Research

Lateral gene transfer and parallel evolution in the history of glutathione biosynthesis genes

Shelley D Copley and Jasvinder K Dhillon

Address: Department of Molecular, Cellular, and Developmental Biology and the Cooperative Institute for Research in Environmental Sciences, University of Colorado at Boulder, Campus Box 216, Boulder, CO 80309, USA.

Correspondence: Shelley D Copley. E-mail: copley@cires.colorado.edu

Published: 29 April 2002

Genome ***Biology*** 2002, **3(5)**:research0025.1–0025.16

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/5/research/0025

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 28 December 2001
Revised: 22 February 2002
Accepted: 5 March 2002

Abstract

Background: Glutathione is found primarily in eukaryotes and in Gram-negative bacteria. It has been proposed that eukaryotes acquired the genes for glutathione biosynthesis from the alpha-proteobacterial progenitor of mitochondria. To evaluate this, we have used bioinformatics to analyze sequences of the biosynthetic enzymes γ-glutamylcysteine ligase and glutathione synthetase.

Results: γ-Glutamylcysteine ligase sequences fall into three groups: sequences primarily from gamma-proteobacteria; sequences from non-plant eukaryotes; and sequences primarily from alpha-proteobacteria and plants. Although pairwise sequence identities between groups are insignificant, conserved sequence motifs are found, suggesting that the proteins are distantly related. The data suggest numerous examples of lateral gene transfer, including a transfer from an alpha-proteobacterium to a plant. Glutathione synthetase sequences fall into two distinct groups: bacterial and eukaryotic. Proteins in both groups have a common structural fold, but the sequences are so divergent that it is uncertain whether these proteins are homologous or arose by convergent evolution.

Conclusions: The evolutionary history of the glutathione biosynthesis genes is more complex than anticipated. Our analysis suggests that the two genes in the pathway were acquired independently. The gene for γ-glutamylcysteine ligase most probably arose in cyanobacteria and was transferred to other bacteria, eukaryotes and at least one archaeon, although other scenarios cannot be ruled out. Because of high divergence in the sequences, the data neither support nor refute the hypothesis that the eukaryotic gene comes from a mitochondrial progenitor. After acquiring γ-glutamylcysteine ligase, eukaryotes and most bacteria apparently recruited a protein with the ATP-grasp superfamily structural fold to catalyze synthesis of glutathione from γ-glutamylcysteine and glycine. The eukaryotic glutathione synthetase did not evolve directly from the bacterial glutathione synthetase.

Background

Aerobic organisms produce intracellular thiols such as glutathione (GSH), homoglutathione [1], γ-glutamylcysteine (γ-Glu-Cys) [2], γ-glutamylcysteinylserine [3] and mycothiol [4] for protection against reactive oxygen species formed as by-products of aerobic metabolism. GSH is the most common of these. In addition to buffering the redox status of the cytoplasm and protecting biomolecules against oxidative damage,

GSH provides reducing equivalents to several enzymes (including ribonucleotide reductase [5], 3′-phosphoadenosine 5-phosphosulfate reductase [6] and arsenate reductase [7]), and serves as a substrate for glutathione-*S*-transferases, which detoxify potentially dangerous electrophiles.

GSH is found primarily in Gram-negative bacteria and eukaryotes, and only rarely in Gram-positive bacteria (Figure 1). It is not found in the Archaea or in amitochondrial eukaryotes such as *Entamoeba histolytica* [8], *Giardia duodenalis* [9], *Trichomonas vaginalis* [10] and *Trichomonas foetus* [9]. This distinctive pattern led to the proposal that the genes for GSH biosynthesis may have been transferred to eukaryotes from bacteria via the progenitor of mitochondria [8,11,12]. If this hypothesis is true, then the genes for GSH biosynthesis in eukaryotes should resemble those from alpha-proteobacteria, the modern relatives of the mitochondrial progenitor [13,14].

The biosynthesis of GSH (Figure 2) requires only two enzymes. γ-Glu-Cys ligase (GshA) catalyzes the formation of a peptide bond between the γ-carboxylate of glutamate and cysteine. GSH synthetase (GshB) catalyzes the subsequent formation of a peptide bond between the cysteinyl carboxylate of γ-Glu-Cys and the amino group of glycine. Each of these reactions requires hydrolysis of ATP to drive formation of the peptide bond.

