University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Vadim Gladyshev Publications | Biochemistry, Department of

1-19-2007

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

Follow this and additional works at: http://digitalcommons.unl.edu/biochemgladyshev

Part of the Biochemistry, Biophysics, and Structural Biology Commons

Fomenko, Dmitri E.; Xing, Weibing; Adair, Blakely M.; Thomas, David J.; and Gladyshev, Vadim, "Supporting Online Material for "High-Throughput Identification of Catalytic Redox-Active Cysteine Residues"" (2007). *Vadim Gladyshev Publications.* Paper 31.
http://digitalcommons.unl.edu/biochemgladyshev/31

This Article is brought to you for free and open access by the Biochemistry, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Vadim Gladyshev Publications by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

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:

Materials and Methods

SOM Text

Figs. S1 to S21

Table S1

References

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'-agttttatgggggaatccctgggaggcgctctg-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

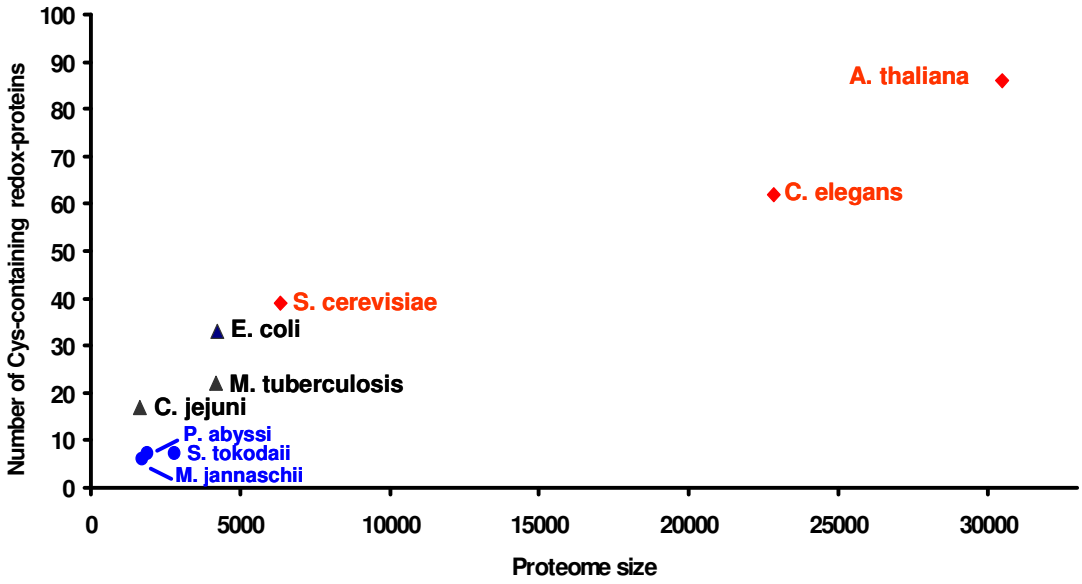
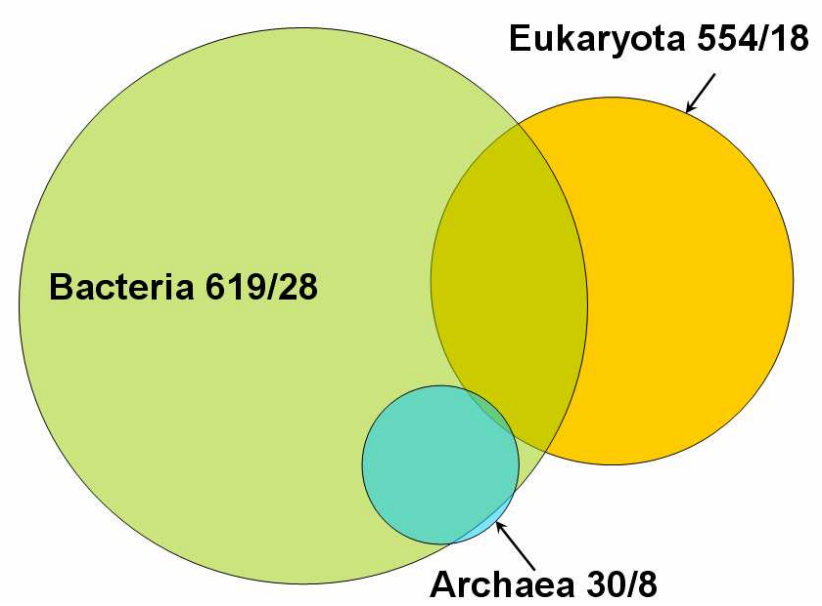

Number of Cys-containing redox-proteins
100
90
80
70
60
50
40
30
20
10
0
A. thaliana
C. elegans
S. cerevisiae
E. coli
M. tuberculosis
C. jejuni
P. abyssi
S. tokodaii
M. jannaschii
0
5000
10000
15000
20000
25000
30000
Proteome size

Fig. S2

a)


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

b)

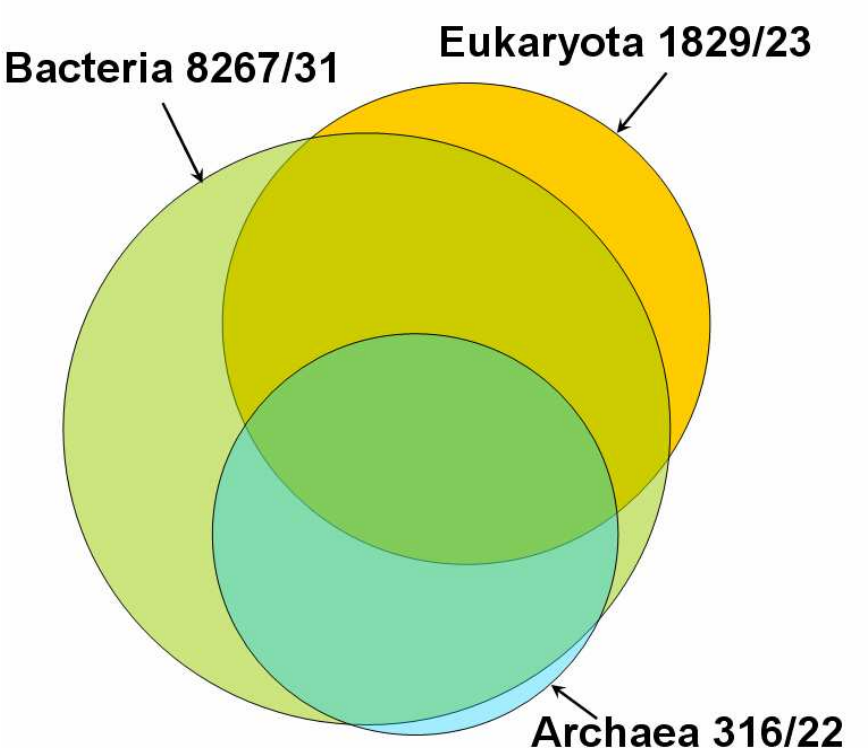

Bacteria 8267/31
Eukaryota 1829/23
Archaea 316/22

Fig. S3

```
42897919 Environmental sequence -MNKLTKSELNNIYPLGENRTGYPIFIDFYADWUGPCRMFEQVLNEYTQQYEG-KIQYYKYDI
44346764 Environmental sequence ---YELTYQNLDGELVNEDLTNYNTIIVFWADYUGTCRQELPVLEANIDKLIEYDVIALAHS
42830445 Environmental sequence ---YELTYQNLDGEFVNENLTNYNTIIVFWADYUGTCRQELPVLEANIDKLIEYDVIALAHS
44454992 Environmental sequence SNIKELVYEDTTGTIVNVDLTNYKTIVVFWADYUGICREELPLLEANLTTISENYNVIALAHS
44323676 Environmental sequence  SDSSFSFINSDLVEDSEDFTNYKTIIVFWADYUSVCRQELPVLEAELENLQEYDIIALAHS
44640430 Environmental sequence FGLYLFNSKRGLVLDNHNFSNFEYTIIEFYSDYULGCTASKFIVNEFKKKHEEIPIVSVNASK
44604412 Environmental sequence FILYLYTSRRGLTIDGHNFSNYYYTILEYYSDYULGCTASKLIVDEFKKKNPEIPIVSINASK
21232053 X. campestris         DKAHYFDVTTDTFETEVLQKSLTTPVLVDFWATWCGPCKSLTPILEKLAADYNG-AFEIAKYDV
23060100 P. fluorescens        SSDLIKHVSDASFEADVLKAEG--AVLVDYWAEWCGPCKMIAPVLDEIAETYKG-KLTVAKINI
28810438 S. pyogenes           WRKKYALEVTDATFVEETKEG---LVLIDFWATWCGPCRMQAPILEQLSQEIDEDELKILKYDV
28900827 V. parahaemolyticus   SPLLDGVPIEGTLDNFSALLESSTPVVVDFWAPWCNPCVGFAPVFSDVAQEQAG-TARFVKIDT
29345628 B. thetaiotaomicron   DKEMFLKDVFDYEKSKEWKYKGDYPAIIDLYADWCGPCRQTAPIMKELAKEYAG-KITIYKVNV
34897150 O. sativa             RVVAVHSTATWDEQWGAHKSNPNYLIVIDFSATWCGPCRFIEPAFKDVAGRFAVAVFFKIDVDE
27763683 C. reinhardtii        DRVVEVTSDQDFSAKLADVAGSGSLYICDFTAKWCGPCRMIAPVFSSLSNKYTDVTFVKIDIDN
41053764 D. rerio               MVGYKVIGNDSDFQAELSGAGSYLTVVKFTMSGCRPCVRIAPAFNMLSNKYPQVVFLEVDVHV
21554313 A. thaliana           NAPNVVDIHSTEEFLSALSGAGELLVIVEFYGIWCASCRALFPKLCKTAVEHPDIVFLKVNFDE
28868894 P. syringae           SGCGVDLGTDQNGQKVASERIKGIWLVVNYWAEWCGPCRTEVPEFNALSEQLKDKKVTVLGVNF
```

