

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

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*. 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.



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 AACY000000000 (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 than 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 than 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'-agatcgtgacatatggctgcttcccagacgctg-3' and 5'-gcgctggccctcagctagcagttttcctcttggccacagcag-3'. The product was ligated into the NdeI/XhoI sites of pET15b. The following primers were used for site-directed mutagenesis: Cys157Ser, 5'-atgatattgcatatccaactctgttatcaacctgttct-3' and 5'-aggaacaagggtgataacagagttggatgacaatatcat-3'; and Cys207Ser, 5'-agttttatgggggaatccctgggagcgctctg-3' and 5'-cagagcgctcccagggattccccccataaaactttg-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 μ L final volume) that contained 5 μ g of wild type AS3MT or Cys157Ser or Cys207Ser mutant AS3MT, 1 mM AdoMet, 3 μ M [73 As]-iAsIII, 0.5 μ M *E. coli* thioredoxin, 0.26 μ M rat liver thioredoxin reductase, 300 μ M NADPH in 100 mM Tris/100 mM sodium phosphate buffer, pH 7.4, were incubated at 37 $^{\circ}$ C for 60 min. Reactions were stopped by addition of 16 μ L of 30% H₂O₂ 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-dehydroquinase dehydratase (PDB 1GQO), which catalyzes dehydration of 3-dehydroquinase to 3-dehydroshikimate in the third step of the shikimate pathway; however, the low level of similarity and the absence of Cys in appropriate dehydroquinase 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 Fe³⁺ 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 Fe³⁺ to Fe²⁺. 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 CDD25924	325/37	11/0	2/0
OsmC-like protein	Thiol peroxidase	CDD11475 CDD11476 CDD29457	254/3	11/0	0/0
Formylmethanofuran dehydrogenase	Oxidation of formylmethanofuran	CDD12595	100/3	13/6	0/0
F420-reducing hydrogenase alpha subunit	Hydrogen oxidation or proton reduction	CDD10549	25/4	17/4	0/0
F420-reducing hydrogenase, delta subunit	Hydrogen oxidation or proton reduction	-	18/8	12/3	0/0
Methylviologen-reducing hydrogenase	Hydrogen oxidation or proton reduction	CDD13801	77/1	11/4	0/0
NADH oxidoreductase	Electron transport	CDD29449	151/70	3/0	0/0
Formate dehydrogenase alpha chain (FDH)	Oxidation of formate to carbon dioxide	CDD12612 CDD17412	8/6	0/0	0/0
Glycine reductase selenoprotein B	Glycine reductase	CDD11108	311/23	0/0	0/0
Arsenate reductase	Reduction of arsenate				
Proteins with predicted redox function and catalytic redox-active Cys					
SeIj	ADP-ribosylation	-	36/0	0/0	5/10
SeIk homologs	Function not known	-	0/0	0/0	16/57
SeIs 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 MoeA 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

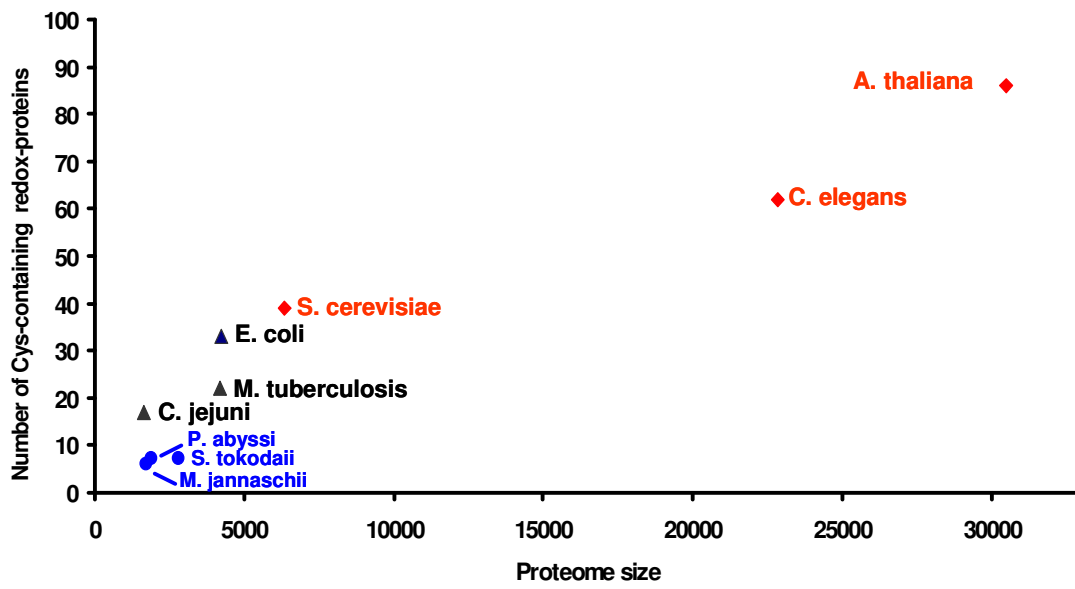
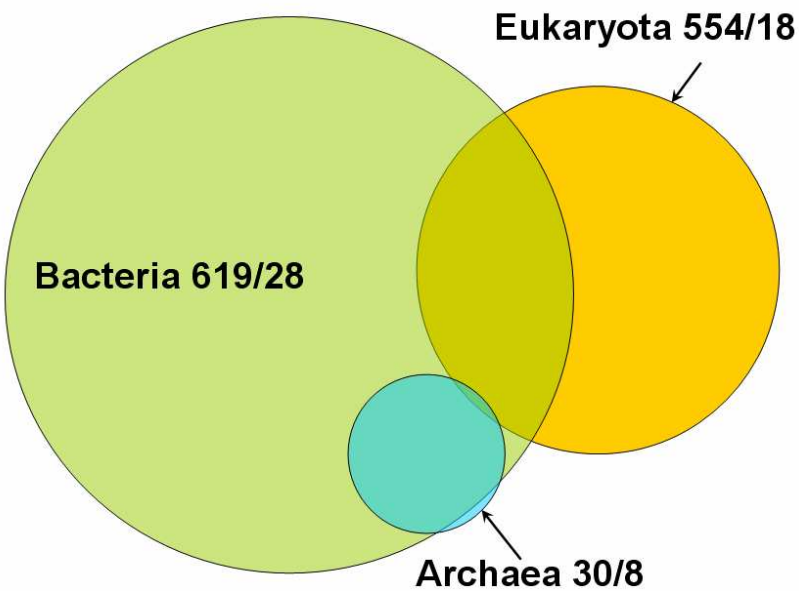


Fig. S2

a)



b)

