions -3, +1 and -3, -1, +1, +2, respectively, in any identified proteins. Gly was found to be enriched in the positions flanking the redox-active Cys. Most other residues were found to be distributed uniformly.

Fig. S21. Secondary structure context of redox-active Cys. a) Secondary structure context was determined for ten residues upstream and 10 residues downstream of the predicted redox-active Cys in the 10-protein sets representing 27 protein families. **b)** Secondary structure context was determined for 17 non-thioredoxin fold protein families, each represented by 10 proteins. In each case, a beta-strand was the predominant secondary structure element upstream of Cys, an alpha-helix was predominantly downstream of Cys, whereas the Cys itself was most often found in the loop.

Table S1. Proteins identified in searches for Cys/Sec pairs in homologous sequences.

Protein family	Comments	Protein family	Bacterial Cys/Sec sequences	Archaeal Cys/Sec sequences	Eukaryotic Cys/Sec sequences
Functionally characterized proteins containing catalytic redox-active Cys					
Methionine-S-sulfoxide reductase (MsrA)	Reduction of methionine-S-sulfoxides	CDD25795	767/1	11/0	47/14
Methionine-R-sulfoxide reductase (MsrB)	Reduction of methionine-R-sulfoxides	CDD25798	1276/0	7/0	23/37
Animal thioredoxin reductase (TR)	Reduction of thioredoxins, glutaredoxins and some biofactors	CDD10363	0/0	0/0	15/34
Deiodinase (includes thyroid hormone deiodinases 1, 2 and 3 and bacterial homologs)	Reductive deiodination of thyroid hormones (in animals). Unknown function (in bacteria)	CDD1392	23/9	0/0	0/24
Glutathione peroxidase (also includes phospholipid hydroperoxide glutathione peroxidase and other homologs)	Reduction of hydroperoxides	CDD10260 CDD25459	182/9	0/0	162/68
Peroxiredoxin (Prx)	Reduction of hydroperoxides	*1	873/84	64/0	298/0
Proline reductase PrdB	Amino acid metabolism	CDD27462	8/52	0/0	0/0
Thioredoxin (includes protein disulfide isomerases, DsbA, DsbC, DsbG, DsbE and other protein families)	Reduction, formation or isomerization of disulfide bonds	*2	1594/48	57/0	843/17
Glutaredoxin (includes glutaredoxin-like proteins)	Reduction of intramolecular disulfides and mixed disulfide bonds involving glutathione	*3	230/5	18/0	124/0
CMD and AhpD domain-containing proteins	Oxidoreduction	CDD25931 CDD10469 CDD11836	325/37	11/0	2/0
OsmC-like protein	Thiol peroxidase	CDD25924 CDD11475 CDD11476	254/3	11/0	0/0
Formylmethanofuran dehydrogenase	Oxidation of formylmethanofuran	CDD29457	0/0	13/6	0/0
F420-reducing hydrogenase alpha subunit	Hydrogen oxidation or proton reduction	CDD12595	100/3	17/4	0/0
F420-reducing hydrogenase, delta subunit	Hydrogen oxidation or proton reduction	CDD10549	25/4	12/3	0/0
Methylviologen-reducing hydrogenase	Hydrogen oxidation or proton reduction	-	18/8	11/4	0/0
NADH oxidoreductase	Electron transport	CDD13801	77/1	3/0	0/0
Formate dehydrogenase alpha chain (FDH)	Oxidation of formate to carbon dioxide	CDD29449	151/70	22/4	0/0
Glycine reductase selenoprotein B	Glycine reductase	CDD12612	8/6	0/0	0/0
Arsenate reductase	Reduction of arsenate	CDD17412 CDD11108	311/23	0/0	0/0
Proteins with predicted redox function and catalytic redox-active Cys					
SelJ	ADP-ribosylation	-	36/0	0/0	5/10
SelK homologs	Function not known	-	0/0	0/0	16/57
SelS homologs	Translocation of misfolded proteins from the ER to cytosol	-	0/0	0/0	12/29
BthD and SelH homologs	Function not known, CxxC/U motif	-	0/0	0/0	4/21
SelM homologs	Function not known, CxxC/U motif	-	0/0	0/0	3/30
SelU homologs	Function not known, CxxC/U motif	-	0/0	0/0	48/15
Selenoprotein P (SelP)	Involved in Se transport, C/UxxC motif	CDD24729 CDD24730	0/0	0/0	0/36
SelT homologs	Function not known, CxxC/U motif	-	0/0	0/0	12/43
Sep15/Fep15 homologs	Function not known, CxC/U motif	-	0/0	0/0	10/52
SelO homologs	Function not known	CDD3203	116/0	0/0	19/14
Selenophosphate synthetase (SPS, SelD) homologs	Synthesis of selenophosphate	CDD10578	104/57	3/3	10/11
SelW-like proteins including SelV homologs	Function not known, CxxC/U motif	CDD16464 CDD12854	28/51	3/0	23/42
AdoMet-dependent methyltransferases (arsenic methyltransferase)	Arsenic detoxification	CDD10371	36/5	6/0	20/0
HesB-like protein	Biosynthesis of iron-sulfur clusters	CDD23223	189/5	4/2	0/0
Heterodisulfide reductase	Reduction of disulfides/sulfur metabolism	CDD10867	12/2	9/4	0/0
Molybdopterin biosynthesis MoeB family proteins	Possible reduction of a disulfide between MoaD and a rhodanese	CDD30111 CDD10349 CDD30117	282/11	6/0	35/0
Subfamily of glutathione S-transferase homologs	Possible glutathione-dependent oxidoreductase	CDD10495	234/8	0/0	52/0
Rhodanese-related sulfurtransferase superfamily	Multiple redox functions	CDD01448 CDD01444 CDD01524 CDD00158	871/25	19/0	46/0
DsrE-like protein	Sulfur oxidation/reduction	CDD15459 CDD11267	66/3	8/0	0/0
Hypothetical protein 1	Function not known	-	1/59	1/0	0/0
Hypothetical protein 2	Cyclic phosphodiesterase, CxC/U motif	CDD10605	8/28	0/0	0/0
Hypothetical protein 3	Possible iron transport/reduction, CxxC/U motif	CDD29747	62/2	0/0	0/0
Total			**8267/619**	**316/30**	**1829/554**