Fig. S4

```
44648881 Environmental sequence ENCEMIYYTWNKTSTCKKTKELLLQNNFDIVDRDIFDDPLSESEIYEFLK
44435231 Environmental sequence --MKVDVFIHPGTFSCRTLEEFLKTFGIDYTHFDLASDKNAAALIESKGI
44629248 Environmental sequence MTEAIRVFWQPHTTSCIRVKDYLKERGIEFESINIVEDGIETLQRLGAKS
43273236 Environmental sequence MTEAIRVFWQPHTTSCIRVKDYLKEKGIEFESINIVEDGIETLERLGAKS
43851003 Environmental sequence MSTEIRVFWQPHTTSCIRVKDYLKEKGIEFESINIVEGGIETLQRLGAKS
44190504 Environmental sequence ISENIRVFWQPHTTSCIRVKDYLKQRGIEYESVNIVEDGVEDLQRLGAKS
43771658 Environmental sequence ISENIRVFWQPHTTSCIRVKDYLKQRGIEYESVNIVEDGVEDLQRLGAKS
33151537 H. ducreyi            --MLVEIYGCLSCLYCVRAKQLAEKMAIELADFEFKFIDMIAEGISKQDL
33519160 A. avenae             AKNKVAVFSKTYCPYCDKAKQALNSFNIKPGALEVVELDKRDDGNEIQDY
33591804 B. pertussis          -MQKVVYYSKDYCPYCARAQALLKQRGVAD--LEIIRIDQDPSQ---RDI
34496581 C. violaceum          -MKPVTMYITAVCPYCVRAKQLLASKGVGG--INELRIDLDPDA---RDK
33146704 O. sativa             SGNAVVVFSASGCCMCHVVKRLLLGLGVGP---AVYELDQLAAAADIQAA
33603860 B. bronchiseptica     MAHDIVVYATPFCAPCERLKKFLQEKQVPFRS---VDLMVDEEAAEKLED
39997905 G. sulfurreducens     EFPDVIIYTLSTCPHCAEAKAYLAKRGIPFTNREVDTDDEYMAELIKIFD
34419297 B. KVP40              ---MITIYIIPNCDGCRIAIALCKFKHVPYTVIELQSQTDIDQLHQKLGG
```

Fig. S5

```
43561023 Environmental sequence --VKKDDFTGRYVLLWWYPKADTPGUTIEGNGFRDRIQDFEDRNASIVGLSYDSPAENGAFRDK
43541593 Environmental sequence IQRNLKDFSGKWLLLFFYPKDDTPGUTIEARAFGESLNQFEQAGYVVLGVSQDTVASHKYRDK
44311362 Environmental sequence NLHQLHDYIGKNLVIYFFPKADTPGUIKQACGFRDEYKNFEKYNISILGVSYDNESALRSFRKK
43508495 Environmental sequence NLHQLDDYKGNNLVIYFFPKADTPGUIRQACGFRDEYKNFEKYNISILGISYDDQSILRLFKKK
44169693 Environmental sequence KIHSLSAYKGQRVILYFFPYADTPGUIKEACGFRNIYQEFIKNNIIVMGVSYDKAQDLRKFKDK
43105956 Environmental sequence KEHNLNSYKGKWILLAFYYSDEFFAUTAQMMALVQSLKEFNKLSAIPFGISSDSEQHSEFAKN
43475719 Environmental sequence QLHRLSDYRGQTVVLAWFPKAFTGCUTAECKSLRESGEKIRAFDVSYFMASCDDVKKNTAFAEK
6435547  R. norvegicus          KDISLSDYKGKKVVFFFPLDFTFVCPTEIIAFSDRAEEFKK-INCQVIGASVDSHFSHLAWIN
13786925 C. fasciculata         KKVSLSSYKGKYVVLFFPKDFTFVCPTEIIQFSDDAKRFAE-INTEVISCSCDSEYSHLQWTS
22267474 M. musculus            DFLG----DSVGLLFSHPRDFTPVCITELGRAAKLAPEFAKRNVKLIALSIDSYEDHLAWSKD
23471338 P. syringae            EWLG----DSVGVLFSHPADFTPVCITELGFTAKLKDEFAKRGVKAIALSVDPVDSHIKWIDD
11139253 A. capsulatus          EFIG----DHYVILFSHPDDFTPICITELGVFAKLEPEFTARRVKLIGLSAVALKSHFEIKD
23028046 M. degradans           KVRLSSFAGEKNVVYFYPKAMTPGCTVQACSLRDSKKELEDVDAVVFGISPDPVSRLEKFIEK
23013762 M. magnetotacticum     GKAALADYKGKVLVLYFYPKDDTSGCTSEAKAFRDAMPEYQAAGYEILGVSKDSVASHAKFRTK
16082618 T. acidophilum         KMRKLSDFKGQKVVLAFFPGAITSVCTKEMCTFRDSMANFNKVNAKVIGISVDSPFSLAEFAKK
16765025 S. typhimurium         SDVSLSQYAGKRKVLNIWFPSIDTGVCAASVRKFNQLATEVEN-TVYLCVSADLPFAQSR-FCGA
```

Fig. S6

```
6680075  M. musculus            LGSLRGKVLLIENVASLVGTTIRDYTEYNDLQKRLGPRGLVVLGFPCNQFGHQENGKNEEILNSLKY
41406084 H. sapiens             LGSLRGKVLLIENVASLVGTTVRDYTQYNELQRRLGPRGLVVLGFPCNQFGHQENAKNEEILNSLKY
14717812 R. norvegicus          LDKYRGCVCLVINVASQVGKTDVNYTQLVDLHARYAECGLRILAFPCNQFGRQEPGSNQEIKEFAAG
42916350 Environmental sequence ------------VNVASLVGKTSQWYKELVALHKELGHRGLECLAFPCNQFGMQEPGGAAEVEACVRA
44608250 Environmental sequence LSTYSGQFCLIVNVASAVGLTP-QYAGLRTLHNETDD-LNVLGFPCNQFGAQEPGSDEEILDFVTN
44102119 Environmental sequence FSEYKEQALLIVNLASQVGLTP-QYTGLCALEKQRDD-LTVLGFPCNQFGSQEPGTDKEILSFAKE
43527571 Environmental sequence --------TLVVNVASQVGLTSQNYKELVQLDNKYEN-LVVAAFPCNQFGAQEPGSAAEIKAFAGK
21913146 C. reinhardtii         FKSLNNRVVLVVNVASKVGLTAANYKEFATLLGKYPATDLTIVAFPCNQFGQEPGTNAEIKAFASA
18026892 H. brasiliensis        LSTYKGKVLLIVNVASQCGLTNSNYTELTQLYQKYKDQGLEILAFPCNQFGSQEPGTNEQIVEFACT
20147455 B. napus               LDKFKGKPLLIVNVASKCGLTSSNYTELSQLYDKYRNQGFEILAFPCNQFGGQEPESNPDIKRFVCT
18028086 R. sativus             LSKFTGKVLLIVNVASKCGLTQGNYKELNILYAKYKTKGLEILAFPCNQFGSQEPGSNKEIKDNICT
19745714 S. pyogenes            LAAYKEKVVLIVNTATKCGLTP-QYQALQALYDTYHDKGFEVLDFPCNQFLNQAPGDAEEINHFCSL
16125974 C. crescentus          LADYRGQVLLIVNTASKCGFTP-QYEGLEALYKAHKDRGFTVLAFPCNQFGAQEPGNAEIANFCSL
15838488 X. fastidiosa          LADWRGQVLLIVNVASRCGFTP-QYAGLEMLWQRYRDAGLIVIGFPCDQFAGQEPGDEAKIAEFCTL
7433111  C. reinhardtii        FKDLEGKAVLIVNVASKCGFTP-QYKGLEELYQQYKDRGLVILGFPCNQFGGQEPGDASAIGEFCQR
15596484 P. aeruginosa          ---VAGKPLVVVNTASHCGFTP-QFKGLEALYQRYKGQGLEVLGVPSDDFK-QEAADTAETAKICYG
23105542 A. vinelandii          LAQFRDKVVLVVNVASKCGLTP-QYAGLEKLHLQYRERGFSVLGLPCNQFAGQEPDNEAIIREFCSL
19553787 C. glutamicum          VADLAGLLLIVNVASKCGLTP-QYEGLQKLYEEYQDRGFFVIGVPCNQFNGQEPGTDAEVCAFAQN
585223   B. taurus              FKQVAGKYILFVNVASYCGLTG-QYVELNALQEELEPFGLVILGFPCNQFGKQEPGENSEILATLKY
1708061  D. immitis             LAEYRGKVLLLVNVATYCAYTF-QYNDFNPYLENNSNGTLKILAFPCNQFLLQEPAENHEILNGLKY
544436   W. bancrofti           LTEYRNKVLLIVNVATYCAYTM-QYRDFNPILGSNSNGTLNILGFPCNQFYLQEPAENHEILNGLKY
17550320 C. elegans             LSQYRGKVILLVNVATFCAYTQ-QYTDFNPYLEKYQAQGLTLVAFPCNQFYLQEPAENHEIMNGLTY
```