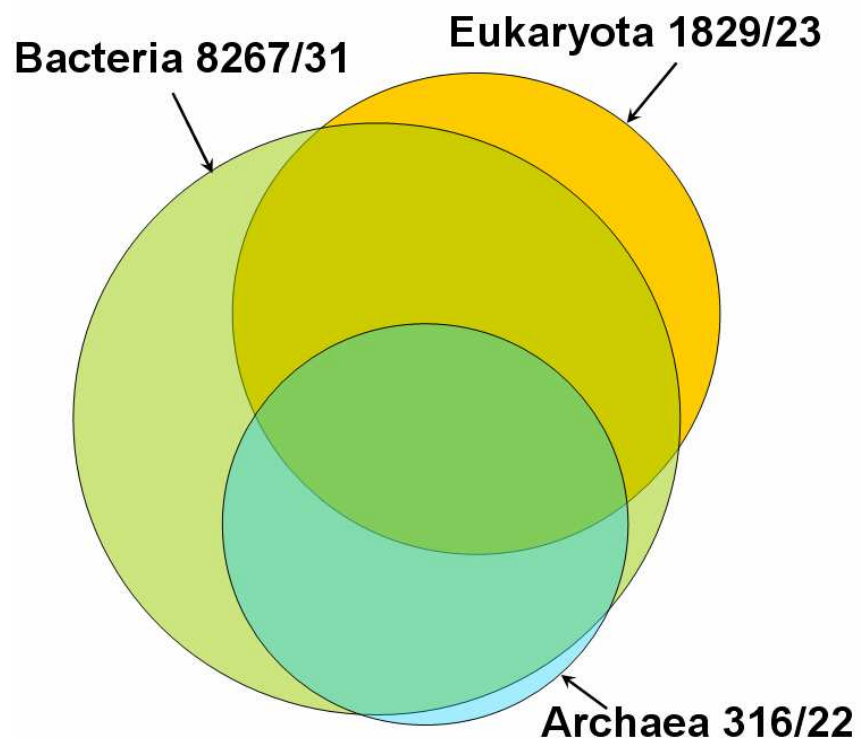


Fig. S3

```

42897919 Environmental sequence -MNKLTKSELNNIYPLGENRTG...P...I...D...F...M...A...D...W...G...F...C...R...M...F...E...V...I...N...E...T...Q...Q...Y...E...G...-...K...I...O...Y...K...I...D...I
44346764 Environmental sequence ---ELTYQNLGDGELVNEEDLTN...N...T...I...V...F...W...A...D...W...G...F...C...R...Q...E...L...F...V...L...E...A...N...I...D...K...L...I...E...Y...D...V...A...L...A...H...S
42830445 Environmental sequence ---ELTYQNLGDGELVNEEDLTN...N...T...I...V...F...W...A...D...W...G...F...C...R...Q...E...L...F...V...L...E...A...N...I...D...K...L...I...E...Y...D...V...A...L...A...H...S
44454992 Environmental sequence SNKLELVIEDTTTIVVNDLTN...K...T...I...V...F...W...A...D...W...G...F...C...R...Q...E...L...F...V...L...E...A...N...I...D...K...L...I...E...Y...D...V...A...L...A...H...S
44323676 Environmental sequence SDSSFSFINSDLVDSSEDFTN...K...T...I...V...F...W...A...D...W...G...F...C...R...Q...E...L...F...V...L...E...A...E...L...E...N...L...Q...E...Y...D...I...A...L...A...H...S
44640430 Environmental sequence PGLYLFNASKRGLVLDGNHNSNFY...T...L...E...Y...S...D...V...L...G...C...R...A...S...K...I...V...N...E...P...K...K...H...E...I...P...I...V...S...V...N...A...S...K
44604412 Environmental sequence PLYLYTSRRGLTIDGHNFSNY...Y...T...L...E...Y...S...D...V...L...G...C...R...A...S...K...I...V...N...E...P...K...K...H...E...I...P...I...V...S...V...N...A...S...K
21232053 X. campestris DKAH...P...D...V...T...T...D...T...F...E...T...E...V...L...K...S...L...T...P...V...I...V...D...F...W...A...D...W...G...F...C...R...Q...E...L...F...V...L...E...K...A...D...Y...N...G...-...A...F...E...A...K...I...D...V
23060100 P. fluorescens SSDL...K...H...V...S...D...A...S...F...E...A...D...V...L...K...A...E...G...-...-...A...V...I...V...D...F...W...A...D...W...G...F...C...R...Q...E...L...F...V...L...E...K...A...D...Y...N...G...-...K...L...T...A...K...I...N...I
28810438 S. pyogenes WRKK...A...L...E...V...T...D...A...T...F...V...E...E...T...K...E...G...-...-...L...V...I...D...F...W...A...D...W...G...F...C...R...Q...E...L...F...V...L...E...Q...S...Q...E...I...D...E...L...K...L...K...I...D...V
28900827 V. parahaemolyticus SPLLDGVP...I...E...G...T...L...D...N...F...S...A...L...L...E...S...T...P...V...V...D...F...W...A...D...W...G...F...C...R...Q...E...L...F...V...L...E...S...D...A...Q...E...Q...A...G...-...T...A...R...F...V...K...I...D...T
29345628 B. thetaiotaomicron DKEMFLK...D...V...F...D...Y...E...K...S...K...E...W...K...Y...G...D...P...A...L...I...D...F...W...A...D...W...G...F...C...R...Q...E...L...F...V...L...E...K...A...E...Y...A...G...-...K...I...T...Y...K...I...N...V
34897150 O. sativa RVVA...H...S...T...A...T...W...D...E...Q...W...G...A...H...K...S...N...P...N...L...I...V...I...D...F...W...A...D...W...G...F...C...R...P...I...B...A...F...K...D...A...G...R...F...A...V...F...F...K...I...D...D...E
27763683 C. reinhardtii DRVVEVT...S...D...Q...D...F...S...A...K...L...A...D...V...A...G...S...G...S...L...M...I...C...P...T...A...W...G...F...C...R...M...I...A...D...P...S...S...I...S...N...K...Y...T...V...T...F...F...K...I...D...I...D...N
41053764 D. rerio MVG...R...V...I...G...N...D...S...D...P...A...E...L...S...G...A...G...S...L...T...V...W...K...P...T...M...S...C...R...P...C...V...R...I...A...D...A...P...N...M...I...S...N...K...Y...Q...V...V...F...L...E...V...D...I...H...V
21554313 A. thaliana NAPN...V...D...I...H...S...T...E...E...P...L...S...A...L...S...G...A...G...E...R...L...V...I...V...E...F...M...C...I...W...C...S...R...A...L...F...E...K...I...C...K...T...A...V...E...H...P...I...V...F...L...K...V...N...F...D...E
28868894 P. syringae SGCG...D...L...G...T...D...Q...N...G...K...V...A...S...E...R...I...K...G...H...L...W...V...N...W...A...D...W...G...F...C...R...T...E...V...E...F...N...A...I...S...E...Q...L...K...I...K...V...I...L...G...I...N...F

```


Fig. S5

```

43561023 Environmental sequence --VKKDDTGRVYVLLWVYFKADTPGUTHEGNGFRRIQIFEDRNASIVGLSYDSPAENGARDK
43541593 Environmental sequence IQRNFKDTSKNNLILLFYPKDDTPGUTHEARFPGSLNDFEQAGVWVLGSDNVAASHKGRDK
44311362 Environmental sequence NLHQHDDYIGKNNVLYFFPKADTPGUIKQACFRPEYKNEKYNSILIGSYDQESALRSRKK
43508495 Environmental sequence NLHQHDDYIGKNNVLYFFPKADTPGUIKQACFRPEYKNEKYNSILIGSYDQESILRLRKK
44169693 Environmental sequence ITHSLSAMKGRVLLAFYSDYADTPGUIKQACFRPNLYQEFIKNNIVMGSYVKAQDLRGRDK
43105956 Environmental sequence ITHSLSAMKGRVLLAFYSDYADTPGUIKQACFRPNLYQEFIKNNIVMGSYVKAQDLRGRDK
43475719 Environmental sequence QLHRSDYRGGTVLAAHFKAPFGUFAKSLRISGKLRAPDQSYFMASTQKKNTAIAEK
6435547 R. norvegicus IDISLSDYRGGTVLAAHFKAPFGUFAKSLRISGKLRAPDQSYFMASTQKKNTAIAEK
13786925 C. fasciculata KVSLSGKGRVLLAFYSDYADTPGUIKQACFRPEYKNEKYNSILIGSYDQESALRSRKK
22267474 M. musculus IDPLG---DSGELLSHDDEPTVCTDELGRAAKLAFPAKRNKILALSYDQESHIAEK
23471338 P. syringae IDPLG---DSGELLSHDDEPTVCTDELGRAAKLAFPAKRNKILALSYDQESHIAEK
11139253 A. capsulatus IDPLG---DSGELLSHDDEPTVCTDELGRAAKLAFPAKRNKILALSYDQESHIAEK
23028046 M. degradans IVRLSSPAGERNVWVYFKAMTPCIVQACLRISKKLEEDVDVWVFGSPDDVSRLEITEK
23013762 M. magnetotacticum GKAAALADYKGRVLLAFYSDYADTPGUIKQACFRPEYKNEKYNSILIGSYDQESALRSRKK
16082618 T. acidophilum IMRKLSDYKGRVLLAFYSDYADTPGUIKQACFRPEYKNEKYNSILIGSYDQESALRSRKK
16765025 S. typhimurium SDVSLSDYKGRVLLAFYSDYADTPGUIKQACFRPEYKNEKYNSILIGSYDQESALRSRKK

```

Fig. S6

```

6680075 M. musculus      LSSLGKGVLLIENVASLIGTHIRVDYHEINDLQKRLGPRGLVVLGFPFCNPFQHOENKNEEINLSLKY
41406084 H. sapiens           LSSLGKGVLLIENVASLIGTHIRVDYQINELORRLGPRGLVVLGFPFCNPFQHOENKNEEINLSLKY
14717812 R. norvegicus      LDRKSGVCLIIENVASOUGKIDVNYIQLVLDLHARLAECCGLRLLAFPCNPFQHOEPGSAEIKGEAAG
42916350 Environmental sequence -----LVNVASLUGKISQWYKELVALHKLGHRLGELCLAFPCNPFQHOEPGSAEIKGEAAG
44608250 Environmental sequence LSTHSQPCLLIENVASOUGLTPQVAGLRTLHNETDDLNLVLFPPCNQFGAQEPGSDDELLDEVTN
44102119 Environmental sequence FSBMRBQALLIENVASOUGLTPQVYGLCALERQRDDLNLVLFPPCNQFGAQEPGSDDELLDEVTN
43527571 Environmental sequence -----LVNVASOUGLTPQVYKELVQLDNKYENLNLVAFPPCNQFGAQEPGSAEIKGEAAGK
21913146 C. reinhardtii      FKSLLNNVILLIENVASOUGLTPAAYKEFATLLGKYPATDFTLVAFPCNPFQHOEPGSAEIKGEAAGK
18026892 H. brasiliensis     LSTYKGRVILLIENVASOUGLTPNSVYBETQLYQKQKQDGLLELAFPCNPFQHOEPGSAEIKGEAAGK
20147455 B. napus             LDRYKGRVILLIENVASOUGLTPSSVYBETQLYQKQKQDGLLELAFPCNPFQHOEPGSAEIKGEAAGK
18028086 R. sativus          LSRVYKGRVILLIENVASOUGLTPGKRPENLYAKKTKKGLLELAFPCNPFQHOEPGSAEIKGEAAGK
19745714 S. pyogenes         LSRVYKGRVILLIENVASOUGLTPQVAGLQALVDYTLHDKGFELELAFPCNPFQHOEPGSAEIKGEAAGK
16125974 C. crescentus       LADYKGRVILLIENVASOUGLTPQVAGLEALYKAKHDKGFELELAFPCNPFQHOEPGSAEIKGEAAGK
15838488 X. fastidiosa       LADYKGRVILLIENVASOUGLTPQVAGLEALYKAKHDKGFELELAFPCNPFQHOEPGSAEIKGEAAGK
7433111 C. reinhardtii      FKDLKGRVILLIENVASOUGLTPQVAGLEALYKAKHDKGFELELAFPCNPFQHOEPGSAEIKGEAAGK
15596484 P. aeruginosa       ---LAKGRVILLIENVASOUGLTPQVAGLEALYKAKHDKGFELELAFPCNPFQHOEPGSAEIKGEAAGK
23105542 A. vinelandii      LADYKGRVILLIENVASOUGLTPQVAGLEALYKAKHDKGFELELAFPCNPFQHOEPGSAEIKGEAAGK
19553787 C. glutamicum       LADYKGRVILLIENVASOUGLTPQVAGLEALYKAKHDKGFELELAFPCNPFQHOEPGSAEIKGEAAGK
585223 B. taurus           FKQYKGRVILLIENVASOUGLTPQVAGLEALYKAKHDKGFELELAFPCNPFQHOEPGSAEIKGEAAGK
1708061 D. immitis          LADYKGRVILLIENVASOUGLTPQVAGLEALYKAKHDKGFELELAFPCNPFQHOEPGSAEIKGEAAGK
544436 W. bancrofti       LADYKGRVILLIENVASOUGLTPQVAGLEALYKAKHDKGFELELAFPCNPFQHOEPGSAEIKGEAAGK
17550320 C. elegans          LSRKGRVILLIENVASOUGLTPQVYDFNPLEKYLQAQGLLVAFPCNPFQHOEPGSAEIKGEAAGK