The first number in each box in the table shows the number of detected Cys-containing sequences, and the second number of Sec-containing sequences

*[1]-Peroxiredoxin family includes the following 6 conserved domains: CDD10324, CDD10547, CDD10943, CDD11785, CDD12957 and CDD24441

*[2]-Thioredoxin protein family includes the following 12 conserved domains: CDD10397, CDD11047, CDD11362, CDD11851, CDD12152, CDD12457, CDD12859, CDD13482, CDD14578, CDD23182, CDD26287 and CDD24359

*[3]-Glutaredoxin family includes the following 5 conserved domains: CDD10152, CDD10564, CDD12342, CDD13716 and CDD15697

Fig. S1

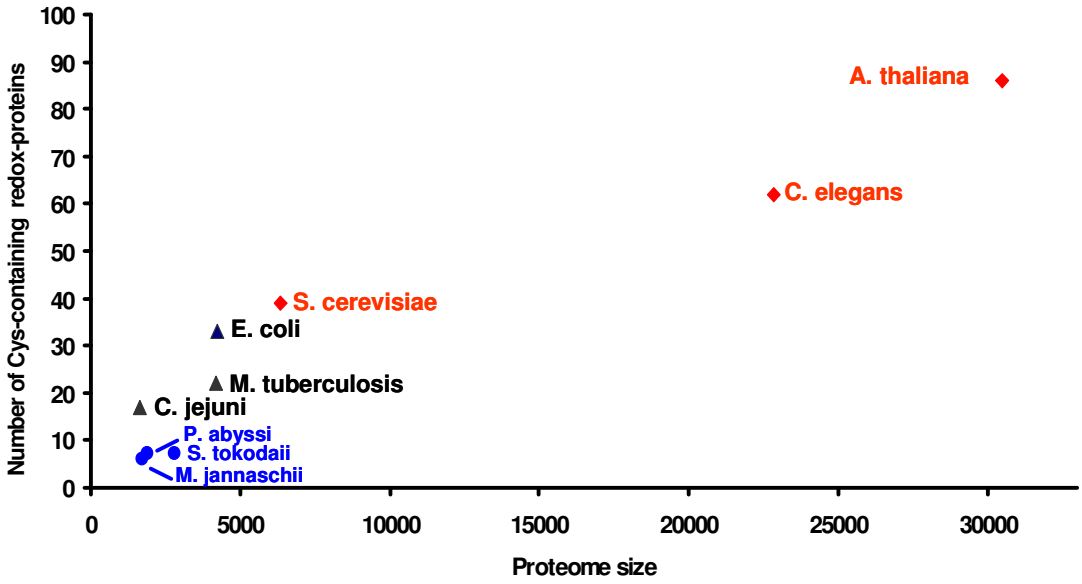
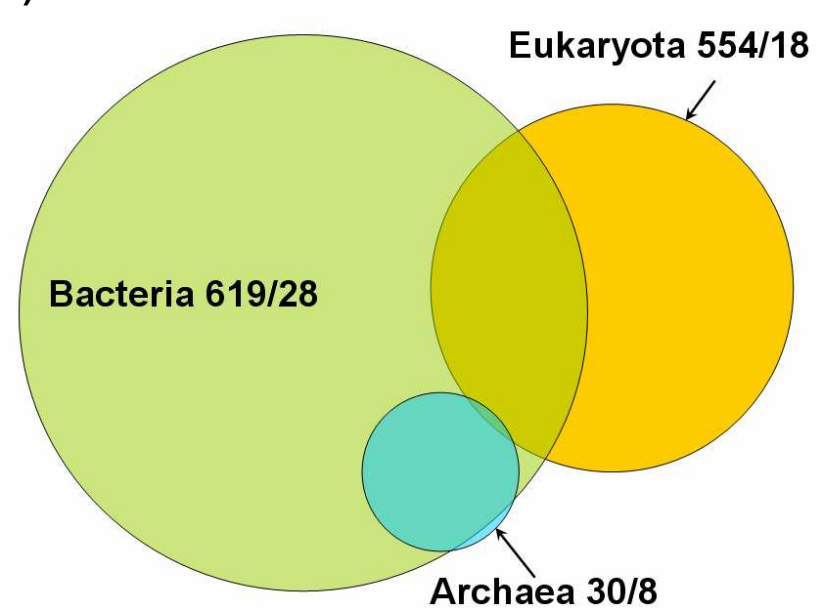

Number of Cys-containing redox-proteins
Proteome size
100
90
80
70
60
50
40
30
20
10
0
0
5000
10000
15000
20000
25000
30000
A. thaliana
C. elegans
S. cerevisiae
E. coli
M. tuberculosis
C. jejuni
P. abyssi
S. tokodaii
M. jannaschii

Fig. S2

a)


Eukaryota 554/18
Bacteria 619/28
Archaea 30/8

b)