Fig. S7

```
44513405 Environmental sequence KYTLYHNPRUGKSRGVYSLLNEYKINYTLVEYLKNPLDVDDVLLLSKKLGIA-PGEFVRKNEKEPKENDLD
43340045 Environmental sequence KYVLYHNPRUGKSRGAVLLLNERNITFDVIEYLKNPLTKEEVLILAEKLGYH-PGEFVRKKEKEVENKLY
43472556 Environmental sequence DIVLYHNPRUGKSRGAVSLLKEKDLEFSIVEYLKTPLTKDEVLSLSKKLGYP-PADFVRRSEADPKANQLK
44484681 Environmental sequence DIVLYHNPRUGKSRGAVSLLKEKDLEFSIVEYLKTPLTKDEVLSLSKKLGYP-PADFVRRSEADPKANQLK
44278052 Environmental sequence -VVLYHNSRUGKSRGAVSLLKEKNIDFIIIEYLKTPLTEGEIILLSQKLGKS-PAEFVRRSERDPKDNQLH
43239390 Environmental sequence DKILYHNPRUGKSREAVLLLEQMNIEFEIIEYLKTPPNLNDIIDIAKKLGKR-PLNFLRVKEKDPKQHDIK
44586697 Environmental sequence KYIIYHNSRUGKSRSALQLLRDSNIEPQIIDYIKNPPSIAELKRISRLLGID-PSKWIRKTEKDYKNNNLK
44565900 Environmental sequence -WTVYHNNRUGKSRKVLQLLRDNNINFTIIEYLKEPISPNELHNICQILAIT-PIELVRTNDKYPKTLNVN
44565893 Environmental sequence -WTVYHNNRUGKSRKVLQLLRDNNINFTVIEYLKEPINHDELHNICQILGIR-PIQLVRTNDKYPKCLNVN
43995796 Environmental sequence NVTIYHNPHUGSSRNALAVAEEMGVEHDVVLYIKEPPDEKTLRAIAEGLEDP-VEDLVRKDSK-PKKLELN
16123247 Y. pestis              DVTIYHNPRCSKSRETLALVEQQGITPQVVLYLETPPSVDKLKELLQQLGFSDARQLMRTKEDLYKTLNLD
1073863  H. influenzae          SVIIYHNPHCSKSRETLALLENKGIQPIIELYLQKQYSVNELQSIAKKLGIDDVRQMMRTKDELYKSLNLD
28868893 P. syringae            DITLYHNPRCTKSRGALELLQARGITPDIILYLETPPDAGTLHDLLGKLGIS-ARQLLRTGEDDYKQLNLA
15675953 N. meningitidis        EIKIFHNPRCSKSRAALSLLEERGIAAEVVKYLDTPPDLSELKDIFNKLGIASARGMMRVKDDLYKELGLD
23014994 M. magnetotacticum     TITIYHNPACGTSRNTLGLIRNSGAEPRVIEYLITPPTRDELVGLIAGIGIA-VRDLLRKKGTPYADLGLD
5915690  A. multivorum          NITIYHNPACGTSRNTLEMIRNSGNEPTVIHYLENPPSRDELVKLIADIGIS-VRALLRKNVEPYEELGLE
26250123 E. coli                NITIYHNPACGTSRNTLEMIRNSGTEPTIIYYLETPPTRDELVKLIADIGIT-VRALLRKNVEPYEELGLA
27378195 B. japonicum           SVTIYHNPECGTSRNTLAMIRQSGAEPTVIEYLKTPPSREILKQLIAAIGIS-VRALLREKGTPYKELGLA
15964828 S. meliloti            TVTIYHNPACGTSRNTLAMIRNAGIEPTVVEYLKNPPSRAELEAIAAAGIT-VRQALREKGTPIAELGLG
16125750 C. crescentus          PITIFHNPACGTSRNTVAMVQAAGYAPQVVEYLKTGWTREQLQDLAAKSGGS-LRALMREKGTPAETLGLL
38505862 Synechocystis sp.      MIVIYHNPDCGTSRNVLQLLEAAGYLPQVIEYVKEGWTKPQLLGLFAAADIT-PRSALRTTKSPAAELNLL
```

Fig. S8

```
1591288  M. jannaschii          DEAKKFILDKLKKANQDKVVIYFEGFAUGGPKFGIAIAHPN-ENDKLIYDNEFKYYIDPIADQW
45047123 M. maripaludis         EEAMVFINEKISDTGSKDLIVFFEGFGUGGPKFGIDTASKILETDEKIYDEDFRIFIDKVARQV
47118322 C. perfringens         AYTEFKSFIQENGVEKFDFRINIAGVGUGGPVFNIVLDEQ-SDNDEVVKIEDITFFVDKELVKD
39996309 G. sulfurreducens      DAAKAVLAPIVGEHPGKILRVVFEGFGUGGPRLGLVLDEP-ADNDARAVLNGIEVAVTSNFRSL
51854827 S. thermophilum        EAAAEIARRLEGKPEKVGFRIFVQGYGUGGPSLGLALEEP-RETDVTIVEGDARVYLEQQVAKW
46579793 D. vulgaris            ARKEIEAYFAD--KQKTPLRVYLAPGGUSGPRLALALDEP-NESDNVFKEGDFTFCVNSDLLSQ
50874889 D. psychrophila        AIDKLKEYVEQN-KIDSALRIALMQGGUAGPSLGLALDEP-KDNDKSFDFDSITFLIEESLLET
48785105 B. fungorum            AADKVKQLIEEEGNADLKLRVFVQGGGCSGFQYGFTFDEAVNEDDTVVNKSGVQLLIDSMSYQY
44554993 Environmental sequence AADKVKQLIEEEGNADLKLRVFVQGGGCSGFQYGFTFDEAVNEDDTVVNKSGVQLLIDSMSYQY
45520124 R. eutropha            AADKVKQLIEEEGNAELKLRVFVQGGGCSGFQYGFTFDEEVNEDDTTVVKNGVTLLIDSMSYQY
54031087 Polaromonas sp.        AAAKVAELIAEEGNPDLKLRVFVQGGGCSGFQYGFTFDEITNEDDTTVTKNGVSLLIDAMSYQY
47572999 R. gelatinosus         AAGKVKELVEEEGNPELKLRVFVQGGGCSGFQYGFTFDEIVNDDDTKVEKNGVMLLIDAMSLQY
52005978 T. denitrificans       AASKVKSLIDEEGNPDLKLRVFVSGGGCSGFQYGFTFDEAQNADDTVVVKNGVTLLVDSMSFQY
33598389 B. parapertussis       AAAKVKDLLAEEGNPELKLRVFVQGGGCSGFQYGFTFDEVVNDDDTVLDKAGVQLLVDPMSFQY
53759505 M. flagellatus         AARKVKELIDEEGSPDLKLRVFVSGGGCSGFQYGFTFEETINEDDTTVDKDGVTLLIDPMSLQY
56475933 Azoarcus sp.           AAMKVKELIEEEGNPELKLRVFVSGGGCSGFQYGFTFDEEVNEDDTAFEKNGVTLLVDPMSYQY
34105000 C. violaceum           ACAKVQDLIAEEGNPDLKLRVFVTGGGCSGFQYGFTFDEIANEDDTAIERQGVTFLVDPMSYQY
43148776 Environmental sequence AVAKVRELIEEENPDLKLRVFVTGGGCSGFQYGFSFDESQEEDDTVIERDGVKLLVDSMSYQY
53758707 M. capsulatus          AAQRVSELLAEEGDQNLKLRVYVTGGGCSGFQYGFTFDEVVNEDDTVYEKNGVSVLVDPMSLQY
32029699 H. somnus              AANKVKTLISEENTELKLRVYITGGGCSGFQYGFTFDEKTNDGDLIVENSGVKLVVDPMSLQY
15602323 P. multocida           AANKVKALISEENPNLKLRVYITGGGCSGFQYGFTFDEKVNDGDLTIEKSGVHLVIDPMSLQY
```

Fig. S9

```
44260962 Environmental sequence EKTIESIKGVGMPSLKELMENILNEKVPIYITGTCAEARGVSESS------
43490440 Environmental sequence DKTIESIKGVGMPSLKELMENILNEKVPIYITGTCAEARGVSESS------
26249943 E. coli                NQLTSPASDE-FDLVRSWQQLNMQTGVALNICVAAALRRGVVDETEAGRL
24053808 S. flexneri            NQLTSPASDE-FDLVRGWQQLNAQTGVALNICVAAALRRGIVDETEAGRL
56415370 S. enterica            NLLTSPASDE-YDLVRAWQKLNTQTGVALNICVAAALRRGIIDETEAGRL
44011039 Environmental sequence NLLTSPASDE-YDLVRAWQLNQQTGVTLNICVAAALRRGVVDETEAKRL
51597998 Y. pseudotuberculosis  NQLTAPASDE-FDLVRAWQQLAAEQAVTLNVCVAAALRRGITDQHEAEQL
50122961 E. carotovora          NQLTSPANDE-FDLVRGWQQLGDVEQVALNVCVAAALRRGIADSQQATQL
37524435 P. luminescens         NQLTSPASDE-FDLVSAWQMWATETRFSWHICIAAALRRGVIDAQQASEL
37681219 V. vulnificus          TELTVPANDE-FHLTKAWQQLAKQTNVRLETCVAAALRRGVVSQSEASQH
48868941 H. influenzae          NALVYPANDE-VNLQKHWQMFSITTNVPLHLCVAASQRRGVVDN--LTTP
15597801 P. aeruginosa          SANVVSGQDE-FDLPAAWRELVERNGTDAVVCIAAALRRGVLNAEEAERY
53612706 A. vinelandii          ADSIVTPQDE-SDLPAQWRAFVERHALDAVVCVAAALRRGVLDSREAQRY
50085033 Acinetobacter sp.      NNLQWVPDDQ-RNLMREWQKLSIR----LPVCVSAALNRGITDEENAKRH
44493175 Environmental sequence TRLATPPQDD-RHIPNRWSELAKENTDLVLCVAAAQRRGMVDPDEMKRN
43771839 Environmental sequence TRLATPPQDD-RHIPNRWSELARENNTDLVLCVAAAQRRGMVDPDEMKRN
44430522 Environmental sequence TRLATPPQDD-RHIPNRWSELAKEKKTDLVLCVAAAQRRGMVDPDEMKRN
14285420 A. vinosum            TRLTTPPQDD-RHIVNRWAELAEQYELDLVVCVAAAQRRGIVDEGEASRN
52006365 T. denitrificans       TRWGEPPQDD-RNVTTRWSKLAEETGLDLVVCVAAALRRGIKD-------
```