```

Fig. S7

```
44513405 Environmental sequence KYIVYHNPRUGKSRGVALLLNEYNITFDVIEYLNKPLRBEVLIAPKLGIA-PGEFVRNKEKRENDID
43340045 Environmental sequence KYIVYHNPRUGKSRGVALLLNEYNITFDVIEYLNKPLRBEVLIAPKLGIA-PGEFVRNKEKVENRLEY
43472556 Environmental sequence DWVYHNPRUGKSRGVALLEKDLKPSVIEYLNKPLRBEVLIAPKLGIA-PADFPASADKFRANQIK
44484681 Environmental sequence DWVYHNPRUGKSRGVALLEKDLKPSVIEYLNKPLRBEVLIAPKLGIA-PADFPASADKFRANQIK
44278052 Environmental sequence -MYVYHNPRUGKSRGVALLEKKNIDPILIEYLNKPLRBEVLIAPKLGIA-PADFPASADKFRANQIK
43239390 Environmental sequence DKLYVYHNPRUGKSRGVALLEKQNNIDPILIEYLNKPLRBEVLIAPKLGIA-PADFPASADKFRANQIK
44586697 Environmental sequence KYIVYHNPRUGKSRGVALLEKDNIDPILIEYLNKPLRBEVLIAPKLGIA-PADFPASADKFRANQIK
44565900 Environmental sequence -MYVYHNPRUGKSRGVALLEKDNIDPILIEYLNKPLRBEVLIAPKLGIA-PADFPASADKFRANQIK
44565893 Environmental sequence -MYVYHNPRUGKSRGVALLEKDNIDPILIEYLNKPLRBEVLIAPKLGIA-PADFPASADKFRANQIK
43995796 Environmental sequence NVTIYHNPRUGKSRGVALLEKEMVIEYLNKPLRBEVLIAPKLGIA-PADFPASADKFRANQIK
16123247 Y. pestis DWIYVYHNPRUGKSRGVALLEKQVIEYLNKPLRBEVLIAPKLGIA-PADFPASADKFRANQIK
1073863 H. influenzae SMTIYVYHNPRUGKSRGVALLEKQVIEYLNKPLRBEVLIAPKLGIA-PADFPASADKFRANQIK
28868893 P. syringae DWIYVYHNPRUGKSRGVALLEKQVIEYLNKPLRBEVLIAPKLGIA-PADFPASADKFRANQIK
15675953 N. meningitidis EIKIYVYHNPRUGKSRGVALLEKQVIEYLNKPLRBEVLIAPKLGIA-PADFPASADKFRANQIK
23014994 M. magnetotacticum TMTIYVYHNPRUGKSRGVALLEKQVIEYLNKPLRBEVLIAPKLGIA-PADFPASADKFRANQIK
5915690 A. multivorum NVTIYVYHNPRUGKSRGVALLEKQVIEYLNKPLRBEVLIAPKLGIA-PADFPASADKFRANQIK
26250123 E. coli NVTIYVYHNPRUGKSRGVALLEKQVIEYLNKPLRBEVLIAPKLGIA-PADFPASADKFRANQIK
27378195 B. japonicum SMTIYVYHNPRUGKSRGVALLEKQVIEYLNKPLRBEVLIAPKLGIA-PADFPASADKFRANQIK
15964828 S. meliloti TMTIYVYHNPRUGKSRGVALLEKQVIEYLNKPLRBEVLIAPKLGIA-PADFPASADKFRANQIK
16125750 C. crescentus PHTIYVYHNPRUGKSRGVALLEKQVIEYLNKPLRBEVLIAPKLGIA-PADFPASADKFRANQIK
38505862 Synechocystis sp. MMTIYVYHNPRUGKSRGVALLEKQVIEYLNKPLRBEVLIAPKLGIA-PADFPASADKFRANQIK
```

Fig. S8

```

1591288 M. jannaschii DEAKKPIIDKLLKANQKIVVIFPEPFAUGSPKSGIAIAHPN-ENRKLIVDIEFKYVITPDEQ
45047123 M. maripaludis EEAMGPIINEKISDTGSKDILVDFPEPQUGSPKSIDTASKEIDTDEKIVDEDFRFDKELRQV
47118322 C. perfringens EITEKSPFVAVGPEKFDKRNIAAGVGGSPVAVIVLQAS-SRNDGAKIEDITFFDKELVKD
39996309 G. sulfurreducens EAVLAPIVGHEHAKILRWFEPGCHGSPRLGLVLDSP-ADNDARVLAGPFAITSNFRSL
51854827 S. thermophilum EAAEIAARRLEAKPEKGGHREIVGKISGSPSLGLALDEP-REEDPTIVEGDARVYIQQVAK
46579793 D. vulgaris EKEEAEYFAI--KQKTPRIVLAPGCHGSPRLGLALDEP-NESNDFKEGDPPTQVNSDLLSQ
50874889 D. psychrophila AIDKLRVYMEON-KIISALRVAIMQKCIKSPSLGLALDEP-KDNDKSFDFDSITPTIESELLIT
48785105 B. fungorum AADKVKQLIIEEGNADLKLRFVIVGGCCSGFOYGFDFDEANEDDTIVMNSVQLLIDMSIQY
44554993 Environmental sequence AADKVKQLIIEEGNADLKLRFVIVGGCCSGFOYGFDFDEANEDDTIVMNSVQLLIDMSIQY
45520124 R. eutropha AADKVKQLIIEEGNADLKLRFVIVGGCCSGFOYGFDFDEANEDDTIVMNSVQLLIDMSIQY
54031087 Polaromonas sp. AADKVKQLIIEEGNADLKLRFVIVGGCCSGFOYGFDFDEANEDDTIVMNSVQLLIDMSIQY
47572999 R. gelatinosus AADKVKQLIIEEGNADLKLRFVIVGGCCSGFOYGFDFDEANEDDTIVMNSVQLLIDMSIQY
52005978 T. denitrificans AADKVKQLIIEEGNADLKLRFVIVGGCCSGFOYGFDFDEANEDDTIVMNSVQLLIDMSIQY
33598389 B. parapertussis AADKVKQLIIEEGNADLKLRFVIVGGCCSGFOYGFDFDEANEDDTIVMNSVQLLIDMSIQY
53759505 M. flagellatus AADKVKQLIIEEGNADLKLRFVIVGGCCSGFOYGFDFDEANEDDTIVMNSVQLLIDMSIQY
56475933 Azoarcus sp. AADKVKQLIIEEGNADLKLRFVIVGGCCSGFOYGFDFDEANEDDTIVMNSVQLLIDMSIQY
34105000 C. violaceum AADKVKQLIIEEGNADLKLRFVIVGGCCSGFOYGFDFDEANEDDTIVMNSVQLLIDMSIQY
43148776 Environmental sequence AADKVKQLIIEEGNADLKLRFVIVGGCCSGFOYGFDFDEANEDDTIVMNSVQLLIDMSIQY
53758707 M. capsulatus AADKVKQLIIEEGNADLKLRFVIVGGCCSGFOYGFDFDEANEDDTIVMNSVQLLIDMSIQY
32029699 H. somnus AADKVKQLIIEEGNADLKLRFVIVGGCCSGFOYGFDFDEANEDDTIVMNSVQLLIDMSIQY
15602323 F. multocida AADKVKQLIIEEGNADLKLRFVIVGGCCSGFOYGFDFDEANEDDTIVMNSVQLLIDMSIQY