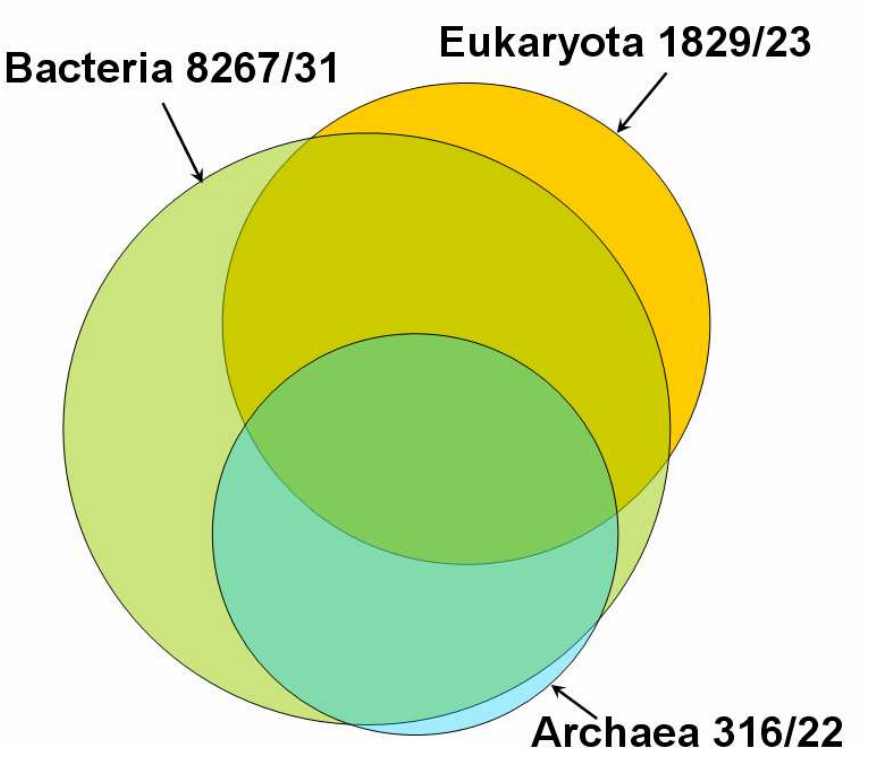

Bacteria 8267/31
Eukaryota 1829/23
Archaea 316/22

Fig. S3


42897919 Environmental sequence -MNKLTKSELNNIYPLGENRTG
44346764 Environmental sequence ---
42830445 Environmental sequence ---
44454992 Environmental sequence SNIKELVYEDTTGTIVNVDLTN
44323676 Environmental sequence SDSSSFSFINSDLVEDSEDFTN
44640430 Environmental sequence FGLYLFNSKRGLVLDNHNFSNFEYT
44604412 Environmental sequence FILYLYTSRRGLTIDGHNFSNY
21232053 X. campestris DKAH
23060100 P. fluorescens SSDL
28810438 S. pyogenes WRKK
28900827 V. parahaemolyticus SPLLDGVPIEGTLDNFSALLESSTP
29345628 B. thetaiotaomicron DKEMFLKDVFDYEKSKEWKYKGD
34897150 O. sativa RVVA
27763683 C. reinhardtii DRVVEVTSDQDFSAKLADVAGSGSL
41053764 D. rerio MVG
21554313 A. thaliana NAPN
28868894 P. syringae SGCG

Fig. S4

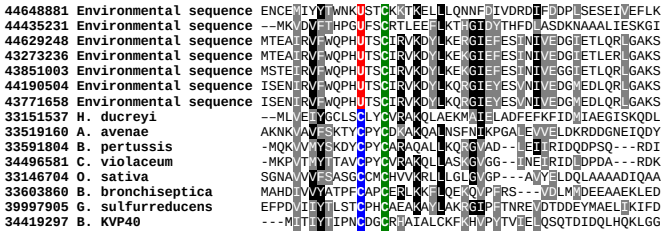

44648881 Environmental sequence
44435231 Environmental sequence
44629248 Environmental sequence
43273236 Environmental sequence
43851003 Environmental sequence
44190504 Environmental sequence
43771658 Environmental sequence
33151537 H. ducreyi
33519160 A. avenae
33591804 B. pertussis
34496581 C. violaceum
33146704 O. sativa
33603860 B. bronchiseptica
39997905 G. sulfurreducens
34419297 B. KVP40

Fig. S5

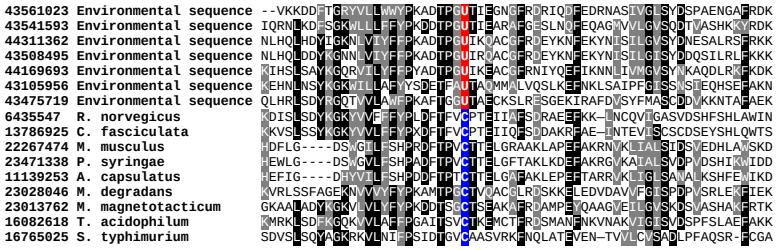

43561023 Environmental sequence
43541593 Environmental sequence
44311362 Environmental sequence
43508495 Environmental sequence
44169693 Environmental sequence
43105956 Environmental sequence
43475719 Environmental sequence
6435547 R. norvegicus
13786925 C. fasciculata
22267474 M. musculus
23471338 P. syringae
11139253 A. capsulatus
23028046 M. degradans
23013762 M. magnetotacticum
16082618 T. acidophilum
16765025 S. typhimurium