Fig. S10

```
44505793 Environmental sequence MTSKYHLISFVTUPWVQRAVIVLRAKNVEFEVTHITAD-NKPDWFLEVSPHGKVPLLMVDQEV
42992680 Environmental sequence MAKNIHLISSVTUPWVQRAVIVLRTKEVEFDVTYINLR-EKPDWFLKISPHGKVPVLKVDDEI
43211849 Environmental sequence YNKYPILYIFRRUPWAIRARMALSESKITIELREISLK-DRPDSLYKISAKGTVPVLQINKDQ
46324676 B. cepacia            STLQYHLVSHVLCPYVQRAVIVLTEKGVPFERTDVDLS-NKPDWFLRISPLGKTPVLVVDGAP
54031150 Polaromonas sp.       MASQITLISHKLCPYVQRAAIVMAEKCISFERQDIDLA-HKPDWFLKVSPLGKTPVLRVDGEA
17547313 R. solanacearum       PDSTITLISHPLCPFVQRAAIVLLEKGVPFERINVDLA-AKPDWFLALSPTGKVPLLKVVRAD
50120521 E. carotovora         LNAQITLISHPLCPFVQRAAIVLLEKNVSFERIDVDLA-AKPDWFLALSPVDKVPLLKVELAD
22038178 A. tauschii           GGDDLKLLGAWPSPFVTRVKLALALKGLSYEDVEEDLY-KKSELLLKSNPVHKKIPVLIHNGA
20143562 O. sativa             GRDELKLLGMWASPYVSRAKLALQLKGVSYEYIEEDLG-NKSDLFLRSNPVHKTVPVLIHNGN
8052535  A. thaliana           MADEVILLDFWPSPFGVRARIALREKGVEFEYREENLR-DKSPLLLQMNPVHKKIPVLIHNGK
1737447  E. globulus           MAEEVILLDFWPSPFGMRAKIALREKGVHFDLREEELLSNKSPLLLQMNPVHKKIPVLIHNGK
11385507 Z. mays               ---AVRLVGSFASPFVHRAEVALRLKGVPYELILEDLG-NKSELLLAHNPVHKLVPVLLHGDR
29290335 P. acutifolius        SQEEVTLLGATGSPFVCRVKIALKLKGIEVKYVEENLA-NKSEQLLKYNPVHKKVPVFVHGDK
47222286 T. nigroviridis       AKDHIRLYSMRFCPFAQRTKLVLIAKGIKHDTININLK-DKPDWFLEKNPLGLVPTLETAAG-
15808378 T. rubripes           PEGHIRLYSMRFCPFAQRTRLVLNAKGIKYETINIHLK-DKPDWFLQKNPLGLVPTLETPAG-
55250043 D. rerio              PNGQIRLYSMRFCPFAQRTRLVLTAKGVKHDIININLV-SKPDWFLKKNPFGTVPVLETSSG-
49900006 X. tropicalis         SEETIRVYSMRFCPYAQRARLVLAAKGIKHEVININLK-NKPDWFIEKSPFGLVPSLETSSG-
50927069 R. norvegicus         PEGVIRIYSMRFCPYSHRTRLVLKAKSIRHEIININLK-NKPDWYYTKHPFGQVPVLENSQC-
21311857 M. musculus           PEGVIRIYSMRFCPYSHRARLVLKAKGIRHEVININLK-SKPDWYYTKHPFGQITVLENSQC-
31873364 H. sapiens            PEGSIRIYSMRFCPFAERTRLVLKAKGIRHEVININLK-NKPEWFFKKNPFGLVPVLENSQG-
46518247 C. gigas              EAGTLRVYSMRFCPYAQRALLVLTYKNIPHEVVNINLK-NKPEWFLQKNPLGRVPTLEKDD--
21355779 D. melanogaster       DDGVLRLYSMRFCPYAQRAHLVLNAKNVPYHSVYINLT-EKPEWLVEVSPLLKVPALQIVAEK
```

Fig. S11

```
44275476 Environmental sequence YLNFINSGKDKTFKTPEQIFEIINNAGVDPSKQIVTYUQ--GGIRAAHVMFVLALVSTFSPNINYDRVKVYDGSWGEWA
43834666 Environmental sequence YLEFIDENNNNKFKSQNSIESINKQNITYSKQIATYUQ--GGIRAAHVFVVLKLIG------YKNIKVYDASWGEWA
44586938 Environmental sequence YFNLMDS-QTHIFRSEEDIKAIIADNGIALSKAIYTYUQ--AGVRAAHANFVLQLIG-------QSEARVYDGSWGEWA
49176232 E. coli              YTELVRE-GELKT--TDSLDAIFFGRGVSYSKPIIVSCG--SGVTAAVLLALATLD-------VPNVKLYDGAWSEWG
37525471 P. luminescens       YTMLVEN-GHFKS--ETSITDIFHKQGVDLNKPVITSCG--SGVTAAVLVLGLDIIG-------KKDVYLYDGSWAEWG
9658033  V. cholerae          SAELITG-HKLKE--QASLRPLITHVLPETAQEYLFSCG--SGVTACIVLLAAYVCG-------YKNISVYDGSWTEWG
24372842 S. oneidensis        SGEVLNG-YKMKS--TTSLQAIFQALVGNKALR-IFSCG--SGITACILILASVVAG-------HKSAVLYDGSWAEWG
54302561 P. profundum         SSQLIKD-GFFID--KELLVNRFNAVS-DISQRIIFSCG--SGVTACVLALGAELAG-------RKMITVYDGSWTEWG
50905511 O. sativa            SLEMFDDAPMLP--ADSIRKKFEQAGISLSRPIVVTCG--SGVTACILALGLYRIG-------KQDIPVYDGSWTEWE
4406372  D. glomerata         SPQILDASQALP--ADSLKKRFDQEGISLSPIVTSCG--TGVTACILALGLHRLG-------KSDVAVYDGSWTEWG
39996033 G. sulfurreducens    YNEA--IPTAVSIPFAELEKNPALITASKDRLLVFYCGGVTUVLSPKSAGLAKKSGYE-------KVRVYLDGEPEWK
45360053 R. xylanophilus      YRQG--HLPGAVRYESAEQVRKLAPQKDAF-IVAY--CSNFNUHSSTRVAREIAAMGYE-------NVYDYEGGKQDWV
34557042 W. succinogenes      SRES--TIPGSIGVSDGKFKELWGRIPMDPNTKVVVFCGGYECELSHSVAGHVAMGVK-------NFMTYSGGTPEWK
48834389 Magnetococcus sp.    YAAG--RLPGAVNIPLKILEANIALIPAGQEVVAY--CRGPWCVLAFDAVARIRARGI-------KARRIQDGLPEWR
46106717 R. xylanophilus      YRAG--IIPGALSVPIERLEAYIAEIPKDQIIVAY--CRGPYCVFADEAVALIRSRGY-------RARRIQEGLPDWR
52007258 T. denitrificans     YQAG--IIPGAVNIPIDELPHHLEAIPQGQIIVAY--CRGPYCVLAFDAVATIRQAGY-------QARRLEDGFPEWK
46317410 B. cepacia           STEG--HLPGALNIPLSELDARVSEIPAGTIIVAY--CRGPYCVFAVEAVAAIRARGF-------KAARLEDGFPEWK
54029015 Polaromonas sp.      STSA--HLPRARSIPVDELKKRLNEIPKDVPIVAY--CRGPFCLMAKDAVELIRKKGY-------RAFHITDGVAEWR
48838221 M. barkeri           YEMM--IIPGANSIPLEELEKHIATIPINQIIVAY--CRGRGCILSVEAVEIIRAHGF-------KAVRLEASVQEWL
54025306 N. farcinica         YSAG--IIPGAINIPIDQLSDRIAEIPADTIIVVY--CRGEYCVFAYDAVRLITERGR-------RAVRIRDGMLEWR
3955039  S. peucetius         YLAG--IIPGAVCIPVAELTDRIGEIAKDTIIVVY--CRGEYCALAYDAVRLITDHGR-------RAIRINDGMLEWR
15607465 M. tuberculosis      YQAG--IIPGALNIPIAELADRIAEITGDRIIVAY--CRGAYCVMAPDAVRIARDAGR-------EVKRLDDGMLEWR
20807605 T. tengcongensis     YEQA--IIKGSISIPLEELPNHLNCISKDKLIVTY--CASYECTSCIEAAELLANYGF-------NVKVYRGGTKEWI
46581156 D. vulgaris          SPIPDMNDWNMAETGDKSQQDFEALIGPDKNRPLVFYCGFVKCTRSHNGAVWAQKLGYT-------NVYRVPGGIVAWK
50874889 D. psychrophila      YTSSEFAIKGAHRANPGKLDTWKSKFAKDKKIVLYCAUP--NESTSASLARKITADGFS-------SVHALKGGWREWS
48847313 G. metallireducens   AVGAINLPNDGPADIERIKQMEIPFTKKDEIIV-YCSUA--GEQASARVALVLIERGFT-------KTYVVRGGRQAVF
53760573 R. eutropha          VPIE--RIPGAVVVVHGPLDALGGQLESRIIVVYCACP--NEISAAILAERLRVAGVG-------KTWALAGGFDEWK
48768248 R. metallidurans     APIE--RIPGSIVEIKGPFDTISGHDASSQFVVYCACP--HEVSAAVLAERLRTAGVP-------NTWALAGGFDEWK
15596406 P. aeruginosa        DEPS--GIPGAIPVELNVSLKDIPGDLRDASIVIYCACP--HEISAAMLAQRLNASGFT-------RTWALAGGLDAWR
33603950 B. bronchiseptica    RDEQ--RIPGAIAVDLRAPLQDLQFDPEAGIIVVYCACP--NEVSAAQLAKKLRAAGYR-------NTFALRGGYEAWR
54029683 Polaromonas sp.      RAGGG-IIPGALVWSDLRKMASLDIPHDAHVVVYCACP--NDASAAQVAKRLMAAGFS-------NVRPLHGGIDAWE
48782625 B. fungorum          RKLDPFVIPGTQFADERQLDEIVATYPRDQKLVIYCSCP--NEISAAWMARQLNEAGFS-------DVLPLRGGMEAWR
44357259 Environmental sequence RALDPFVIPGSQFADERQLDEIVATYPHDQKVVIYCSCP--NEISAAWMAKQVNEAGFA-------DVLPLRGGMEAWR
44624393 Environmental sequence RKLDPFTIPGAQFADERQIGDIVSRYPFSQKFVVYCSCP--NEFTAALMAKRLLDAGFT-------DALALRGGLDAWR
21242497 X. axonopodis        RQLQPYTIPGAVFADERQLAQIIASVPRDRSVVIYCACP--DEVSAAWLAARVRERGYR-------DVRPLLGGLDAWR
17548617 R. solanacearum      RMSQPHRIPGALLYDVSAKDGPLEIEGPDRIIVIYCACP--NEASAVMLARTLMGRGFR-------RVRPLHGGIDAWM
47573188 R. gelatinosus       AGLDLRIIPGAWRVELSEVATHASQLPRDRIIVLYCNCP--NEASAATAAQALRAAGLP-------RVRPLAGGLEGWA
53730919 D. aromatica         VAETG-PITGATVAEHDRLLDAVGEWPKNLPIVTLCACP--EDAGAIQAARQLLNAGFL-------SVRPLKGGYEAWL
50874889 D. psychrophila      YENG--IIPGAKLIPVGQLESRLDELP--KSKPIVVYUA--IGGRSRVAVQLLAGKGFS-------KIYNLSGGINAWE
46112894 Exiguobacterium sp.  SKGN--IIKGFKNIPLQVLPTQLDKLP--KSKEVIVIGQ--SGVRSKQAVKQLKKAGYT-------QVTEVSGGMNAWR
46142555 M. burtonii          SNSG--HLEGAVNIEVSQLGTRLNEAP--ADKVILVYCR--TGVRSVRASKTLVNAGYT-------DVYNVKGGIMAWM
46198460 T. thermophilus      SAGE--IIQGAVNIPIRELPKRVGEIP--KGKPIIVYCK--VGHRSSMAMMFLRGQGY-------NVQSISGGLDGWK
48847131 G. metallireducens   SGQG--RLQGAVLIPINEVERRIGEIP--RNRPVVVYCA--VGSRSGLVAGFLSRKGYR-------EVYNVADGIVGWY
46580382 D. vulgaris          YAEG--IIPGAWLVPLADLADGVRQLP--AENPLLVYCA--IGGRSRIAAQLLAGNGFS-------KVMNLSGGFKAWN
53691784 D. desulfuricans     YRQG--HLPGARLVPVGELSDRIDEIE--RSGPTLVYCA--IGGRSRVAAQMLAGKGFK------HVINVAGGFKDWE
43887282 Environmental sequence YEIC--SLPDSKLIPLGELTSRVHEID--TADDIIVYCH--HGVRSLQAARMLKGMGYK-------KVRNLAGGIDAWA
52006282 T. denitrificans     YAAG--TIPKAKHIPLGQLQSRLSEIDKHKNKPVLVTCR--SGNRSAHACRILKKAGFE-------SVYNQAGGILAWE
53757119 M. capsulatus        SAEG--IIEGAYHIPLGKLEERASEIAQYKSKPVIVTCQ--QGTRSPSACKTLTKQGFS-------RIYEVRGGMLAWR
53729577 D. aromatica         YASG--HLPDAKNIPVAKLADRIGEIEKFKSKPIVCCA--TGVRSNKACAELKKQGFD-------KLHNLAGGVDAWV
56420601 G. kaustophilus      YAFG--IIPGAVSIPLGELENRVAEIP--KSKTIYVVCR--TGTRSDLAAQKLAEKGFD-------RVRNVIPGMSQWN
23099356 O. iheyensis         SDKG--IILGARNIPVTQVKQRIIEIR--KSKPIYLYCQ--GSSRSARAAQLLHKKGYK-------EIYQLKGGFKKWT
```