```


Fig. S9

```

44260962 Environmental sequence EKIESIKGVGMPSELMELENILNEKVPYIUGCPEARGSSS-----
43490440 Environmental sequence DKIESIKGVGMPSELMELENILNEKVPYIUGCPEARGSSS-----
26249943 E. coli NQLTSPISDE-FDIIVRAMQQLAAEQAVTLN/CVAAALRRGLIDPEAKREI
24053808 s. flexneri NQLTSPISDE-FDIIVRAMQQLAAEQAVTLN/CVAAALRRGLIDPEAKREI
56415370 S. enterica NQLTSPISDE-YDIIVRAMQQLAAEQAVTLN/CVAAALRRGLIDPEAKREI
44011039 Environmental sequence NQLTSPISDE-YDIIVRAMQQLAAEQAVTLN/CVAAALRRGLIDPEAKREI
51597998 Y. pseudotuberculosis NQLTAPISDE-FDIIVRAMQQLAAEQAVTLN/CVAAALRRGLIDPEAKREI
50122961 E. carotovora NQLTSPISDE-FDIIVRAMQQLAAEQAVTLN/CVAAALRRGLIDPEAKREI
37524435 P. luminescens NQLTSPISDE-FDIIVRAMQQLAAEQAVTLN/CVAAALRRGLIDPEAKREI
37681219 V. vulnificus TELIVPISDE-FDIIVRAMQQLAAEQAVTLN/CVAAALRRGLIDPEAKREI
48868941 H. influenzae NALIVPISDE-VNIQSHMMPFSITHNVPLHI/CVAAALRRGLIDPEAKREI
15597801 P. aeruginosa SANVVSQDE-FDIIVRAMQQLAAEQAVTLN/CVAAALRRGLIDPEAKREI
53612706 A. vinelandii ADSIVTPEDE-SDLPAQRAPFVERHDAV/CVAAALRRGLIDPEAKREI
50085033 Acinetobacter sp. NNLOWVPDD-RNIPNRMSSELAKENMIDLV/CVAAALRRGLIDPEAKREI
44493175 Environmental sequence TRLAIPPDD-RHIPNRMSSELAKENMIDLV/CVAAALRRGLIDPEAKREI
43771839 Environmental sequence TRLAIPPDD-RHIPNRMSSELAKENMIDLV/CVAAALRRGLIDPEAKREI
44430522 Environmental sequence TRLAIPPDD-RHIPNRMSSELAKENMIDLV/CVAAALRRGLIDPEAKREI
14285420 A. vinosum TRLAIPPDD-RHIPNRMSSELAKENMIDLV/CVAAALRRGLIDPEAKREI
52006365 T. denitrificans TRIGEPDD-RNITTRMSKLAEEHSDLV/CVAAALRRGLIDPEAKREI

```

Fig. S10

```
44505793 Environmental sequence MTSKYLLSPVTIDVORANIVLQAKMVEEVYTHLAD-NKPDWFEVSPHRSKVPFLADQEV
42992680 Environmental sequence MAKNIHLLSSTVTDVORANIVLQAKMVEEVYTHLAD-NKPDWFEKISPHRSKVPFLADDEI
43211849 Environmental sequence YNKYPIILDFWSPVRAKIALALQKGVSEYREBNLE-NKSPDLLQKNSPHRSKVPFLADKIDQ
46324676 B. cepacia STLQVLLSHPLCDFVORANIVLQAKMVEEVYTHLAD-NKPDWFLKISPHRSKVPFLADGEP
54031150 Polaromonas sp. MASQVLLSHPLCDFVORANIVLQAKMVEEVYTHLAD-NKPDWFLKISPHRSKVPFLADGEA
17547313 R. solanacearum PDSTVLLSHPLCDFVORANIVLQAKMVEEVYTHLAD-NKPDWFLKISPHRSKVPFLADWRD
50120521 E. carotovora LNAQVLLSHPLCDFVORANIVLQAKMVEEVYTHLAD-NKPDWFLKISPHRSKVPFLADKELD
22038178 A. tauschii GGDDLLLGAWSPDFVTRVGLALAKGSLSDVEEFLY-NKSEDLKSNPHRSKVPFLAHNFA
20143562 O. sativa GRDELLLGAWSPDFVTRVGLALAKGSLSDVEEFLY-NKSEDLKSNPHRSKVPFLAHNFA
8052535 A. thaliana MADEVLLDFWSPDFVRAKIALALQKGVSEYREBNLE-NKSPDLLQKNSPHRSKVPFLAHNFK
1737447 E. globulus MAEEVLLDFWSPDFVRAKIALALQKGVSEYREBNLE-NKSPDLLQKNSPHRSKVPFLAHNFK
11385507 Z. mays ---AVRLLGSPASDFVRAKIALALQKGVSEYREBNLE-NKSEDLKSNPHRSKVPFLAHGDR
29290335 P. acutifolius SQEEVLLGATGSPDFVRAKIALALQKGVSEYREBNLE-NKSEDLKSNPHRSKVPFLAHGDK
47222286 T. nigroviridis AKDHVLLGATGSPDFVRAKIALALQKGVSEYREBNLE-NKSPDLLQKNSPHRSKVPFLAHGDK
15808378 T. rubripes PEGHIVLLGATGSPDFVRAKIALALQKGVSEYREBNLE-NKSPDLLQKNSPHRSKVPFLAHGDK
55250043 D. rerio PNGQVLLGATGSPDFVRAKIALALQKGVSEYREBNLE-NKSPDLLQKNSPHRSKVPFLAHGDK
49900006 X. tropicalis SEETVLLGATGSPDFVRAKIALALQKGVSEYREBNLE-NKSPDLLQKNSPHRSKVPFLAHGDK
50927069 R. norvegicus PEGVIVLLGATGSPDFVRAKIALALQKGVSEYREBNLE-NKSPDLLQKNSPHRSKVPFLAHGDK
21311857 M. musculus PEGVIVLLGATGSPDFVRAKIALALQKGVSEYREBNLE-NKSPDLLQKNSPHRSKVPFLAHGDK
31873364 H. sapiens PEGVIVLLGATGSPDFVRAKIALALQKGVSEYREBNLE-NKSPDLLQKNSPHRSKVPFLAHGDK
46518247 C. gigas EACTVLLGATGSPDFVRAKIALALQKGVSEYREBNLE-NKSPDLLQKNSPHRSKVPFLAHGDK
21355779 D. melanogaster DDGVVLLGATGSPDFVRAKIALALQKGVSEYREBNLE-NKSPDLLQKNSPHRSKVPFLAHGDK
```