We have analyzed the sequences of genes encoding GshA and GshB and have discovered that the evolutionary history of these genes is more complex than expected. Our results are consistent with two possible explanations for the distribution of GshA genes. The most likely possibility is that GshA arose in the bacterial domain, and the gene was transferred to eukaryotes at an early stage in their evolution. A second, less appealing, possibility is that a GshA gene was present in the last common ancestor and was subsequently lost from many organisms, primarily those that live under anaerobic conditions. GshB appears to have arisen independently within bacteria and eukaryotes subsequent to the acquisition of the GshA gene. Notably, in each domain, the scaffold typical of the ATP-grasp superfamily was utilized to provide GshB. Multiple examples of lateral transfer of the GshA gene are evident, the most dramatic being a trans-domain transfer from an alpha-proteobacterium to a plant sometime before 300 million years ago.

Results and discussion

GshA sequences fall into three distinct groups

We assembled an initial set of GshA sequences by searching the NCBI Protein database [15] using either GshA or glutamate-cysteine ligase as query words. The set was expanded using the output of BLAST [16] searches with the sequences from *Escherichia coli*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana*. Some of the sequences found in the BLAST searches correspond to hypothetical proteins whose functions have not been experimentally verified. For organisms that are known to synthesize GSH, these sequences are likely to encode GshA. However, the assignment of function is uncertain for the sequences from *Mycobacterium tuberculosis, Streptomyces coelicolor* and *Clostridium acetobutylicum* because these organisms are not known to synthesize GSH [4,17].

The sequences in the GshA set fall into three distinct groups (Table 1) that have no significant relationship to each other on the basis of pairwise sequence identities. The first group consists primarily of gamma-proteobacteria, the second of non-plant eukaryotes and the third primarily of flowering plants and alpha-proteobacteria. Pairwise sequence identities within each group range from 24 to 93% for group 1, from 32 to 98% for group 2, and from 45 to 93% for group 3. (Group 1 corresponds to the GCS entry in the Pfam database [18].) The clustering of GshA sequences into three groups was confirmed using PSI-BLAST [16]. PSI-BLAST is an iterative form of BLAST in which sequence information from hits found with an initial query sequence is incorporated into a profile that is used in subsequent searches. We carried out PSI-BLAST searches using GshA sequences from *E. coli* (group 1, gi12517129), *Drosophila melanogaster* (group 2, gi7290879), and *Mesorhizobium loti* (group 3, gi13475748). In each case, PSI-BLAST converged within a few iterations (three, two and five, respectively), and no members of the other groups were found, even with statistically insignificant scores (data not shown). Cyanobacterial sequences (not shown in Table 1) are found in the PSI-BLAST output for group 3, but are quite distantly related to the other group 3 sequences (typically less than 20% sequence identity).

The lack of significant overall sequence identity between the members of these groups could indicate that these proteins arrived at a common function by convergent evolution from different progenitors, or that the sequences have diverged so far that evolutionary relationships are no longer readily apparent. For truly related but very divergent proteins, it is often possible to identify locally conserved regions that are important for structure and/or function. Therefore, our next approach was to search for such motifs using Block Maker [19,20]. (It would be ideal to search for motifs with a divergent set of proteins in which the pairwise identities are less than about 40%, so that the motifs found are highly conserved because they are important for structure and/or function. However, the high pairwise identities between plant and alpha-proteobacterial sequences in group 3 made this impossible.) Using a query set containing six group 1 sequences, seven group 2 sequences, seven group 3 sequences, and four cyanobacterial sequences, Block Maker identified three blocks of conserved sequence (Figure 3). A search of the non-redundant database using a PSSM (position-specific scoring matrix) generated from these blocks retrieved a total of 81 sequences with *E*-values less than 10.