Fig. S6

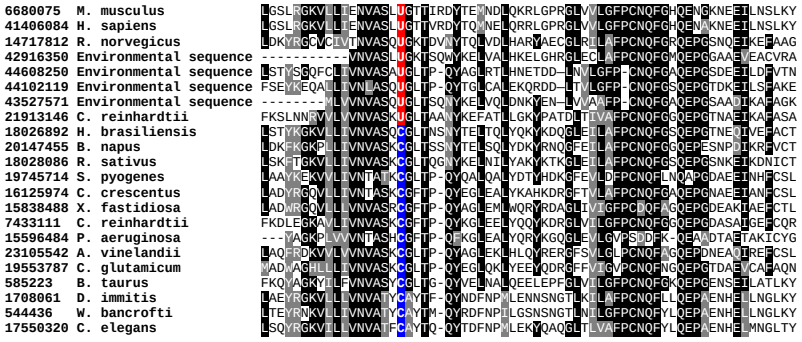

6680075 M. musculus
41406084 H. sapiens
14717812 R. norvegicus
42916350 Environmental sequence
44608250 Environmental sequence
44102119 Environmental sequence
43527571 Environmental sequence
21913146 C. reinhardtii
18026892 H. brasiliensis
20147455 B. napus
18028086 R. sativus
19745714 S. pyogenes
16125974 C. crescentus
15838488 X. fastidiosa
7433111 C. reinhardtii
15596484 P. aeruginosa
23105542 A. vinelandii
19553787 C. glutamicum
585223 B. taurus
1708061 D. immitis
544436 W. bancrofti
17550320 C. elegans

Fig. S7

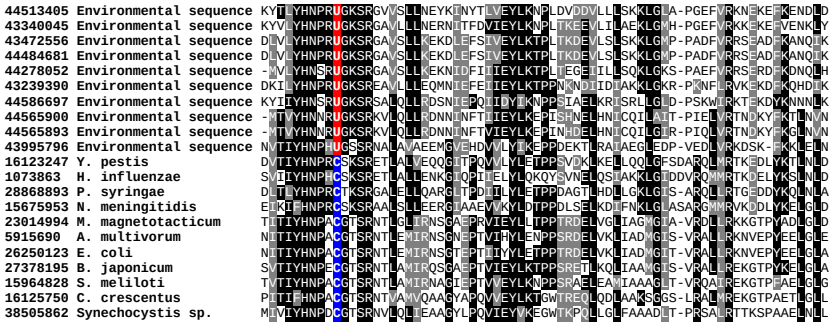

44513405 Environmental sequence
43340045 Environmental sequence
43472556 Environmental sequence
44484681 Environmental sequence
44278052 Environmental sequence
43239390 Environmental sequence
44586697 Environmental sequence
44565900 Environmental sequence
44565893 Environmental sequence
43995796 Environmental sequence
16123247 Y. pestis
1073863 H. influenzae
28868893 P. syringae
15675953 N. meningitidis
23014994 M. magnetotacticum
5915690 A. multivorum
26250123 E. coli
27378195 B. japonicum
15964828 S. meliloti
16125750 C. crescentus
38505862 Synechocystis sp.

Fig. S8

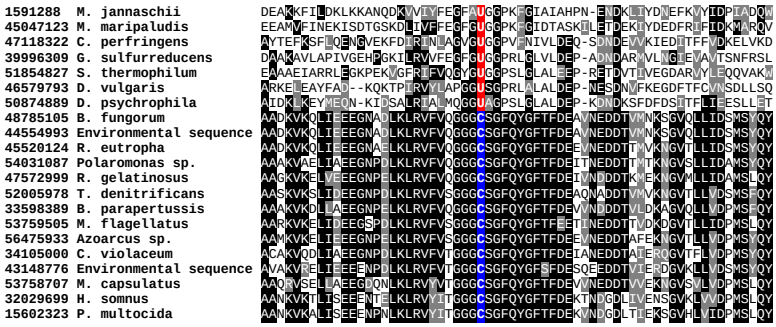

1591288 M. jannaschii
45047123 M. maripaludis
47118322 C. perfringens
39996309 G. sulfurreducens
51854827 S. thermophilum
46579793 D. vulgaris
50874889 D. psychrophila
48785105 B. fungorum
44554993 Environmental sequence
45520124 R. eutropha
54031087 Polaromonas sp.
47572999 R. gelatinosus
52005978 T. denitrificans
33598389 B. parapertussis
53759505 M. flagellatus
56475933 Azoarcus sp.
34105000 C. violaceum
43148776 Environmental sequence
53758707 M. capsulatus
32029699 H. somnus
15602323 P. multocida

Fig. S9

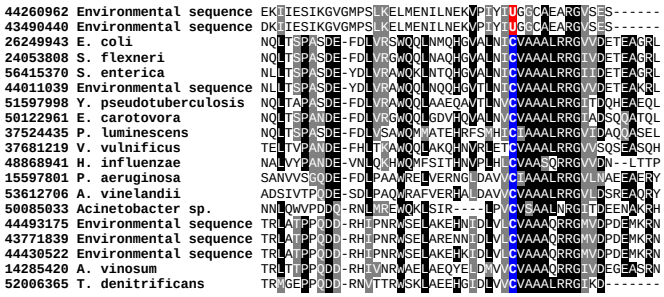

44260962 Environmental sequence
43490440 Environmental sequence
26249943 E. coli
24053808 S. flexneri
56415370 S. enterica
44011039 Environmental sequence
51597998 Y. pseudotuberculosis
50122961 E. carotovora
37524435 P. luminescens
37681219 V. vulnificus
48868941 H. influenzae
15597801 P. aeruginosa
53612706 A. vinelandii
50085033 Acinetobacter sp.
44493175 Environmental sequence
43771839 Environmental sequence
44430522 Environmental sequence
14285420 A. vinosum
52006365 T. denitrificans