Fig. S11

44275476 Environmental sequence ALNFINIGKDKTKFTPEQFEIENNAGVDPKQIVTYQ--GGRAAHVMFVLALVSTFSFPNINYDRVKVYDSSSGEWA
43834666 Environmental sequence ALFIDENNNKFKSQNDESIENKQNIYTKQIATYQ--GGRAAHVFFVLLKLS-----YKMKVYDSSSGEWA
44586938 Environmental sequence WFNLMDE-QTHFRSEEDKAIADNGIALKAVIYQ--AGVRAAHVNFVLLQIT-----QSEARVYDSSSGEWA
49176232 E. coli WTELVRE-GELKT--ETDIDAIFFGRGVSYKPIIVS--SGVRAAVVLLALATLD-----VPMVKLYDGSWSEWG
37525471 P. luminescens WTMLVEN-GHFKS--ETDIDAIFFKQVLDLNPKITSG--SGVRAAVVLLGDIIT-----KKDYYLYDGSWAEWG
9658033 V. cholerae FAELITG-HKLKE--QALDRPLTHMLPETAQEYLFSG--SGVRAACIVLLAAVVG-----YKMSVYDGSWTEWG
24372842 S. oneidensis FGEVLNG-YMKKS--TTLQAIFAQVGNKALR-IFSG--SGVRAACIVLLASVVG-----HKSAVLYDGSWADWG
54302561 P. profundum FSQLIKD-GFFID--KELVNRFNALS-DIQRIIFS--SGVRAACIVLLGAEELG-----RKMIVYDGSWTEWG
50905511 O. sativa FLEMFDADMLLP--ADLIRKRFQAGISLRPIVVTG--SGVRAACIVLLGAYRIG-----KQDIPVYDGSWTEWG
4406372 D. glomerata FPQILDASQLLP--ADLIRKRFQAGISLRPIVVTG--SGVRAACIVLLGAYRIG-----KSDVAVYDGSWTEWG
39996033 G. sulfurreducens WNEA--HIPGANSPPFALEKKNPALTASKRPIVVFYCGVTVLSPKSAAGLAKKSGE-----KVRVYLDSEEPKWK
45360053 R. xylanophilus WYRQ--HIPGARYESAVQVRKAPQKDAF-IVAV--CSNFNHSTRVARELAAMGE-----NWYDYEGCKQDMW
34557042 W. succinogenes WRRES--HIPGSHGSDGKFKELWGRIPMDPNTKIVVFCGYECELSSHVSHVAMGK-----NFMITYSGGTEPKK
48834389 Magnetococcus sp. WYAG--HIPGANIPPKLEANALPAGQVAVAV--CRGFWCVLAFDVARLRRAREI-----KARRVLDSEEPWR
46106717 R. xylanophilus WYRAG--HIPGANSPIPEREAYAEIPKQDIIVAY--CRGFWCVLAFDVARLRRAREI-----KARRVLDSEEPWR
52007258 T. denitrificans WYQAG--HIPGANIPDLDLPHHEEAPQGGIIVAY--CRGFWCVLAFDVARLRRAREI-----KARRVLDSEEPWR
46317410 B. cepacia WTEG--HIPGANIPSHLDARSEIPAGTIIVAY--CRGFWCVLAFDVARLRRAREI-----KARRVLDSEEPWR
54029015 Polaromonas sp. WYTSA--HIPGARSPIVDELKRRNEIPKDVPIVAY--CRGFWCVLAFDVARLRRAREI-----KARRVLDSEEPWR
48838221 M. barkeri WYEMM--HIPGANSPIPEDEKHTATPINQIIVAY--CRGFWCVLAFDVARLRRAREI-----KARRVLDSEEPWR
54025306 N. farcinica WYASG--HIPGANIPDQSDRPAEIPADTEIVAY--CRGFWCVLAFDVARLRRAREI-----KARRVLDSEEPWR
3955039 S. peucetius WYLAG--HIPGANIPVAVLIDRGEAKDTIVVAY--CRGFWCVLAFDVARLRRAREI-----KARRVLDSEEPWR
15607465 M. tuberculosis WYQAG--HIPGANIPFARLADRPAEIPDREIVAY--CRGFWCVLAFDVARLRRAREI-----KARRVLDSEEPWR
20807605 T. tuberculosus WYEQA--HIPGANSPIPELPHNCLSKDKLIVAY--GASYECSIEEAEIIVANYG-----NWKVYRSGGTEPKK
46581156 D. vulgaris WPIPDMDNMAETGDKSQDFEALGPDKNRPIVVFYCFVKTIRSHNGAVWAQKLSIT-----NMYRPPGGIVAMK
50874889 D. psychrophila WTSSEFTRSHRANPKRDTWKSXFAKDKKIVLYCAQ--NESTASLARNDTADSG-----SVHAFKGGWRMS
48847313 G. metallireducens AVGAINLPNDGPDADIERIKQMEPFTKKDEIIVLYCA--GEQASARVALVLERGET-----KTYVVRGGQAVF
53760573 R. eutropha WPIE--HIPGANIPDHPGLDAGGLESLRDIIVLYCAQ--NESTASLARNDTADSG-----KRWALDGGFDEWK
48768248 R. metallidurans WAPIE--HIPGANIPVEINSLKDPGLDRDASIVLYCAQ--HEISSAVLAEERLTAGIP-----NTWALDGGFDEWK
15596406 P. aeruginosa WDEPS--HIPGANIPVEINSLKDPGLDRDASIVLYCAQ--HEISSAVLAEERLTAGIP-----NTWALDGGFDEWK
33603950 B. bronchiseptica WRDEQ--HIPGANIPDHRAPLQDQFDPEAGIIVLYCAQ--NEISSAVLAEERLTAGIP-----NTWALDGGFDEWK
54029683 Polaromonas sp. WRAGG--HIPGANIPVWSDLRKMASLDPHDAHVVLYCAQ--NDASDAQVRRKRAAGS-----NVRPFLGGIDAWR
48782625 B. fungorum WRKLDPFVIPGTFADERQDEIIVATYPRDQKLVLYCAQ--NEISSAVLAEERLTAGIP-----NVRPFLGGIDAWR
44357259 Environmental sequence RALDPFVIPGTFADERQDEIIVATYPRDQKLVLYCAQ--NEISSAVLAEERLTAGIP-----NVRPFLGGIDAWR
44624393 Environmental sequence RKLDPFVIPGTFADERQDEIIVSRYPFSQKVVLYCAQ--NEFTDALMARRLIDAGT-----DALALRGGIDAWR
21242497 X. axonopodis RQLQPYTIPGAFADERQDAIIVASIPDRSRVVLYCAQ--DEISSAVLAEERLTAGIP-----NVRPFLGGIDAWR
17548617 R. solanacearum RMSQPHRIPGANLYDSAKDGPVIEGPDRIIVLYCAQ--NEASSAVLAEERLTAGIP-----NVRPFLGGIDAWR
47573188 R. gelatinosus WAGLDLRHIPGANRVELSEVATHASQPRDRIVLYCAQ--NEASSAVLAEERLTAGIP-----NVRPFLGGIDAWR
53730919 D. aromatica WVAETG--HIPGANIPVAVLIDRGEAKDTIVVAY--CRGFWCVLAFDVARLRRAREI-----KARRVLDSEEPWR
50874889 D. psychrophila WYENG--HIPGAKLIPVGOESRDELP--KIKPVIIVYCA--IGGRSRVAVQLLAGKES-----KHYNLGGIDAWR
46112894 Exiguobacterium sp. WYKGN--HIPGAKLIPVGOESRDELP--KIKPVIIVYCA--IGGRSRVAVQLLAGKES-----KHYNLGGIDAWR
46142555 M. burtonii WYNSG--HIPGANNEVSLGTRNEAP--AKKVLVYCA--IGGRSRVAVQLLAGKES-----KHYNLGGIDAWR
46198460 T. thermophilus WYAGE--HIPGANIPVAVLIDRGEAKDTIVVAY--CRGFWCVLAFDVARLRRAREI-----KARRVLDSEEPWR
48847131 G. metallireducens WYGGG--HIPGANIPVAVLIDRGEAKDTIVVAY--CRGFWCVLAFDVARLRRAREI-----KARRVLDSEEPWR
46580382 D. vulgaris WYAEQ--HIPGANIPVAVLIDRGEAKDTIVVAY--CRGFWCVLAFDVARLRRAREI-----KARRVLDSEEPWR
53691784 D. desulfuricans WYRQG--HIPGARLPVAVLIDRGEAKDTIVVAY--CRGFWCVLAFDVARLRRAREI-----KARRVLDSEEPWR
43887282 Environmental sequence WYEIC--HIPGARLPVAVLIDRGEAKDTIVVAY--CRGFWCVLAFDVARLRRAREI-----KARRVLDSEEPWR
52006282 T. denitrificans WYRQG--HIPGARLPVAVLIDRGEAKDTIVVAY--CRGFWCVLAFDVARLRRAREI-----KARRVLDSEEPWR
53757119 M. capsulatus WYAEQ--HIPGANIPVAVLIDRGEAKDTIVVAY--CRGFWCVLAFDVARLRRAREI-----KARRVLDSEEPWR
53729577 D. aromatica WYASG--HIPGANIPVAVLIDRGEAKDTIVVAY--CRGFWCVLAFDVARLRRAREI-----KARRVLDSEEPWR
56420601 G. kaustophilus WYAPG--HIPGANSPIPELENRALP--KIKPVIIVYCA--IGGRSRVAVQLLAGKES-----KHYNLGGIDAWR
23099356 O. iheyensis WYDKG--HIPGANIPVAVLIDRGEAKDTIVVAY--CRGFWCVLAFDVARLRRAREI-----KARRVLDSEEPWR

Fig. S12

```

44612327 Environmental sequence LVSGLFRFGQGVHFDPSGSPCYRCLNSEPPPAALVPSUAVGVGLGDPGVGLQATEVTKLLLEN
43231982 Environmental sequence LVSGLFRFGQGVHFDPSGSPCYRCLNSEPPPAALVPSUAVGVGLGDPGVGLQATEVTKLLLDN
43257063 Environmental sequence LVSGLFRFGQGVHFDPSGSPCYRCLNSEPPPAALVPSUAVGVGLGDPGVGLQATEVTKLLLEN
43262145 Environmental sequence LVSGLFRFGQGVHFDPSGSPCYRCLNSEPPPAALVPSUAVGVGLGDPGVGLQATEVTKLLLEN
42946779 Environmental sequence LVSGLFRFGQGVHFDPSGSPCYRCLNSEPPPAALVPSUAVGVGLGDPGVGLQATEVTKLLLVE
2950364 Synechococcus sp. NVVGSIFRFQGVHFDPSGSPCYRCLNSEPPPAALVPSUAVGVGLGDPGVGLQATEVTKLLLSQ
22299946 T. elongatus NVVGSIFRFQGVHFDPSGSPCYRCLNSEPPPAALVPSUAVGVGLGDPGVGLQATEVTKLLLSK
37522981 G. violaceus NVVGSIFRFQGVHFDPSGSPCYRCLNSEPPPAALVPSUAVGVGLGDPGVGLQATEVTKLLLSI
45545541 R. xylanophilus NVVGSIFRFQGVHFDPSGSPCYRCLNSEPPPAALVPSUAVGVGLGDPGVGLQATEVTKLLLSI
53796678 C. aurantiacus NVVGSIFRFQGVHFDPSGSPCYRCLNSEPPPAALVPSUAVGVGLGDPGVGLQATEVTKLLLSI
53610625 A. vinelandii NVVGSIFRFQGVHFDPSGSPCYRCLNSEPPPAALVPSUAVGVGLGDPGVGLQATEVTKLLLSI
48834858 T. fusca YVVGSIFRFQGVHFDPSGSPCYRCLNSEPPPAALVPSUAVGVGLGDPGVGLQATEVTKLLLSI
54659256 C. hominis LLDSGTEGENHSRHLIPGET-SVDEKTMGLNVODTNFDCIKKEFPPTPIHCLYAFIYEDEQD--
21464561 A. thaliana MVEKSTGTEGKCHARVLLPGVT-PCDEKNIYLFPPQVKFPLCLNLETPRNAACLEYAHLIWEVYHRSK
6855414 L. major LLDSEGLTGTKCMQPAIPFVT-ESVSS--SYDPPKGLDCLNKNFPNAEHTIQWARDLPHLFLVSV
17539268 C. elegans LLDSEGLTGTKCNTQVYVYLYT-ESVSS--SYDPPKGLDCLNKNFPNAEHTIQWARDLPHLFLVSV