Fig. S12

```
44612327 Environmental sequence LVSAAIFRFQGQVIIFDPESGGPCYRCLYSEPPPAALVPSUAVAGVLGVLPGVVGLIQATEVIKLILEN
43231982 Environmental sequence LVSASIFRFQGQIIIFDPESGGPCYRCLYSEPPPAALVPSUAVAGVLGVLPGVVGLIQATEVIKLILDN
43257063 Environmental sequence LVSASIFRFQGQIIIFDPESGGPCYRCLYSEPPPAALVPSUSVAGVLGVLPGVVGLIQATEVIKLILEN
43262145 Environmental sequence LVSASIFRFHGQIIIFDPDSGGPCYRCLYSEPPPAALVPSUAVAGVLGVLPGVVGLIQATE IKLILEN
42946779 Environmental sequence LVSASIFRFQGQLSVFDPASGGPCYRCLYSQPPPASLVPSUAVAGVLGVLPGAVGLMQATEVIKLVLVE
2950364  Synechococcus sp.     NVYGSIFRFEGQATVFNYEDG-PNYRDLYPEPPPPGMVPSCAEGGVLGILPGHIGVIQATETVKIILGQ
22299946 T. elongatus          NVYGSIFRFEGQATVFNYEGG-PNYRDLYPEPPPPGLVPSCAEGGVLGILPGHIGVIQATETIKIILGK
37522981 G. violaceus          NVYGSIFRFEGQSTVFHYEDG-PCYRCLYPEPPPPGLVPSCAEGGVLGILPGVIGVIQATEAVKIILGI
45545541 R. xylanophilus       NVYGSIFRFEGQASVFWAEEG-PCYRCLYPEPPPPGLVPSCAEGGVLGILPGAIGVIQATETVKLILGI
53796678 C. aurantiacus        NVYGSIFRFEGQATVFSARDGGPCYRCLYPEPPPPGLVPSCAEGGVLGVLPGVIGTIQATEVIKLLTGI
53610625 A. vinelandii         NVHASIFRFEGQASVFYAKEG-PCYRCLYPEPPPPGEIPGCAEGGVLGVLPGLLGVVQATEVIKRILGI
48834858 T. fusca              YVWGSIFRFDGQVSVFWNEYG-PNYRDLYPEPPPPGMVPSCAEGGVLGVLCATIGSVMANEAIKLITGI
54659256 C. hominis            LLDSGTEGFNGHSRHIIPGET-SCVEETMGLNVQDTNFPLCEIKEFPRTPIHGIYAVFIYEEDEQD--
21464561 A. thaliana           MVDGGTEGFKGHARVILPGVT-PCLEETIYLFPPQVKFPLCTLDETPRNAAHGIEYAHLIQWETVHRGK
6855414  L. major             LLESGTLGTKGVMQPAIPFVT-ESYSS--SYDPPEKGIPLCTLKNFPNAIEHTIQWARDLFHILFYSVP
17539268 C. elegans            LLESGTMGTKGNTQVVYPYLT-ESVSS--SVDPPEKEIPVCTLKNFPNEIQHTIQWAREQFETFFAQPG
```

Fig. S13

```
12642418 C. hydrogenoformans    ITKKALVIGGGIAGIQAALDIADAGYQVILVEKEPTIGGKMAHVDKTFPTLDCSSUISTPKMAAAAHDP
15669884 M. jannaschii          VDKSGLIIGGGIAGIQAALDLGDQGYKVYLVEKEPSIGGRMAQLAKTFPTDDCALUILAPKMVSVANHP
45359260 M. maripaludis         VDKSGMVIGNGIAGIQSALDLGDQGFKVYLVDKDESIGGRMAQLAKTFPTDDCAMUILAPKMVSAANHP
20093689 M. kandleri            VENSVLIIGGGIAGIQAALDLADQGFKVYLVEKEPTIGGNMARLAKTFPTDDCAMUILAPKMVQVGNHP
15679380 M. thermautotrophicus  VDDYALVIGGGVAGIQTALDLADWGFKTYMVEKRPSISGRMFQLDKTFPTLDCSMCILAPKMVDVGKHD
46580807 D. vulgaris            VTKRVLVIGGGVAGIQAALDCADGGVEVVLVERESTIGGKMAKLDKTFPTVDCSSCILGPKMVDVAQHP
23475491 D. desulfuricans       VTRRALVIGGGVAGIQAALDCADAGIEVIMVEREPSIGGKMAKLDKTFPTVDCSSCILGPKMVDVAQHP
48847469 G. metallireducens     VTKRSLVIGGGIAGIQAALDIADAGHQVVLVEREPSIGGHMAQLSETFPTLDCSQCIMTPKMVDVANHP
39995201 G. sulfurreducens      VTKRALVIGGGIAGIQAALDIADAGHKVVLVEREPSIGGHMAQLSETFPTLDCSQCIMTPKMVDVANHP
21674069 C. tepidum             ITRRALVIGGGIAGIQAALDIAGAGREVILVEREPSIGGHMEQLSETFPTLDCSQCILTPRMVEAIQHP
20091692 M. acetivorans         ASRNVLIIGGGVAGIEAALNLAEAGFPVTMVEKESTIGGKMALWNEVFPTNDCSICVLAPKMTEVQNHP
21226158 M. mazei               ASRNVLIIGGGVAGIEAALNLAEAGFPVTMVEKESTIGGKMALWNEVFPTNDCSICVLAPKMTEVQNHP
48838060 M. barkeri             ASRNVLIIGGGVAGIEAALNLAEAGFPVTMVEKESTIGGKMALWNEVFPTNDCSICVLAPKMTEVQNHP
53731435 M. burtonii            ANKDVLVIGGGVAGIEAALTLADSGTHVYMVEKEPTIGGKMALLNEVFPTNDCSICVLAPKMTDVQNHT
11498837 A. fulgidus            IEKSVAVIGGGVAGIEAALTLADSGIKVYLIEKNPTIGGHMATLNEVFPTNDCSICILAPKMSDVWNHE
```

Fig. S14

```
83952997 Sulfitobacter sp.       VRHYTKPTFAKPCPDTLRHQIRDECDFLVEGLADUGSCTTCSMHDTVWFEIQGIPSVSIASSVFGQAAETQRKALGMEGARYVLVPHPI
43365817 Environmental sequence VKRYKKPTFARVAPKELNQQISVECDAVIEGLADUGSCTSCSLHDIMDLEGRGIPSGFIASCEFEQAANAQGKSLGISPAR-VFVAHPI
85762167 Environmental sequence VVRAIKSNYSAPADAHIVEEA-ENWDAVIAGIGDUGSCSSCSLHDAVKGEKIGKPSVGIMTTKFTSAAKLMAKVLGADGYPFVVIDHPI
60069660 Environmental sequence VVKRTKRSASMPAPEAILGELTKECDLVITGSGDUGSCTSWSIHDSVEATKRGVAALTVCSTAFVTLGRAQASALGSAGLPIAVVPHPF
85750743 Environmental sequence -YYVKPHGFSNDAPEKLLNQIAEENDIVLTAIGDCGSCCSCCIRDAVSLEDRGIPSAAIITTEFVNETRLTRKAIGMPDLRPIVIEHPV
2622661  M. thermautotrophicus  LRFLWSEKPAGAPASEDQIRAGEGDLCILALGDCGSCTTWVILDAIRLEGMGVPTISICSDTFREYAEKLAAAHGMPGLRIVEIEHPV
```