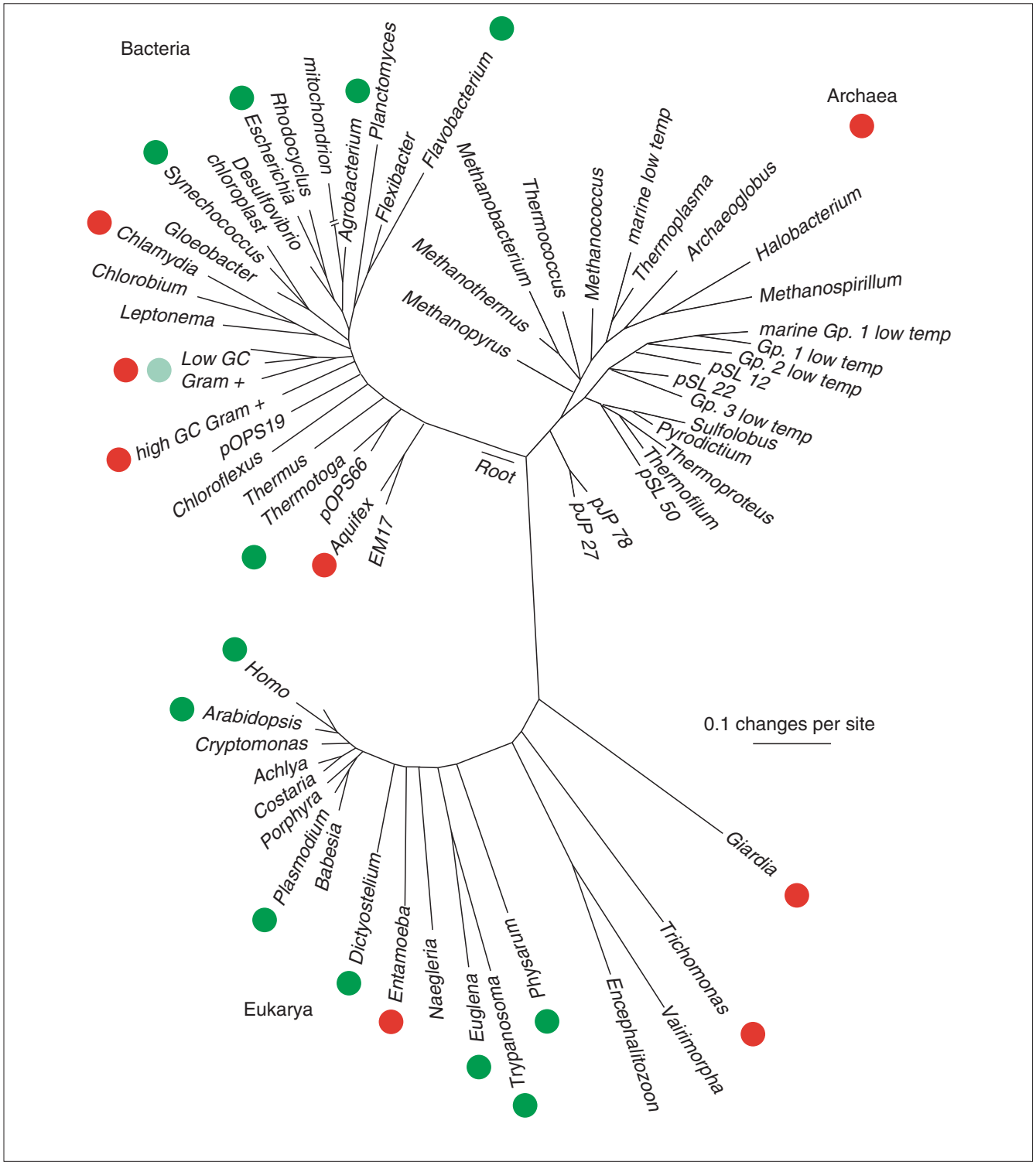

Figure 1
Distribution of glutathione mapped onto a universal tree of life based on 16S rRNA sequences. Information was obtained from an extensive search of the literature on the occurrence of glutathione in various organisms and was supplemented by data obtained from the sequence database. Green dots indicate that glutathione is present in the indicated organisms, whereas red dots indicate that glutathione is absent. The light green dot indicates that glutathione is found in some strains of some low-GC Gram-positive bacteria (*Streptococcus*, *Enterococcus*, *Clostridium* and *Listeria*) but not in others (*Staphylococcus*) [4]. Some strains of Gram-positive bacteria (for example, *Streptococcus thermophilus*, *Streptococcus agalactiae* and *Enterococcus faecalis* [4]) appear to synthesize GSH, whereas others (*Streptococcus mutans* [58]) just import it from the medium. In others (*Clostridium* and *Listeria*), this question has not been resolved [4].