```

Fig. S13

12642418	<i>C. hydrogeniformans</i>	HTKRALVIGGGVAGIQAALEADKGIQVRLVEKEPILGGIMHIDKTFPPTDCSSQI ¹ STPRMAAALHP
15669884	<i>M. jannaschii</i>	VDKSLIIGGGVAGIQAALEADKGIQVRLVEKEPSIGGIMAKLAKTFPPTDCSSQI ¹ LLAPKMSLVANHP
45359260	<i>M. maripaludis</i>	VDKSLIIGGGVAGIQAALEADKGIQVRLVEKEPSIGGIMAKLAKTFPPTDCRAMV ¹ LLAPKMSLVANHP
20093689	<i>M. kandleri</i>	VENSVLIGGGVAGIQAALEADKGIQVRLVEKEPILGGIMHIDKTFPPTDCRAMV ¹ LLAPKMSLVANHP
15679380	<i>M. thermototrophicus</i>	VDDALVIGGGVAGIQAALEADKGIQVRLVEKEPSIGGIMAKLAKTFPPTDCSSQI ¹ LLAPKMSLVANHP
46580807	<i>D. vulgaris</i>	VTRBALVIGGGVAGIQAALEADKGIQVRLVEKEPSIGGIMAKLAKTFPPTDCSSQI ¹ LLAPKMSLVANHP
23475491	<i>D. desulfuricans</i>	VTRBALVIGGGVAGIQAALEADKGIQVRLVEKEPSIGGIMAKLAKTFPPTDCSSQI ¹ LLAPKMSLVANHP
48847469	<i>G. metallireducens</i>	VTKRSLVIGGGVAGIQAALEADKGIQVRLVEKEPSIGGIMAKLAKTFPPTDCSSQI ¹ LLAPKMSLVANHP
39995201	<i>G. sulfurreducens</i>	VTKRSLVIGGGVAGIQAALEADKGIQVRLVEKEPSIGGIMAKLAKTFPPTDCSSQI ¹ LLAPKMSLVANHP
21674069	<i>C. tepidum</i>	VTRBALVIGGGVAGIQAALEADKGIQVRLVEKEPSIGGIMAKLAKTFPPTDCSSQI ¹ LLAPKMSLVANHP
20091692	<i>M. acetivorans</i>	ASRNVLVIGGGVAGIEAALNLAAGHPVDMVEKPSIIGGIMAKLAKTFPPTDCSSQI ¹ LLAPKMSLVANHP
21226158	<i>M. mazei</i>	ASRNVLVIGGGVAGIEAALNLAAGHPVDMVEKPSIIGGIMAKLAKTFPPTDCSSQI ¹ LLAPKMSLVANHP
48838060	<i>M. barkeri</i>	ASRNVLVIGGGVAGIEAALNLAAGHPVDMVEKPSIIGGIMAKLAKTFPPTDCSSQI ¹ LLAPKMSLVANHP
53731435	<i>M. burtonii</i>	ANKDVLVIGGGVAGIEAALNLAAGHPVDMVEKPSIIGGIMAKLAKTFPPTDCSSQI ¹ LLAPKMSLVANHP
11498837	<i>A. fulgidus</i>	TEISVAVIGGGVAGIEAALNLAAGHPVDMVEKPSIIGGIMAKLAKTFPPTDCSSQI ¹ LLAPKMSLVANHP

Fig. S14

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

Fig. S15

```

85801586 Environmental sequence R VMTGNVHGQIDPCGK--KNPLGGLSRRLVRIQEMRI--ASDDP VLLDAGNIPFSPNIHEGNI RSE--MHEANSILKGYERIGGDAIN
42966760 Environmental sequence K VVITCSVHGQIDPCGK--KNPLGGLSRRLVYVRIQEMRI--DSDDP VLLDAGNIPFSPNINKNVLRSE--KHCEITLSEIYERIGGDAIN
44228873 Environmental sequence D IIMVCSVHGQIDPCGK--KNPLGGLSRRLVYVRIQEMRI--ESDDP VLLDAGNIPFSPNINQNMKSE--EYAGALIEIYERIGGDAIN
85772737 Environmental sequence H IIMVGNVHGQIDPCGK--KNPLGGLSRRLVYVRIQEMRI--ESDDP VLLDAGNIPFSPNITQVYEQAE--KLRANAIIKGYERIGGDAIN
50874889 D. psychrophila R IIFSSNLSGKNTGDA--KNPLGGLSRRLASLANKTADSSVLPVYEEGNLEKISSSKNTIAQK--MARKLAAAEKIGYERIGGDAIN
43122652 Environmental sequence S IISSTINWYSFVYDCCD--KNPLGGLSRRTFPLKIMMP--DSDSFLDAGNIPDSDNINPDIISINKRFLARNPITLIEIGGDAIN
44504140 Environmental sequence S IISSTINWYSFVYDCCD--KNPLGGLSRRTFPLKIMMP--DSDSFLDAGNIPDSDNINPDIISINKRFLARNPITLIEIGGDAIN
32445059 R. baltica H IIFVGNVHGVDPCCGNLEKQKGLARIMTLKQLRI--KQVMAPIADAGNIA RRYGRQAEIKPHRS-----EALRKYDYVYIG
87311211 B. marina A IIFVGRQHGVDPCCGNLEKQKGLARIMPTLQQLRI--KQVMAPIADAGNIA RRYGRQAEIKFQTT-----ANLIRKYDYVYIG

```

Fig. S16

```

85788322 Environmental sequence -----V[AERLKEGLGTYE[DAE[KAAPEDL[FOAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
85800620 Environmental sequence -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
88802823 P. irgensii -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
55378797 H. marismortui -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
76800687 N. pharaonis -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
84497447 Janibacter sp. -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
71368962 Nocardioides sp. -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
32447814 R. baltica -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
83852439 O. alexandrii -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
88812304 N. mobilis -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
17129888 Nostoc sp. -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
75702880 A. variabilis -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
23125528 N. punctiforme -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
71673789 T. erythraeum -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
35211894 G. violaceus -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
67924076 C. watsonii -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
22295701 T. elongatus -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
83815755 S. ruber -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
84788069 E. litoralis -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
16413520 L. innocua -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
71845400 D. aromatica -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
47572949 R. gelatinosus -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
68213330 M. flagellatus -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
86160446 A. dehalogenans -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
45659220 L. interrogans -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
56381325 G. kaustophilus -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
10176026 B. halodurans -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG
66797343 D. geothermalis -----S[OAL[CVVA[SSRQ[AEVSV[GA-KEPDI[SLDPL[HICQ[VED[NR[SR[AG