Fig. S10

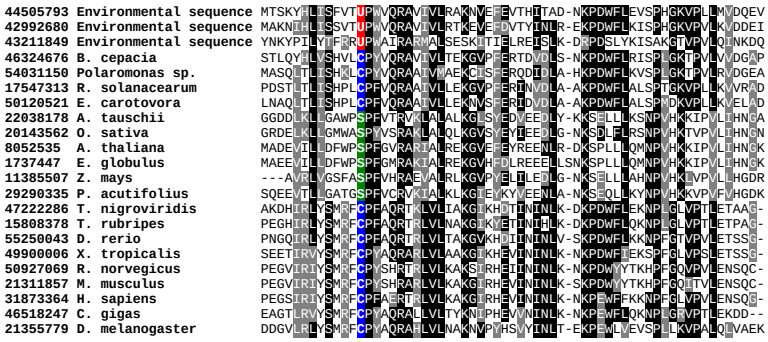

44505793 Environmental sequence
42992680 Environmental sequence
43211849 Environmental sequence
46324676 B. cepacia
54031150 Polaromonas sp.
17547313 R. solanacearum
50120521 E. carotovora
22038178 A. tauschii
20143562 O. sativa
8052535 A. thaliana
1737447 E. globulus
11385507 Z. mays
29290335 P. acutifolius
47222286 T. nigroviridis
15808378 T. rubripes
55250043 D. rerio
49900006 X. tropicalis
50927069 R. norvegicus
21311857 M. musculus
31873364 H. sapiens
46518247 C. gigas
21355779 D. melanogaster

Fig. S11

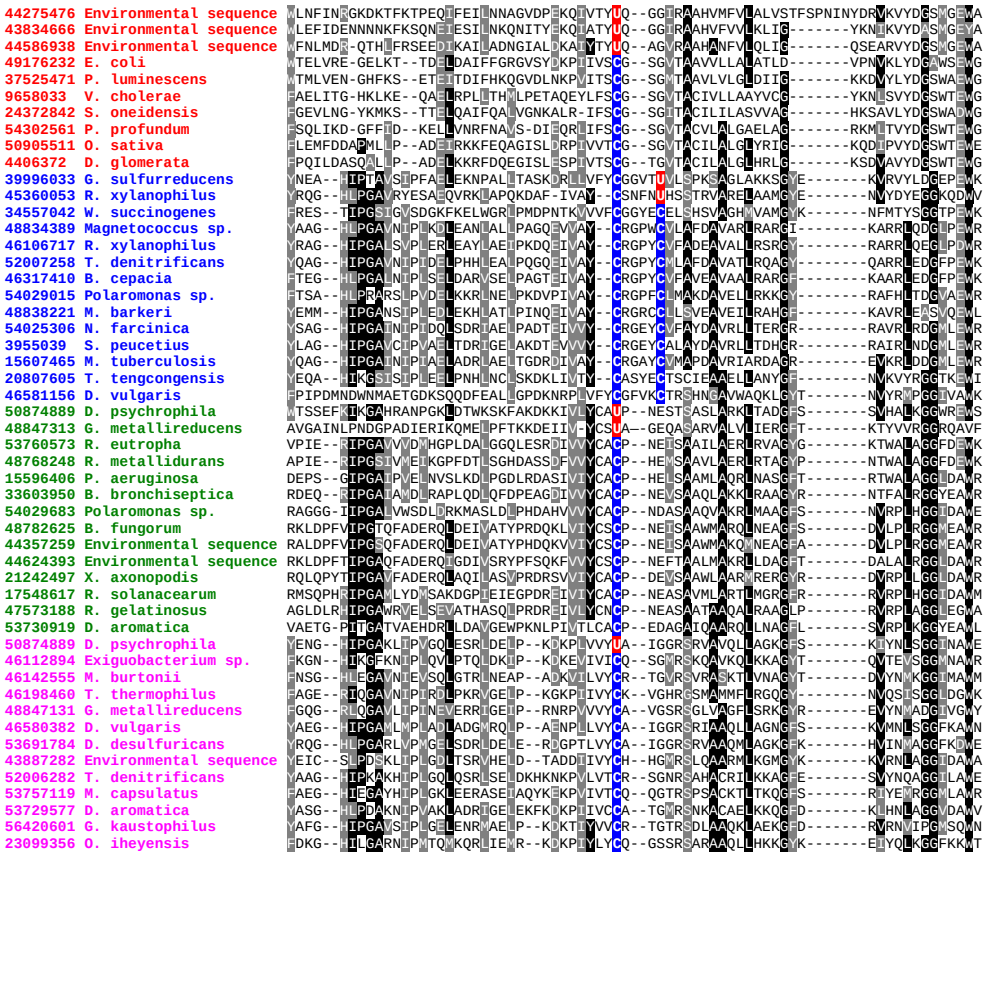

44275476 Environmental sequence
43834666 Environmental sequence
44586938 Environmental sequence
49176232 E. coli
37525471 P. luminescens
9658033 V. cholerae
24372842 S. oneidensis
54302561 P. profundum
50905511 O. sativa
4406372 D. glomerata
39996033 G. sulfurreducens
45360053 R. xylanophilus
34557042 W. succinogenes
48834389 Magnetococcus sp.
46106717 R. xylanophilus
52007258 T. denitrificans
46317410 B. cepacia
54029015 Polaromonas sp.
48838221 M. barkeri
54025306 N. farcinica
3955039 S. peucetius
15607465 M. tuberculosis
20807605 T. tengcongensis
46581156 D. vulgaris
50874889 D. psychrophila
48847313 G. metallireducens
53760573 R. eutropha
48768248 R. metallidurans
15596406 P. aeruginosa
33603950 B. bronchiseptica
54029683 Polaromonas sp.
48782625 B. fungorum
44357259 Environmental sequence
44624393 Environmental sequence
21242497 X. axonopodis
17548617 R. solanacearum
47573188 R. gelatinosus
53730919 D. aromatica
50874889 D. psychrophila
46112894 Exiguobacterium sp.
46142555 M. burtonii
46198460 T. thermophilus
48847131 G. metallireducens
46580382 D. vulgaris
53691784 D. desulfuricans
43887282 Environmental sequence
52006282 T. denitrificans
53757119 M. capsulatus
53729577 D. aromatica
56420601 G. kaustophilus
23099356 O. iheyensis