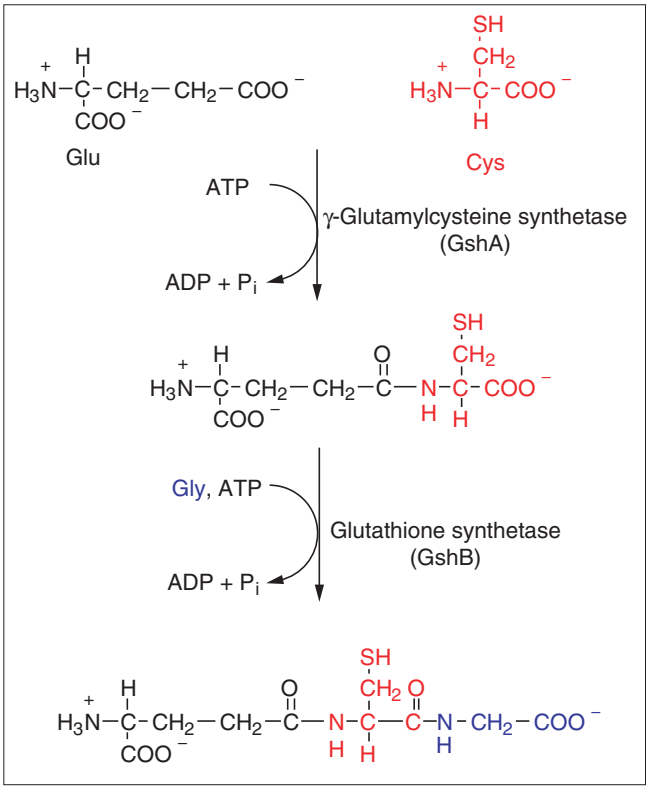

Figure 2
Pathway for the biosynthesis of glutathione.

No proteins with identified functions other than GshA were found with *E*-values less than 7. Thus, the PSSM was very effective at discriminating between GshA and non-GshA sequences in the database. Many of the sequences were duplicate entries of a single sequence or sequences from multiple strains of the same organism. Taking this into account, the search retrieved 44 sequences of known or putative GshAs with *E*-values less than 1. A subset of these sequences representing all three groups and cyanobacteria is shown in Table 2. The *p*-values for each occurrence of the three motifs, as well as the *E*-values for each sequence, are very low in all three groups. These data provide support for a real but distant relationship between the three groups of sequences, as well as the cyanobacterial sequences.

It is difficult to generate reliable phylogenetic trees in cases such as this in which only small blocks of conserved sequence can be identified. Figure 4 shows a tree generated from the three conserved blocks using the parsimony method in PAUP 4.0b [21]. The tree topology supports the findings reported above in that the sequences cluster into three distinct groups, with the cyanobacterial sequences being most closely associated with the group 3 (plant and alpha-proteobacteria) sequences. The available information does not provide reliable information about the order of branching within each of these groups. Similar findings were obtained using the neighbor-joining method in PAUP. A general association of eukaryotic and alpha-proteobacterial sequences that would support the hypothesis that the eukaryotic gene arose by transfer from the mitochondrial progenitor is not seen. The striking association of plant sequences with alpha-proteobacterial sequences will be addressed below.

Possible explanations for the distribution of GshA: lateral transfer versus massive gene loss

Figure 5 shows the distribution of known and putative GshA genes mapped onto a universal tree of life generated using rRNA sequences. Two possible scenarios could account for the observed pattern of distribution. First, a GshA gene might have arisen early in either the bacterial or eukaryotic lineage and undergone a lateral transfer to the other domain. Alternatively, a GshA gene could have been present in the last common ancestor, but have been lost in nearly all Archaea, the deepest-branching eukaryotes and many bacteria. Although the possibility of massive loss of GshA at first glance seems unlikely, careful perusal of Figures 1 and 5 reveals that most of the genera that lack GSH and/or GshA lack aerobic metabolism and therefore do not need GSH for protection against reactive oxygen species. Many Archaea (all of the methanogens, as well as *Archaeoglobus*, *Pyrodictium*, *Thermoproteus* and *Thermofilum*) are anaerobes. *Thermoplasma*, although facultatively aerobic, respires sulfur rather than oxygen. The deepest-branching eukaryotes (*Giardia, Trichomonas, Vairimorpha* and *Encephalitozoon)* lack mitochondria, and therefore aerobic metabolism. *Entamoeba histolytica* has apparently lost its mitochondria and now lacks aerobic respiratory pathways. Bacteria such as *Aquifex* and *Thermotoga* are anaerobes. Thus, the possibility of massive gene loss in these lineages is quite plausible. However, a problem with this hypothesis is that it requires that the GshA gene, which appears to be important only to aerobes, must have originated in the last common ancestor before the emergence of O_2, which occurred only after the evolution of cyanobacteria about 2.6 billion years ago [22].