```


Fig. S17.

```

44183053 Environmental sequence HPAPLAVVRLHIFULLTQIKRYASMRKVITSAKHVELIYIKGSVKQGTENKVTERRSDFT
43725511 Environmental sequence FPAALTYRSLHIFULLTRIKRYASMKKISIKSAQKIELFVIFGSLVDENRSGVSEKRSFFE
43560497 Environmental sequence YFEMDHLHIFULLTQIKRYAHMLKMEIKSGKCHVEEHLHGSVFKTIVVDHOGFTHHE
26247773 E. coli GTNPEELIAAAHACCFSMALS-LMLG-DAEPTIPSIDTAAVYSDDKVDASPAHRIKIAKSEVAE
24052078 S. flexneri GTNPEELIAAAHACCFSMALS-LMLG-DAEPTIPSIDTAAVYSDDKVDASPAHRIKIAKSEVAE
50121147 E. carotovora GTNPEELIAAAHACCFSMALS-LMLG-DESHKPSIDTAAVYSDDKVDGSPAHRIKIAHSTVTE
44496725 Environmental sequence GTNPEELIAAAHACCFSMALS-LALG-DAEPTADKIDTAAVYSDEVDGSPAHRIALHETARIS
28850626 P. syringae GTNPEELIAAAHACCFSMALS-MLLG-DAEKLKADSIDTAAVYSDEVEGSPATSAVHIVLAKLE
56180660 I. loihensis GTNPEELIAAAHACCFSMALS-LMLG-DESYEDSDTAAVYSDEEDGSPSAKIHKVAASAP
44014688 Environmental sequence GTNPEELIAAAHACCFSMALS-LILG-KAEPTAQDITAAVYSDEEQGDEWPSIALHLEAAAP
43985334 Environmental sequence GTNPEELIAAAHACCFSMALS-FALG-KAEPTAQDITAAVYSDEEQGDEWPSIALHLEAAAP
42522952 B. bacteriovorus GTNPEELIAAAHACCFSMALS-GALG-KAEPTAQDITAAVYSDEEQGDEWPSIALHLEAAAP
48855516 C. hutchinsonii GTNPEELIAAAHACCFSMALS-FQLG-GANPTPKAAEASITVQVENGSRKFKSIHLEHATVP
21243642 X. axonopodis GTNPEELIAAAHACCFSMALS-AQLT-DAEPPASIDTAAVYSDEEQGDEWPSIALHLEAAAP
46199563 T. thermophilus GTNPEELIAAAHACCFSMALS-ASLG-RECFPPKRSIDTAAVYSDEVDGKPTTRIRIDLTBAEAP
15806548 D. radiodurans GTNPEELIAAAHACCFSMALS-ALLG-DESHKPSIDTAAVYSDEVDGKPTTRIRIDLTBAEAP
17549328 R. solanacearum GTNPEELIAAAGYSACPLGAMK-FVATRDKLRLPADTSVQDSVGGAIPINSFGLE---VDAISAP
50086039 Acinetobacter sp. GTNPEELIAAAGYSACPLGAMK-FVATRDKLRLPADTSVQDSVGGAIPINSFGLE---VDAISAP
56178201 I. loihensis GTNPEELIAAAGYSACPLGALK-FAAQKVKLEADTAAVYSDEEQGDEWPSIALHLEAAAP
21220863 S. coelicolor GTNPEELIAAAGYSACPLGALG-VVARQKQDIDSG-STVTAAVYSDEEQGDEWPSIALHLEAAAP
16263744 S. meliloti GTNPEELIAAAGYSACPLGALG-LAASKHLTLPAETAVDAVYDAKSDGSRPQ---ARAVSAP

```

Fig. S18

```

23452038 C. reinhardtii  [AEI]ATFALGFWHPPEASFL--N[PGV]WRTVGVYGG--SRPNPTYES[CAG-DE]TEARRW[EDPA]L[ST]ED[LRQ]E[RE]HDP-----TQKSKCQ[MSA]W[YS]AHGT
34792792 S. purpuratus      DLE[AT]TFALGFWHPPEACFL--CAPGV[RT]VGVYGG--TKK[PT]YH----S[CG]HTEV[Q]E[ETKT]S[ER]K[ER]K[PA]NH[SS]-----TTCHKRQ[MSA]H[ED]K[Q]I
89467751 F. lividus            SLE[AT]TFALGFWHPPEACFL--CAPGV[RT]VGVYGG--TKK[PT]YH----S[CG]HTEV[Q]E[ETKT]S[ER]K[ER]K[PA]NH[SS]-----TTCHKRQ[MSA]H[ED]K[Q]I
66511346 B. floridae          --E[AT]TFALGFWHPPEACFL--CAPGV[RT]VGVYGG--TKK[PT]YH----A[CG]HTEV[Q]E[ETKT]S[ER]K[ER]K[PA]NH[SS]-----SACVSRQ[MSA]H[ED]K[Q]I
89138292 A. millepora        ETK[AT]TFALGFWHPPEACFL--CADGV[RT]VGVYGG--SKK[PT]YH----S[CG]HTEV[Q]E[ETKT]S[ER]K[ER]K[PA]NH[SS]-----TANHKRQ[MSA]H[ED]K[Q]I
51995432 H. magnipapillata  QLR[AT]TFALGFWHPPEACFL--CAPGV[RT]VGVYGG--KHP[PT]YH----N[CG]HTEV[Q]E[ETKT]S[ER]K[ER]K[PA]NH[SS]-----TANVTKRQ[MSA]H[ED]K[Q]I
49561997 B. microplius       PVK[AT]TFALGFWHPPEACFL--CAPGV[RT]VGVYGG--TKK[PT]YH----N[CG]HTEV[Q]E[ETKT]S[ER]K[ER]K[PA]NH[SS]-----TACHKRO[MSA]H[ED]K[Q]I
21640047 A. variegatum       NVK[AT]TFALGFWHPPEACFL--CAPGV[RT]VGVYGG--TSK[PT]YH----C[CG]HTEV[Q]E[ETKT]S[ER]K[ER]K[PA]NH[SS]-----TACHKRO[MSA]H[ED]K[Q]I
63523529 I. scapularis      FVK[AT]TFALGFWHPPEACFL--SAPGV[RT]VGVYGG--TKK[PT]YH----N[CG]HTEV[Q]E[ETKT]S[ER]K[ER]K[PA]NH[SS]-----TACHKRO[MSA]H[ED]K[Q]I
71547479 S. fumaroxidans    FVE[AT]TFALGFWHPPEACFL--C[LD]GV[RT]VGVYGG--RKK[PT]YH----D[CG]HTEV[Q]E[ETKT]S[ER]K[ER]K[PA]NH[SS]-----VPPKSKRQ[MSA]H[ED]K[Q]I
66861845 B. natans          PANT[AT]TFALGFWHPPEACFL--PKNV[GV]L[ST]VGVYGG--KKK[PT]YH----S[CG]HTEV[Q]E[ETKT]S[ER]K[ER]K[PA]NH[SS]-----SKPYCRQ[MSA]Y[Q]NE[Q]E
55234261 A. gambiae        PFER[AT]TFALGFWHPPEACFL--ATKGV[RT]VGVYGG--STES[PT]YH----K[CG]HTEV[Q]E[ETKT]S[ER]K[ER]K[PA]NH[SS]-----TTRMKRQ[MSA]H[ED]K[Q]I
33089114 A. mellifera       QAKR[AT]TFALGFWHPPEACFL--V[PG]V[RT]VGVYGG--QKES[PT]YH----N[CG]HTEV[Q]E[ETKT]S[ER]K[ER]K[PA]NH[SS]-----TTKIKRQ[MSA]H[ED]K[Q]I
76258265 C. aurantiacus    PLE[AT]TFALGFWHPPEACFL--Q[SGV]KDVV[SY]YGG--YVP[N]PT[YSR]V[CDG]T[GH]A[AV]R[IT]D[PAQ]S[ER]D[LM]I[F]P[ATH]D[PT]LNRQ[GD]V[GT]Q[YSR]A[H]Y[TP]T[PS]Q[RA]
66799168 D. geothermalis  MNE[AT]TFALGFWHPPEACFL--S[IS]GV[RT]VGVYGG--TLP[N]PT[YSR]V[CDG]T[GH]A[AV]R[IT]D[PAQ]S[ER]D[LM]I[F]P[ATH]D[PT]LNRQ[GD]V[GT]Q[YSR]A[H]Y[TP]T[PS]Q[RA]
88933851 Dehalococcoides sp. TTQ[AT]TFALGFWHPPEACFL--K[IK]GV[RT]VGVYGG--A[IE]N[PT]Y[Q]Q[CSG]KT[GH]A[AV]R[IT]D[PAQ]S[ER]D[LM]I[F]P[ATH]D[PT]LNRQ[GD]V[GT]Q[YSR]A[H]Y[TP]T[PS]Q[RA]
78171127 C. chlorochromatii MHQR[AT]TFALGFWHPPEACFL--KRP[GV]L[ST]VGVYGG--DIP[AT]Y[NHGT]----FAG[GE]H[V]D[PT]V[TS]Y[RH]I[EP]P[Q]I[HD]PTLNRQ[GD]V[GT]Q[YSR]A[H]Y[TP]T[PS]Q[RA]
33150552 H. sapiens         STQ[AT]TFALGFWHPPEACFL--V[IK]GV[ST]VGVYGG--YTS[N]PT[YSR]V[CDG]T[GH]A[AV]R[IT]D[PAQ]S[ER]D[LM]I[F]P[ATH]D[PT]LNRQ[GD]V[GT]Q[YSR]A[H]Y[TP]T[PS]Q[RA]
55251370 D. rerio          FLQ[AT]TFALGFWHPPEACFL--R[OK]GV[ST]VGVYGG--YTFAN[PT]Y[EE]V[CDG]KT[GH]A[AV]R[IT]D[PAQ]S[ER]D[LM]I[F]P[ATH]D[PT]LNRQ[GD]V[GT]Q[YSR]A[H]Y[TP]T[PS]Q[RA]
78709015 O. sativa        ONE[AT]TFALGFWHPPEACFL--R[IP]GV[RT]VGVYGG--N[LD]N[PT]Y[ED]V[CDG]AT[YSR]V[CDG]T[GH]A[AV]R[IT]D[PAQ]S[ER]D[LM]I[F]P[ATH]D[PT]LNRQ[GD]V[GT]Q[YSR]A[H]Y[TP]T[PS]Q[RA]
7580480 L. esculentum     BLE[AT]TFALGFWHPPEACFL--R[IG]GV[RT]VGVYGG--N[HD]P[AT]Y[LL]CSG[TT]E[PA]R[IT]D[PAQ]S[ER]D[LM]I[F]P[ATH]D[PT]LNRQ[GD]V[GT]Q[YSR]A[H]Y[TP]T[PS]Q[RA]
21593143 A. thaliana      SQQ[AT]TFALGFWHPPEACFL--R[IP]GV[RT]VGVYGG--I[VD]N[PT]Y[ED]V[CDG]T[GH]A[AV]R[IT]D[PAQ]S[ER]D[LM]I[F]P[ATH]D[PT]LNRQ[GD]V[GT]Q[YSR]A[H]Y[TP]T[PS]Q[RA]
1136793 B. napus         SQQ[AT]TFALGFWHPPEACFL--R[IP]GV[RT]VGVYGG--F[VD]N[PT]Y[ED]V[CDG]T[GH]A[AV]R[IT]D[PAQ]S[ER]D[LM]I[F]P[ATH]D[PT]LNRQ[GD]V[GT]Q[YSR]A[H]Y[TP]T[PS]Q[RA]