Fig. S12

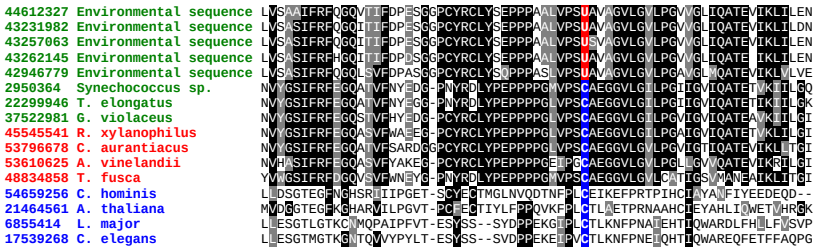

44612327 Environmental sequence
43231982 Environmental sequence
43257063 Environmental sequence
43262145 Environmental sequence
42946779 Environmental sequence
2950364 Synechococcus sp.
22299946 T. elongatus
37522981 G. violaceus
45545541 R. xylanophilus
53796678 C. aurantiacus
53610625 A. vinelandii
48834858 T. fusca
54659256 C. hominis
21464561 A. thaliana
6855414 L. major
17539268 C. elegans

Fig. S13


12642418 C. hydrogenoformans
15669884 M. jannaschii
45359260 M. maripaludis
20093689 M. kandleri
15679380 M. thermautotrophicus
46580807 D. vulgaris
23475491 D. desulfuricans
48847469 G. metallireducens
39995201 G. sulfurreducens
21674069 C. tepidum
20091692 M. acetivorans
21226158 M. mazei
48838060 M. barkeri
53731435 M. burtonii
11498837 A. fulgidus

Fig. S14

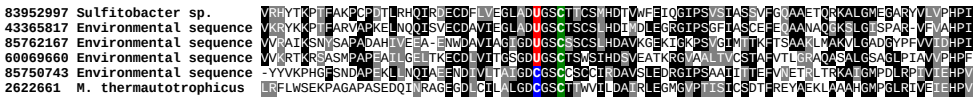

83952997 Sulfitobacter sp.
43365817 Environmental sequence
85762167 Environmental sequence
60069660 Environmental sequence
85750743 Environmental sequence
2622661 M. thermautotrophicus

Fig. S15


85801586 Environmental sequence
42966760 Environmental sequence
44228873 Environmental sequence
85772737 Environmental sequence
50874889 D. psychrophila
43122652 Environmental sequence
44504140 Environmental sequence
32445059 R. baltica
87311211 B. marina

Fig. S16

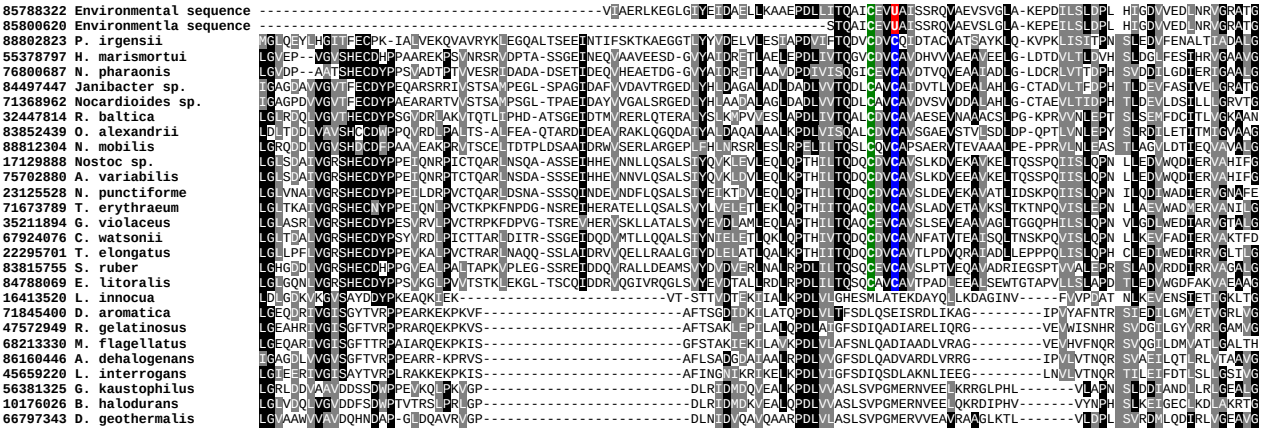

85788322 Environmental sequence
85800620 Environmentla sequence
88802823 P. irgensii
55378797 H. marismortui
76800687 N. pharaonis
84497447 Janibacter sp.
71368962 Nocardioides sp.
32447814 R. baltica
83852439 O. alexandrii
88812304 N. mobilis
17129888 Nostoc sp.
75702880 A. variabilis
23125528 N. punctiforme
71673789 T. erythraeum
35211894 G. violaceus
67924076 C. watsonii
22295701 T. elongatus
83815755 S. ruber
84788069 E. litoralis
16413520 L. innocua
71845400 D. aromatica
47572949 R. gelatinosus
68213330 M. flagellatus
86160446 A. dehalogenans
45659220 L. interrogans
56381325 G. kaustophilus
10176026 B. halodurans
66797343 D. geothermalis

Fig. S17.