The possibility that GshA arose early in the bacterial lineage and then was transferred to the eukaryotic lineage and into at least one archaeon (*Halobacterium*) is perhaps more appealing. GshA may have arisen in cyanobacteria concomitant with the origin of oxygenic photosynthesis, a particularly attractive hypothesis because this would have been the first time protection against reactive oxygen species would have been important. Later, as proteobacteria developed aerobic metabolic processes that took advantage of the growing O_2 concentration in the atmosphere, the consequent production of reactive oxygen species would have given a selective advantage to microbes that acquired a GshA gene by lateral gene transfer from cyanobacteria. GshA would have been advantageous to aerobic eukaryotes as well, providing selective pressure for the acquisition and retention of a bacterial GshA gene. Transfer of a GshA gene from the

Table 1

Three groups of known and putative GshAs

Group 1 Primarily gamma-proteobacteria		Group 2 Eukaryotes (except plants)		Group 3 Primarily alpha-proteobacteria and plants*	
Escherichia coli *Salmonella typhimurium* *Vibrio cholerae* *Pseudomonas aeruginosa* *Proteus mirabilis* *Pasteurella multocida* *Buchnera sp.* APS	Gamma-proteo-bacteria	*Homo sapiens* *Mus musculus* *Rattus norvegicus*	Mammals	*Mesorhizobium loti* *Sinorhizobium meliloti* *Agrobacterium tumefaciens* *Caulobacter crescentus* *Zymomonas mobilis*	Alpha-proteo-bacteria
(*Clostridium acetobutylicum*)	Low-GC Gram-positive	*Onchocerca volvulus* *Caenorhabditis elegans*	Nematode	*Arabidopsis thaliana* *Pisum sativum* *Picea abies* *Glycine max* *Lycopersicon esculentum*	Plants
		Drosophila melanogaster	Fly	*Xylella fastidiosa*	Gamma-proteo-bacterium
		Schizosaccharomyces pombe *Candida albicans* *Saccharomyces cerevisiae*	Fungi	*(Mycobacterium tuberculosis)* *(Streptomyces coelicolor)*	High-GC Gram-positive
		Plasmodium falciparum	Apicom-plexa	(*Streptococcus pneumoniae* TIGR4)	Low-GC Gram-positive
		Leishmania tarentolae *Trypanosoma cruzi*	Kineto-plastids	*Halobacterium* sp. NRC-1	Archaeon

Members of each group have no statistically significant relationship with members of other groups on the basis of pairwise sequence identities. Parentheses designate cases for which assignment of function is tentative because the ability of these organisms to synthesize γ-Glu-Cys has not been demonstrated. *PSI-BLAST outputs for the Group 3 GshAs include proteins not listed here that may be distant homologs in *Archaeoglobus fulgidus* (gi11499888), *E. coli* (gi15800294), *S. typhimurium* (gi16763960), *S. enterica* (gi16759543), and *P. aeruginosa* (gi15597377). These proteins have only about 15% identity to group 3 GshAs over part of their lengths, and their functions are unknown.

alpha-proteobacterial progenitor of mitochondria, as postulated previously [8,11,12], is one possible mechanism for the spread of GshA into eukaryotes. Our analysis does not show a significant association between eukaryotic and alpha-proteobacterial sequences that would support this mechanism (Figure 4). However, the extreme divergence of the sequences limits our ability to resolve phylogenetic relationships, so this mechanism remains a possibility. There are other possible mechanisms for an early trans-domain lateral gene transfer, as well. For example, Doolittle has proposed that protists that consume bacteria as food may incorporate bacterial genes into their genomes [23].