Fig. S15

```
85801586 Environmental sequence RYVNTGNVHGQLDPCGVK--KNPLGGLSRKLVKTQEYRD-AGEDPYILDAGDLFFSTPNLHEGNLRSE--MHRANSILKGYEKIGCDAIN
42966760 Environmental sequence KIVLTGSVHGQLDPCGVK--KNPLGGLSRKYVVIKQMRD-DGQDPIVLDAGDMFFSITSLNKNNLESE--KHRCETILSGYEKIGCDGLN
44228873 Environmental sequence TLIMSGSVHGQLDPCGVK--KNPMGGLSRRYVKIKEMKE-EGKDPIILDAGDFFFSITNLNQNNLKSE--EYRAGAILEGYDKIGCDAIN
85772737 Environmental sequence TLLVSNNVRAQLDPCGVK--KDPLGGLSRKSTVKQLKE-DGKNLVILDAGDLLFSAAIITGVNEQAD--KLRANAILKGYERIGCHAVN
50874889 D. psychrophila        RYIFSSNLLGQIGNTGVA--INKLGGLSRHASLLNKLTADSSKIPRIVESGNLLFKTSQSSKNTLAQK----KARKIARARRKNGYTAIC
43122652 Environmental sequence SIVSTTNVYSEFYDCGCP--KNPLGGLARKTFFLKNMMP--DRDSILIDAGNALFDSNQINPDNLSIENKRFKARNFVKTLEFLGQDVVN
44504140 Environmental sequence SIVSTTNVYSEFYDCGCP--KNPLGGLARKTFFLKNMMP--GRDSFLIDAGNSLFDSNQINPDNLSIENKRFKARNFVKTLEFLGQDVVN
32445059 R. baltica             TLFVTGQQHGYIEPCGCTGLENQKGGVARRMTFLKQLRE-KGWEMAPIDAGNLVRRYGRQSEIKFHRS---------LEALRKMDYVSVG
87311211 B. marina              ALFLTGEQHGYIEPCGCTGLTNQKGGLNRRFTFLQQLREERGWDVVAIDAGNQVRRFGRQAEIKFQTT---------ANAIKQMKYDAIT
```

Fig. S16

```
85788322 Environmental sequence ----------------------------------------VIAERLKEGLGIYEIDAELLKAAEPDLLITQAICEVVAISSRQVAEVSVGLA-KEPDILSLDPL HIGDVVEDLNRVGRATG
85800620 Environmentla sequence ------------------------------------------------------------------STQAICEVVAISSRQVAEVSLGLA-KEPEILSLDPL HIGDVVEDLNRVGRATG
88802823 P. irgensii            VGLQEYLHGITFECPK-IALVEKQVAVRYKLEGQALTSEEINTIFSKTKAEGGTIYYVDELVLESIAPDVIFTQDVCDVCQLDTACVATSAYKLQ-KVPKLISITPN SLEDVFENALTIADAIG
55378797 H. marismortui         LGVEP--VGVSHECDHPPAAREKPSVNRSRVDPTA-SSGEINEQVAAVEESD-GVYAIDRETLAELEPDLIVTQGVCDVCAVDHVVVAEAVEELG-LDTDVLTLDVH SLDGLFESIHRVGAAVG
76800687 N. pharaonis           LGVDP--ATTSHECDYPPSVADTPTVVESRIDADA-DSETIDEQVHEAETDG-GVYAIDRETLAAVDPDIVISQGICEVCAVDTVQVEAAIADLG-LDCRLVTIDPH SVDDILGDIERIGAALG
84497447 Janibacter sp.         IGAGIAVVGVTFECDYPEQARSRRIVSTSAVPEGL-SPAGIDAFVVDAVTRGEDLYHLDAGALADLDADLVVTQDLCAVCAIDVTLVDEALAHLG-CTADVLTFDPH TLDEVFASIVELGRATG
71368962 Nocardioides sp.       IGAGPDVVGVTFECDYPAEARARTVVSTSAVPSGL-TPAEIDAYVGALSRGEDLYHIAAIALAGLDADLVVTQDLCAVCAVDVSVVDDALAHLG-CTAEVLTIDPH TLDEVLDSILLLGRVTG
32447814 R. baltica             LGIRDQLVGVTHECDYPSGVDRIAKVTQTLIPHD-ATSGEIDTMVRERLQTERALYSLKVPVVESLAPDLIVTQALCDVCAVAESEVNAAACSLPG-KPRVVNLEPT SLSEVFDCITLVGKAAN
83852439 O. alexandrii          LDLTIDLVAVSHGCDWPQVRDLPALTS-ALFEA-QTARDIDEAVRAKLQGQDAIYALDAQALAALKPDLVISQALCDVCAVSGAEVSTVLSDLDP-QPTLVNLEPY SLRDILETITMIGVAAG
88812304 N. mobilis             LGRQDLVGVSHDCDFPAAVEAKPRVTSCEITDTPLDSAAIDRWVSERLARGEPLFHLNRSRLESLRPELILTQSLCQVCAPSAERVTEVAAALPE-PPRVLNLEAS TLAGVLDTIEQVAVALG
17129888 Nostoc sp.             LGLSDAIVGRSHECDYPPELQNRPICTQARLNSQA-ASSEIHHEVNNLLQSALSIYQVKLEVLEQLQPTHILTQDQCDVCAVSLKDVEKAVKELTQSSPQIISLQPN LLEDVWQDIERVAHIFG
75702880 A. variabilis          LGLSDAIVGRSHECDYPPELQNRPTCTQARLNSDA-SSSEIHHEVNNVLQSALSIYQVKLDVLEQLKPTHILTQDQCDVCAVSLKDVEEAVKELTQSSPQIISLQPN LLEDVWQDIERVAHIFG
23125528 N. punctiforme         LGLVNAIVGRSHECDYPPELLDRPVCTQARLDSNA-SSSQINDEVNDFLQSALSIYEIKTDVLEQLQPTHILTQDQCDVCAVSLDEVEKAVATLIDSKPQIISLQPN ILQDIWADIERVGNAFE
71673789 T. erythraeum          LGLTKAIVGRSHECNYPPELQNLPVCTKPKFNPDG-NSREIHERATELLQSALSVYLVELEILEKLQPTHIITQAQCDVCAVSLADVETAVKSLTKTNPQVISLEPN LLAEVWADIERVANILG
35211894 G. violaceus           LGLASRLVGRSHECDYPESVRVLPVCTRPKFDPVG-TSREVHERVSKLLATALSVYEVDLAMLEQLAPTHILTQAQCEVCAVSLSEVEAAVAGLTGGQPHILSLQPN VLGDIWEDIARVGTALG
67924076 C. watsonii            LGLTDALVGRSHECDYPSYVRDLPICTTARLDITR-SSGEIDQDVMTLLQQALSIYNIELEILQKLQPTHIVTQDQCDVCAVNFATVTEAISQLTNSKPQVISLQPN LLKEVFADIERVAKTFD
22295701 T. elongatus           LGLLPFLVGRSHECDYPPEVKALPVCTRARLNAQQ-SSLAIDRVVQELLRAALGIYDEELATLQALKPTHIITQDQCDVCAVTLPDVQRAIADLLEPPPQLISLQPH GLEDIWEDIRRVGLTIG
83815755 S. ruber               LGHGDDLVGRSHECDTPPGVEAPALTAPKVPLEG-SSREIDDQVRALLDEAMSVYDVDVERLNALRPDLILTQSQCEVCAVSLPTVEQAVADRIEGSPTVVALEPR SLADVRDDIRRVAGALG
84788069 E. litoralis           LGLGQNLVGRSHECDYPPSVKGLPVVTSTKLEKGL-TSCQIDDRVQGIVRQGLSVYEVDTALLRDLRPDLILTQSQCAVCAVTPADLEEALSEWTGTAPVLLSLAPD TLEDVWGDFAKVAEAAG
16413520 L. innocua             LDLGDKVKDVSAYDDYPKEAQKLEK-------------------------VT-STTVDTEKIALKPDLVLGHESMLATEKDAYQLLKDAGINV-----FVVPDAT NLKEVENSIETIGKLTG
71845400 D. aromatica           LGEQDRIVGISGYTVRPPEARKEKPKVF------------------------AFTSGDIDKILATQPDLVLTFSDLQSEISRDLIKAG---------IPYYAFNTR SIEDILGMVETVGRLVG
47572949 R. gelatinosus         LGEAHRIVGISGFTVRPPRARQEKPKVS------------------------AFTSAKLEPILALQPDLAIGFSDIQADIARELIQRG---------VEVWISNHR SVDGILGYVRRLGAMVG
68213330 M. flagellatus         LGEQARIVGISGFTTRPAIARQEKPKIS------------------------GFSTAKLEKILAVKPDLVLAFSNLQADIAADLVRAG---------VEVHVFNQR SVQGILDMVATLGALTH
86160446 A. dehalogenans        IGAGDLVVGVSGFTVRPPEARR-KPRVS------------------------AFLSADGDAIAALRPDLVVGFSDLQADVARDLVRRG---------IPVLVTNQR SVAEILQTLRLVIAAVG
45659220 L. interrogans         LGIEERIVGISAYTVRPLRAKKEKPKIS------------------------AFINGNIKRIKELKPDLVIGFSDIQSDLAKNLIEEG---------LNVLVTNQR TILEIFDTLSLLGSIVG
56381325 G. kaustophilus        LGRLDDVAVDDSSDWPPEVKQLPKVGP-------------------------DLRIDNDQVEALKPDLVVASLSVPGMERNVEELKRRGLPHL-------VLAPN SLDDIANDILRLGEALG
10176026 B. halodurans          LGLVDQLVGVDDFSDWPTVTRSLPRVGP------------------------DLRIDNDKVEALQPDLVVASLSVPGMERNVEELQKRDIPHV-------VYNPH SLKEIGECLKDLAKRTG
66797343 D. geothermalis        LGVAAWVVAVDQHNDAP-GLDQAVRVGP------------------------DLNIDVQAVQAARPDLVLASLSVPGMERVVEAVRAAGLKTL-------VLDPL SVRDVLQDIRLVGEAVG
```