```

Fig. S19

```

7706511 H. sapiens SRSKIAHSSPPAFETTHAASIA--KRPEHNRSEAIVSCKCNGLGHVFDGPKFC--QSRITUPSSSLKFFVFKGK
55730640 P. pygmaeus SRSKIAHSSPPAFETTHAASIA--KRPEHNRRAEIVSCKCNGLGHVFDGPKFC--QSRITUPSSSLKFFVFKGK
27807643 M. musculus SRSKIAHSSPPAFETTHAASIA--KQPEKNRPEAIVSCKCNGLGHVFDGPKFC--QSRITUPSSSLKFFVFKGK
29648559 D. rerio SRSKIAHSSPPAFETTHAASIS--KQEE--RWGAYRSEIVSCKCNGLGHVFDGPKFC--LSTITUPSSSLKFFVFKV
57335062 Suberites sp. STKFPVHSSPPAFETTHSNSIS--KYNES--TSAIVRSEIVSCKCNGLGHVFDGPKFC--QSRITUPSSSLKFFVFKKGE
56314788 Azoarcus sp. SEHKFVSGCGWPSFTAAAPDNIE-TAIDRSHFVQREVLGHECQHLGHVFDGPPFC--LRYCINSASLLEKFEAS
85715266 Nitrobacter sp. SDAKFVSGCGWPSFTAAVDSHID-EHDTLSHGVRVREVLSSRCQHLGHVFDGPPFC--LRYCINSASLLEKFEK--
68246156 Magnetococcus sp. SDAKFVSGCGWPSFTAAVDSHID-EHDTLSHGVRVREVLSSRCQHLGHVFDGPPFC--LRYCINSASLLEKFEK--
71661450 T. cruzi SEMKFRSGCGWPSFTAAVDSHID-EHDTLSHGVRVREVLSSRCQHLGHVFDGPPFC--LRYCINSASLLEKFEK--
92877819 M. truncatula SITKFRSGCGWPSFTAAVDSHID-EHDTLSHGVRVREVLSSRCQHLGHVFDGPPFC--LRYCINSASLLEKFEK--
91216521 P. torquis SQSKFVSGCGWPSFTAAVDSHID-EHDTLSHGVRVREVLSSRCQHLGHVFDGPPFC--LRYCINSASLLEKFEK--
92907076 Mycobacterium sp. SSEKFRSGCGWPSFTAAVDSHID-EHDTLSHGVRVREVLSSRCQHLGHVFDGPPFC--LRYCINSASLLEKFEK--
32397214 R. baltica AKDKFVSGCGWPSFTAAVDSHID-EHDTLSHGVRVREVLSSRCQHLGHVFDGPPFC--LRYCINSASLLEKFEK--
56127980 S. enterica SHTRKFSGCGWPSFTAAVDSHID-EHDTLSHGVRVREVLSSRCQHLGHVFDGPPFC--LRYCINSASLLEKFEK--
91210995 E. coli SQTRKFSGCGWPSFTAAVDSHID-EHDTLSHGVRVREVLSSRCQHLGHVFDGPPFC--LRYCINSASLLEKFEK--
46133494 H. influenzae SNDKFRSGCGWPSFTAAVDSHID-EHDTLSHGVRVREVLSSRCQHLGHVFDGPPFC--LRYCINSASLLEKFEK--
57637573 S. epidermidis SEDKFRSGCGWPSFTAAVDSHID-EHDTLSHGVRVREVLSSRCQHLGHVFDGPPFC--LRYCINSASLLEKFEK--
46190343 B. longum SRDKFRSGCGWPSFTAAVDSHID-EHDTLSHGVRVREVLSSRCQHLGHVFDGPPFC--LRYCINSASLLEKFEK--

```

Fig. S20

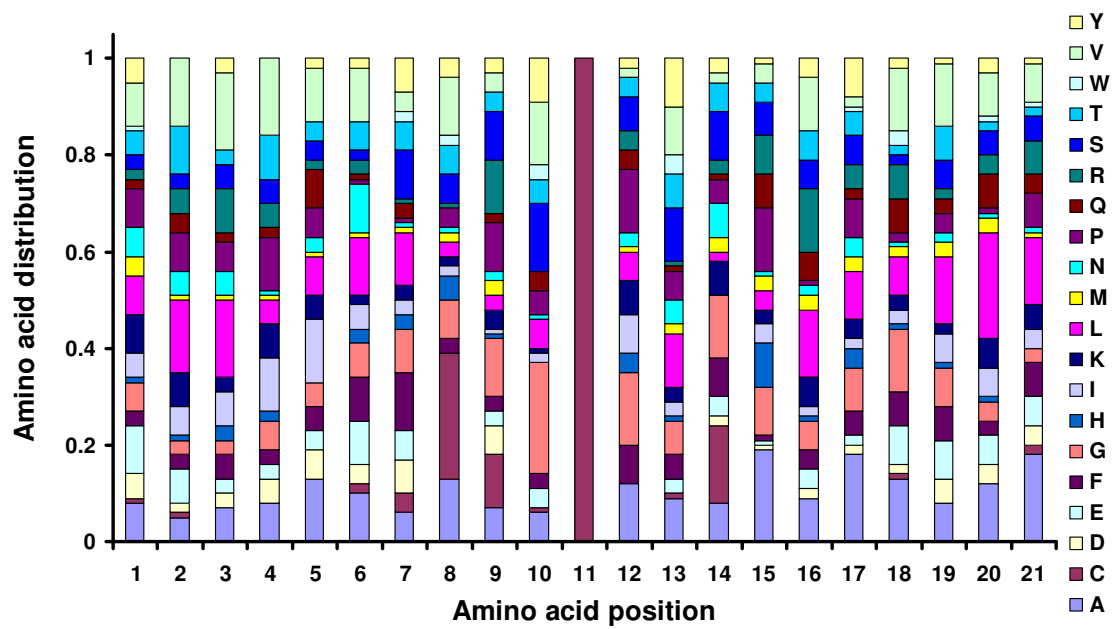
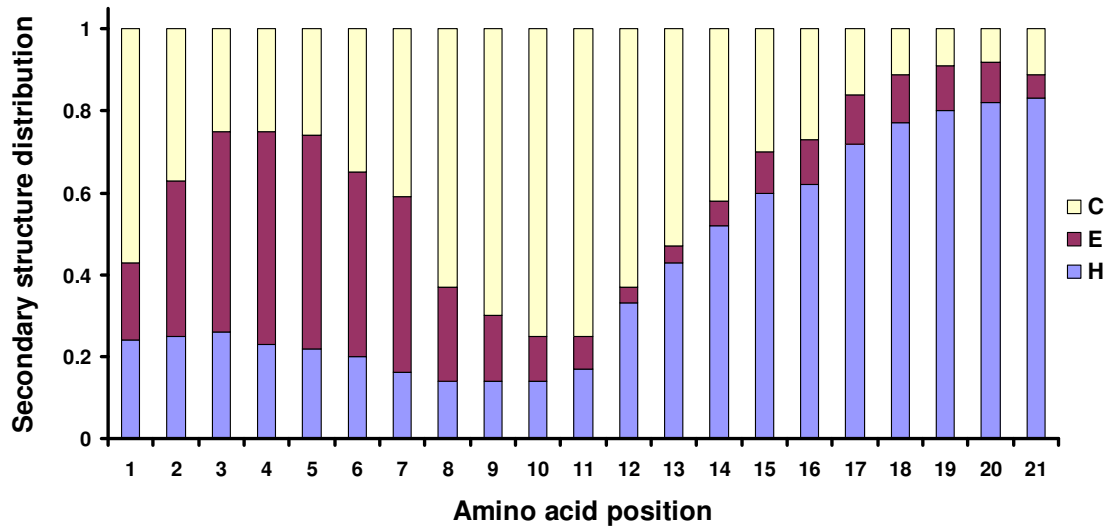


Fig. S21

a)



b)