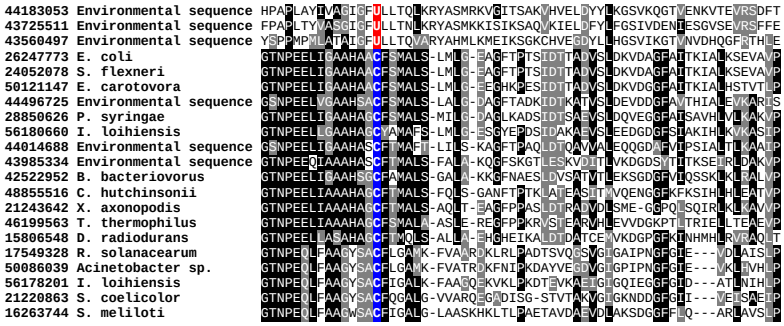

44183053 Environmental sequence
43725511 Environmental sequence
43560497 Environmental sequence
26247773 E. coli
24052078 S. flexneri
50121147 E. carotovora
44496725 Environmental sequence
28850626 P. syringae
56180660 I. loihiensis
44014688 Environmental sequence
43985334 Environmental sequence
42522952 B. bacteriovorus
48855516 C. hutchinsonii
21243642 X. axonopodis
46199563 T. thermophilus
15806548 D. radiodurans
17549328 R. solanacearum
50086039 Acinetobacter sp.
56178201 I. loihiensis
21220863 S. coelicolor
16263744 S. meliloti

Fig. S18

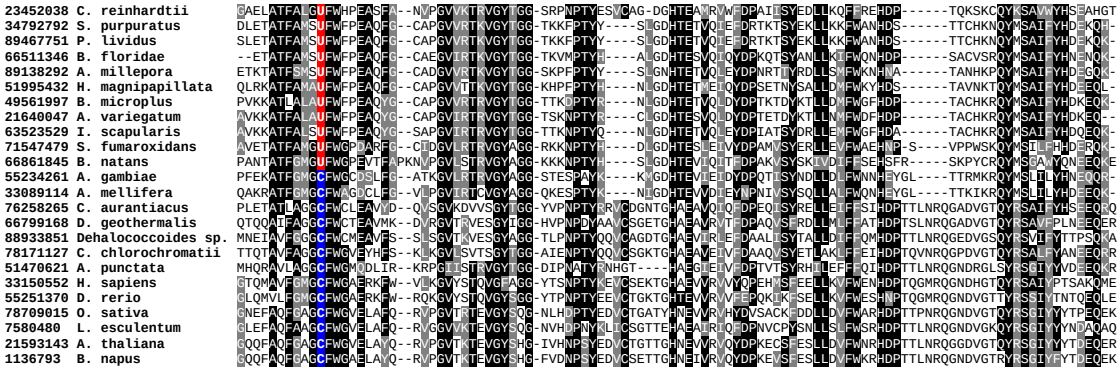

23452038 C. reinhardtii
34792792 S. purpuratus
89467751 P. lividus
66511346 B. floridae
89138292 A. millepora
51995432 H. magnipapillata
49561997 B. microplus
21640047 A. variegatum
63523529 I. scapularis
71547479 S. fumaroxidans
66861845 B. natans
55234261 A. gambiae
33089114 A. mellifera
76258265 C. aurantiacus
66799168 D. geothermalis
88933851 Dehalococcoides sp.
78171127 C. chlorochromatii
51470621 A. punctata
33150552 H. sapiens
55251370 D. rerio
78709015 O. sativa
7580480 L. esculentum
21593143 A. thaliana
1136793 B. napus

Fig. S19

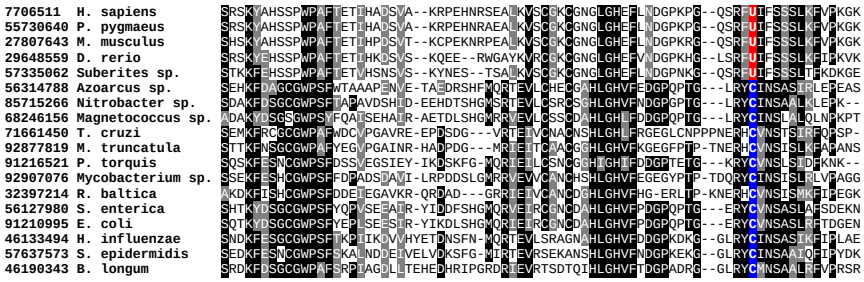

7706511 H. sapiens
55730640 P. pygmaeus
27807643 M. musculus
29648559 D. rerio
57335062 Suberites sp.
56314788 Azoarcus sp.
85715266 Nitrobacter sp.
68246156 Magnetococcus sp.
71661450 T. cruzi
92877819 M. truncatula
91216521 P. torquis
92907076 Mycobacterium sp.
32397214 R. baltica
56127980 S. enterica
91210995 E. coli
46133494 H. influenzae
57637573 S. epidermidis
46190343 B. longum