Fig. S17.

```
44183053 Environmental sequence HPAPLAYIVAGIGFULLTQLKRYASMRKVGITSAKYHVELDYYLKGSVKQGTYENKVTEVRSDFT
43725511 Environmental sequence FPAPLTYVASGIGFULLTNLKRYASMKKISIKSAQYKIELDFYLFGSIVDENIESGVSEVRSFFE
43560497 Environmental sequence YSPPMPWLATAIGFULLTQVARYAHMLKMEIKSGKCHVEGDYLLHGSVIKGTVNVDHQGFRTHLE
26247773 E. coli                GTNPEELIGAAHAACFSMALS-LMLG-EAGFTPTSIDTTADVSLDKVDAGFAHTKIALKSEVAVP
24052078 S. flexneri            GTNPEELIGAAHAACFSMALS-LMLG-EAGFTPTSIDTTADVSLDKVDAGFAHTKIALKSEVAVP
50121147 E. carotovora          GTNPEELIGAAHAACFSMALS-LMLG-EEGHKPESIDTTADVSLDKVDGGFAHTKIALHSTVTLP
44496725 Environmental sequence GSNPEELVGAAHSACFSMALS-LALG-DAGFTADKIDTKATVSLDEVDDGFAVTHIALEVKARTS
28850626 P. syringae            GTNPEELIGAAHAGCFSMALS-MILG-DAGLKADSIDTSAEVSLDQVEGGFAHSAVHLVLKAKVP
56180660 I. loihiensis          GTNPEELLGAAHAGCVAMAFS-LMLG-ESGYEPDSIDAKAEVSLEEDGDGFSHAKIHLKVKASIP
44014688 Environmental sequence GSNPEELIGAAHASCFTMAFT-LILS-KAGFTPAQLDTQAVVALEQQGDAFVHPSIALTLKAAIP
43985334 Environmental sequence GTNPEEQIAAAHASCFTMALS-FALA-KQGFSKGTLESKVDITLVKDGDSYTHTKSELRLDAKVP
42522952 B. bacteriovorus       GTNPEELIGAAHSGCFAMALS-GALA-KKGFNAESLDVSATVTLEKSGDGFVHQSSKLKLRALVP
48855516 C. hutchinsonii        GTNPEELIAAAHAGCFTMALS-FQLS-GANFTPTKLATEASITVVQENGGFKFKSIHLHLEATVP
21243642 X. axonopodis          GTNPEELIAAAHAGCFTMALS-AQLT-EAGFPPASLDTRADVDLSME-GGPQLSQIRLKLKAVVP
46199563 T. thermophilus        GTNPEELIAAAHAGCFSMALA-ASLE-REGFPPKRVSTEARVHLEVVDGKPTLTRIELLTEAEVP
15806548 D. radiodurans         GTNPEELLASAHAGCFTMQLS-ALLA-EHGHEIKALDTDATCELVKDGPGFKTNHMHLRVRAQLT
17549328 R. solanacearum        GTNPEQLFAAEYSACFLGALK-FVAARDKLRLPADTSVQGSVGHGAIPNGFGHE---VDLAISLP
50086039 Acinetobacter sp.      GTNPEQLFAAEYSACFLGALK-FVATRDKFNIPKDAYVEGDVGHGPIPNGFGHE---VKLINVHLP
56178201 I. loihiensis          GTNPEQLFAAGYSACFIGALK-FAAEQEKVKLPKDTEVKAEIGHGQIEGGFGHD---ATLNIHLP
21220863 S. coelicolor          GTNPEQLFAAGYSACFQGALG-VVARQEGMDISG-STVTAKVGHGKNDDGFGHI---VEHSNEIP
16263744 S. meliloti            GTNPEQLFAAGWSACFIGALG-LAASKHKLTLPAETAVDAEVDLAKSDGGFFLQ---ARLAVSLP
```

Fig. S18

```
23452038 C. reinhardtii      GAELATFALGUFWHPEASFA--NVPGVVKTRVGYTGG-SRPNPTYESVGAG-DGHTEATRVWFDPAIISYEDLLKQFREHDP------TQKSKCQYKSAVWYHSEAHGT
34792792 S. purpuratus       DLETATFALSUFWFPEAQFG--CAPGVVRTKVGYTGG-TKKPPTYY----SLGDHTETVQIEFDRTKTSYEKLLKKFWANHDS------TTCHKNQYKSAIFYHDEKQH-
89467751 P. lividus          SLETATFALSUFWFPEAQFG--CAPGVVRTKVGYTGG-TKKPPTYY----SLGDHTETVQIEFDRTKTSYEKLLKKFWANHDS------TTCHKNQYKSAIFYHDEKQK-
66511346 B. floridae         --ETATFALSUFWFPEAQFG--CAEGVIRTKVGYTGG-TKVMPTYH----ALGDHTESVQIQYDPKQTSYANLLKIFWQNHDP------SACVSRQYKSAIFYHNENQK-
89138292 A. millepora        ETKTATFSLSUFWFPEAQFG--CADGVVRTKVGYTGG-SKPFPTYY----SLGNHTETVQLEYDPNRTIYRDLLSMFWKNHNA------TANHKPQYKSAIFYHDEGQK-
51995432 H. magnipapillata   QLRKATFAMAUFWFPEAQFG--CAPGVVTTKVGYTGG-KHPFPTYH----NLGDHTETVEIQYDPSETNYSALLDMFWKYHDS------TAVNKTQYKSAIFYHDEEQL-
49561997 B. microplus        PVKKATLALAUFWFPEAQYG--CAPGVVRTRVGYTGG-TTKNPTYR----NLGDHTETVQLDYDPTKTDYKTLLDMFWGFHDP------TACHKRQYKSAIFYHDKEQK-
21640047 A. variegatum       AVKKATFALAUFWFPEAQYG--CAPGVIRTRVGYTGG-TSKNPTYR----CLGDHTESVQLDYDPTETDYKTLLNMFWDFHDP------TACHKRQYKSAIFYHDKEQ--
63523529 I. scapularis       AVKKATFALSUFWFPEAQYG--SAPGVIRTRVGYTGG-TTKNPTYQ----NLGDHTETVQLEYDPIATSYDRLLEMFWGFHDA------TACHKRQYKSAIFYHDQEQK-
71547479 S. fumaroxidans     AVETATFAMGUFWGPDARFG--CTDGVLRTRVGYAGG-RKKNPTYT----DLGDHTESLEIVYDPAMYSYERLLEVFWAEHNP-S----VPPWSKQYKSIIFHHDERQK-
66861845 B. natans           PANTATFGMGUFWGPEVTFA-PKNVPGVLSTRVGYAGG-KKKNPTYH----SLGDHTEVIQITDPAKVSYSKAYDIFFSEHSFR-----SKPYCRQYKSEAWYQNEEQKE
55234261 A. gambiae          PFEKATFGMGCFWGCDSLFG--ATKGVLRTRVGYAGG-STESPAYK----KVGDHTEVIEIDYDPQTISYNDILDLFWNNHEYGL----TTRMKRQYKSLILYHNEQQR-
33089114 A. mellifera        QAKRATFGMGCFWAGDCLFG--VLPGVIRTCVGYAGG-QKESPTYK----NLGDHTEVDIEYNPNIVSYSQLLALFWQNHEYGL----TTKIKRQYKSLILYHDEEQK-
76258265 C. aurantiacus      PLETATLAGGCFWCLEAVLD--QVSGVKDVVSGYTGG-YVPNPTYRRVCDGNTGHAEAVQIQFDPEQISYRELLEIFFSIHDPTTLNRQGADVGTQYRSAIFYHSEEQQ
66799168 D. geothermalis     QTQQATFAGGCFWCTEAVMK--DVRGVTRVESGYTGG-HVPNPDYAAVCSGETGHAEAVRVTFDPAQVSFRDLLKLFFATHDPTSLNRQGADVGTQYRSAVFPLNEEQER
88933851 Dehalococcoides sp. MNEIAVFGGGCFWCMEAVFS--SLSGVIKVESGYAGG-TLPNPTYQQVCAGDTGHAEAVIRLEFDAALISYTALLDIFFQMHDPTTLNRQGEDVGSQYRSVIFYTTPSQKA
78171127 C. chlorochromatii  TTQTAVFAGGCFWGVEYHFS--KLKGVLSVTSGYTGG-AIENPTYQQVCSGKTGHAEAVEIVFDAAQVSYETLAKLFFEIHDPTQVNRQGPDVGTQYRSALFYANEEQRR
51470621 A. punctata         MHQRAVLAGGCFWGMEDLIR--KRPGILSTRVGYTGG-DIPNATYRNHGT----AEGIEIVFDPTVTSYRHITEFFQIHDPTTLNRQGNDRGLSYRSGIYYVDEEQKR
33150552 H. sapiens          GTQMAVFGMGCFWGAERKFW--VLKGVYSTQVGFAGG-YTSNPTYKEVCSEKTGHAEVVRVVYQPEHISFEELLKVFWENHDPTQGMRQGNDHGTQYRSAIYPTSAKQME
55251370 D. rerio            GLQMVLFGMGCFWGAERKFW--RQKGVYSTQVGYSGG-YTPNPTYEEVCTGKTGHTEVVRVVFEPQKIKFSELLKVFWESHNPTQGMRQGNDVGTTYRSGIYTNTQEQLE
78709015 O. sativa           GNEFAQFGAGCFWGVELAFQ--RVPGVTRTEVGYSQG-NLHDPTYEDVCTGATYHNEVVRVHYDVSACKFDDLLDVFWARHDPTTPNRQGNDVGTQYRSGIYYYTPEQEK
7580480  L. esculentum      GLEFAQFAAGCFWGVELAFQ--RVGGVVKTEVGYSQG-NVHDPNYKLICSGTTEHAEAIRHQFDPWVCPYSNILSLFWSRHDPTTLNRQGNDVGKQYRSGIYYYNDAQAQ
21593143 A. thaliana         GQQFAQFGAGCFWGVELAYQ--RVPGVTKTEVGYSHG-IVHNPSYEDVCTGTTGHNEVVRVQYDPKECSFESLLDVFWNRHDPTTLNRQGGDVGTQYRSGIYYYTDEQER
1136793  B. napus            GQQFAQFGAGCFWGAELAYQ--RVPGVTKTEVGYSHG-FVDNPSYEDVCSETTGHNEIVRVQYDPKEVSFESLLDVFWKRHDPTTLNRQGNDVGTRYRSGIYFYTDEQEK
```

Fig. S19

```
7706511  H. sapiens          SRSKYAHSSPWPAFTETIHADSVA--KRPEHNRSEALKVSCGKCGNGLGHEFLNDGPKPG--QSRFUIFSSSLKFVPKGK
55730640 P. pygmaeus         SRSKYAHSSPWPAFTETIHADSVA--KRPEHNRAEALKVSCGKCGNGLGHEFLNDGPKPG--QSRFUIFSSSLKFVPKGK
27807643 M. musculus         SHSKYAHSSPWPAFTETIHPDSVT--KCPEKNRPEALKVSCGKCGNGLGHEFLNDGPKRG--QSRFUIFSSSLKFVPKGK
29648559 D. rerio            SRSKYEHSSPWPAFTETIHKDSVS--KQEE--RWGAYKVRCGKCGNGLGHEFVNDGPKHG--LSRFUIFSSSLKFIPKVK
57335062 Suberites sp.       STKKFEHSSPWPAFTETVHSNSVS--KYNES--TSALKVSCGKCGNGLGHEFLNDGPNKG--QSRFUIFSSSLTFKDKGE
56314788 Azoarcus sp.        SEHKFDAGCGWPSFWTAAAPENVE-TAEDRSHFWQRTEVLCHEGGAHLGHVFEDGPQPTG---LRYCINSASIRLEPEAS
85715266 Nitrobacter sp.     SDAKFDSGCGWPSFTAPAVDSHID-EEHDTSHGWSRTEVLCSRCSGHLGHVFNDGPGPTG---LRYCINSAALKLEPK--
68246156 Magnetococcus sp.   ADAKYDSGSGWPSYFQAISEHAIR-AETDLSHGWRRVEVLCSSCDAHLGHLFDDGPQPTG---LRYCINSLALQLNPKPT
71661450 T. cruzi            SEMKFRCGCGWPAFWDCVPGAVRE-EPDSDG---VRTEIVCNACNSHLGHLFRGEGLCNPPPNERHCVNSTSIRFQPSP-
92877819 M. truncatula       STTKFNSGCGWPAFYEGVPGAINR-HADPDG---MRIEITCAACGGHLGHVFKGEGFPTP-TNERHCVNSISLKFAPANS
91216521 P. torquis          SQSKFESNCGWPSFDSSVEGSIEY-IKDSKFG-WQRIEILCSNCGGHIGHIFDDGPTETG---KRYCVNSLSIDFKNK--
92907076 Mycobacterium sp.   SSEKFESHCGWPSFFDPADSDAVI-LRPDDSLGWRRVEVVCANCHSHLGHVFEGEGYPTP-TDQRYCINSISLRLVPAGG
32397214 R. baltica          AKDKFISHCGWPSFDDEIEGAVKR-QRDAD---GRRIEIVCANCDGHLGHVFHG-ERLTP-KNERHCVNSISMKFIPEGK
56127980 S. enterica         SHTKYDSGCGWPSFYQPVSEEAIR-YIDDFSHGWQRVEIRCGNCDAHLGHVFPDGPQPTG---ERYCVNSASLAFSDEKN
91210995 E. coli             SQTKYDSGCGWPSFYEPLSEESIR-YIKDLSHGWQRIEIRCGNCDAHLGHVFPDGPQPTG---ERYCVNSASLRFTDGEN
46133494 H. influenzae       SNDKFESGCGWPSFTKPIIKDVVHYETDNSFN-WQRTEVLSRAGNAHLGHVFDDGPKDKG--GLRYCINSASIKFIPLAE
57637573 S. epidermidis      SEDKFESNCGWPSFSKALNDDEIVELVDKSFG-WIRTEVRSEKANSHLGHVFNDGPKEKG--GLRYCINSAAIQFIPYDK
46190343 B. longum           SRDKFDSGCGWPAFSRPIAGDLLTEHEDHRIPGRDRIEVRTSDTQIHLGHVFTDGPADRG--GLRYCMNSAALRFVPRSR
```

Fig. S20

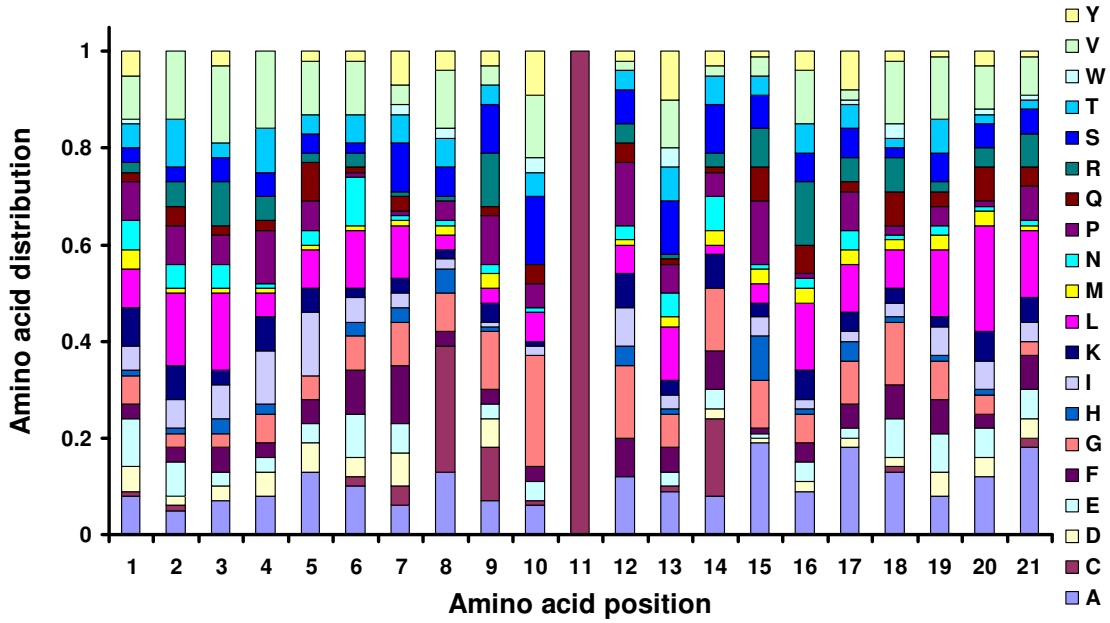
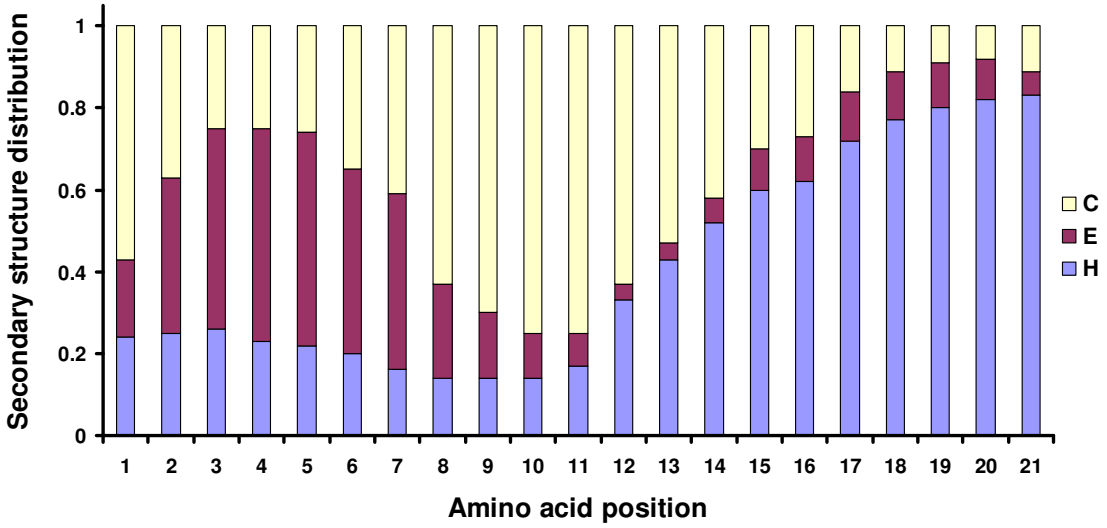
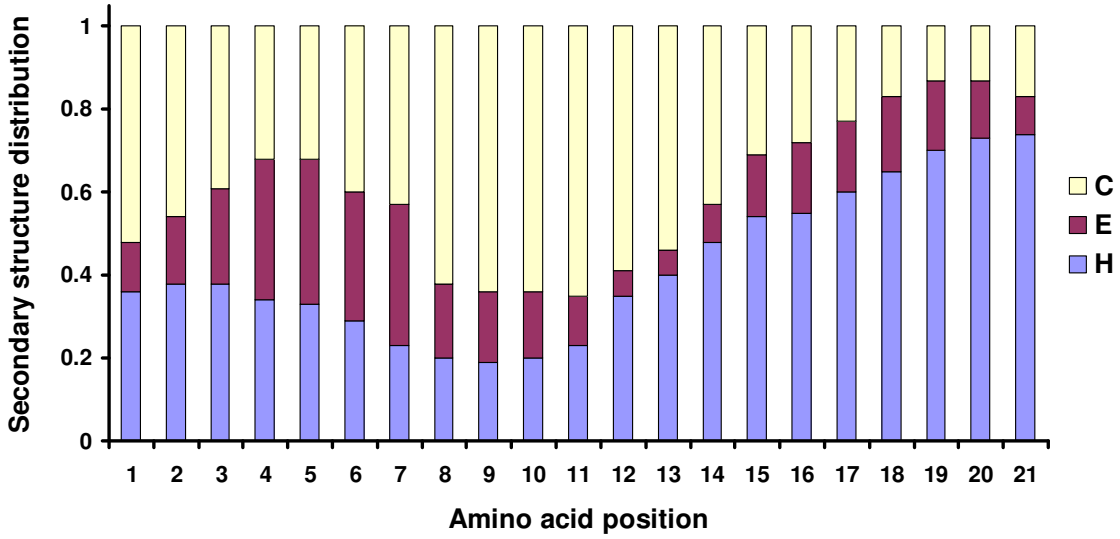

Amino acid distribution
0
0.2
0.4
0.6
0.8
1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
Amino acid position
Y
V
W
T
S
R
Q
P
N
M
L
K
I
H
G
F
E
D
C
A

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