Fig. S20

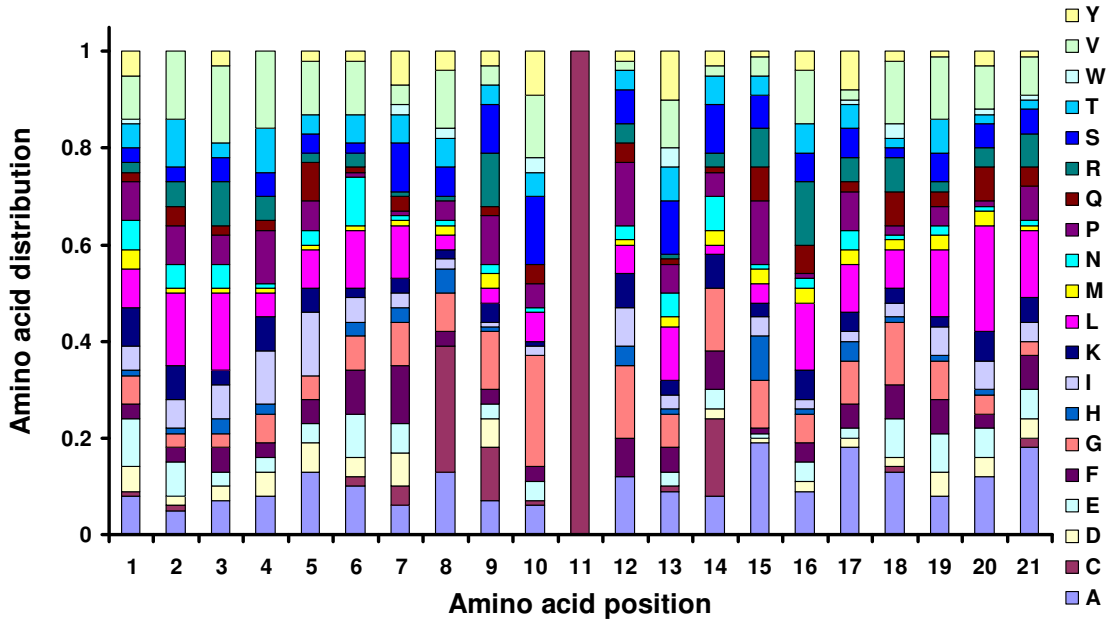
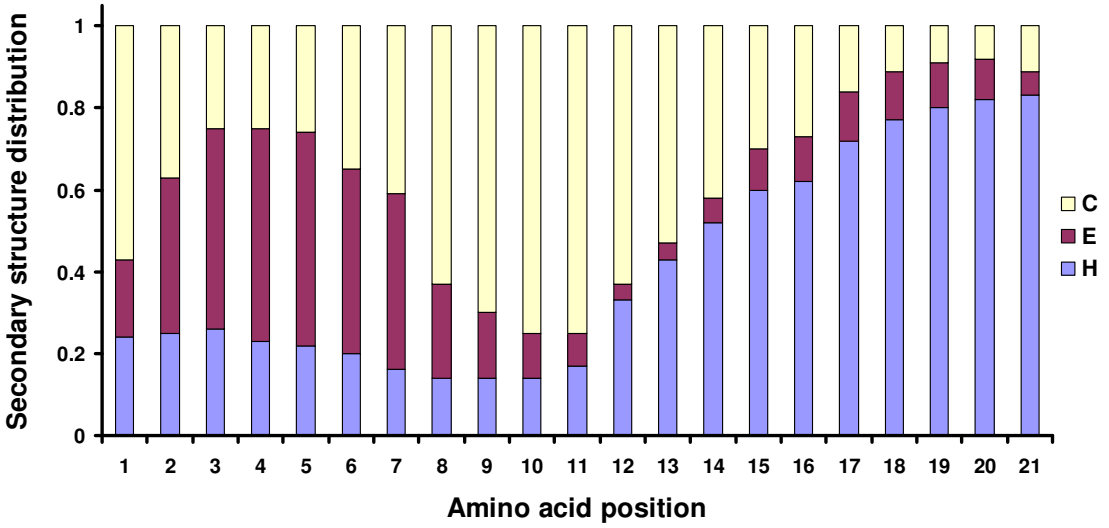
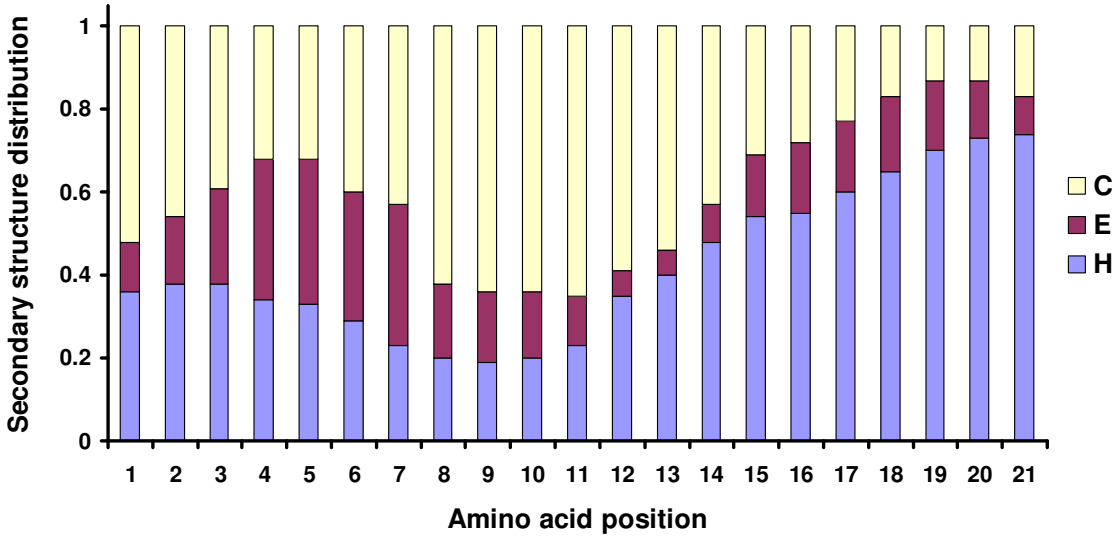

Amino acid distribution
0
0.2
0.4
0.6
0.8
1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
Amino acid position
Y
V
W
T
S
R
Q
P
N
M
L
K
I
H
G
F
E
D
C
A

Fig. S21

a)


Secondary structure distribution
0
0.2
0.4
0.6
0.8
1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
Amino acid position
C
E
H

b)


Secondary structure distribution
0
0.2
0.4
0.6
0.8
1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
Amino acid position
C
E
H