Lateral transfer of GshA genes

Lateral transfer of genes between bacteria is recognized to occur very frequently, especially for genes involved in metabolism [24-26]. Lateral transfers of genes into eukaryotic nuclear DNA from the alpha-proteobacterial progenitor of mitochondria and into plant nuclear DNA from the cyanobacterial progenitor of chloroplasts are well documented. Only a few other cases of lateral transfers between domains have been reported. Examples include transfer of the phosphoglucose isomerase gene from a eukaryote to a bacterium [27] and displacement of several aminoacyl-tRNA synthetases in parasitic and symbiotic bacteria by eukaryotic and archaeal counterparts [28]. Our results show clear evidence of a trans-domain transfer of a GshA gene from alpha-proteobacteria to plants. Figure 4 shows that the data provide 100% bootstrap support to the clustering of plant sequences with the group 3 alpha-proteobacterial sequences, and Table 3 illustrates the strikingly high pairwise sequence identities between plant and alpha-proteobacterial sequences. Many of the alpha-proteobacteria are either plant pathogens (for example, *Agrobacterium tumefaciens*) or plant commensals (for example, *Sinorhizobium meliloti* and *Zymomonas mobilis*). The close physical association between these bacteria and plants apparently provided an opportunity for lateral transfer of the GshA gene, perhaps more than 300 million years ago. A partial sequence is available for GshA from the conifer *Picea abies* (gi12580873). This sequence has 77% identity to the *A. thaliana* sequence and thus belongs to group 3. Conifers diverged from the lineage leading to flowering plants 300-350 million years ago [29]. Thus, lateral transfer of the alpha-proteobacterial GshA gene may have occurred before the divergence of conifers. Additional sequences from other

```
BGSHa.Ec  25 GLERETLRVNADGT (107) ISGVHYNFSL (123) IDKDGKRLQINSNVLQIENELYAPIRPK
BGSHa.Pa  27 GIERECLRVDSDGK (107) IAGIHYNFSL (122) TKQDGEWVQLNTNILQIENEYYSSIRPK
BGSHa.Ps  27 GIEREALRVDVQGE (107) IAGIHYNYSL (122) VFAQGEWRQLNANLLQLDSEYYALARPK
BGSHa.Ba  25 GIERETLRIQKNGH (107) ISGIHYNFSL (124) KDEHGNFKQLNTNILQIENELYTQIRPK
BGSHa.Ca  27 GVERESQRVNYSGD (107) ISGIHYNFSF (130) IYKDGVQIQLNGNLLQSESEFYAPIRFK
Bunk.Pm   22 GLEKESQRVTADGA (104) VSGIHYNFQL (208) LLSMLEQIGAEPELFEIVKEKLTQFTDP
EGSHa.Ca  48 GDEVEYMLVDFDET (181) IYMDSMGFGM (186) INQDNNLENDHFENIQSTNWQTLRFKPP
EGSHa.Hs  48 GDEVEYMLVSFDHE (174) IYMDAMGFGM (141) IHLDDANESDHFENIQSTNWQTMRFKPP
EGSHa.Pb  48 GDEIEYIIIRNDDK (402) VYLDAMFFGM (183) YKEKVLSSHQHFENFQSTNWNSVRFKPP
EGSHa.Sc  48 GDELEYMVVDFDDK (191) IYMDSMGFGM (163) LNQDNKTSSNHFENIQSTNWQTLRFKPP
EGSHa.Sp  48 GDEIECIVVSMDDK (191) IYMDSMGFGM (135) ILQDNSVSNAHFENLNSTNWQSMRFKPP
EGSHa.Tc  51 GEEVEHQLVRVEGG (256) IYMDCMAFGM (135) IDIDDTTHTEHFENIQSTNWQSVRLKPP
hyp.Ce    48 GDEIEYTIVKFDDA (171) IYMDHMGFGM (141) IEQDDEKSSEHFETIQSSNWMNMRFKPP
EGSHa.At111 GTEHEKFGFEVNTL (125) TCTVQVNLDF ( 22) LFANSPFTEGKPNGFLSMRSHIWTDTDK
BGSHa.Cc  34 GAEHEKFGFYLGSH (127) TCTVQANLDF  (22) LFANSPFTEGKPNGFLSARANVWTDTDP
BGSHa.Xf  35 GTEHEKFGFRLDDL (127) TCTVQVNLDY ( 22) LFANSPFTEGKPNGYLSYRSHIWTDTDP
BGSHa.At  34 GTEHEKFAFFRKDN (130) TCTIQVNLDF  (22) LFASSPFTEGKPNGLLSWRGDIWRDTDN
hyp.Halo  16 GVEEEFFVVDEHGV (103) TAGLHIHVGV ( 22) LSANSPYWNGFDTGLASARAKIFEGLPN
hyp.Mt    21 GVEWEFALVDSQTR (100) IWGVHVHVGI ( 22) LSASSPWWGGEDTGYASNRAMMFQQLPT
hyp.Stc    5 GVEEELLLVDPATG (102) VLGCHVHVSV ( 72) TGTVLDDGMVYFDVRLSQRYPTVEFRVA
hyp.Nostoc6 GFEIEIYTGTPQGE ( 96) TASVHINIGI ( 22) LSASSPFLDGKTTGYHSTRWGLFPQTPS
hyp.Prochl9 GFEVELFTGRFSGE ( 95) TTSVHINLGL ( 22) LSASSPFLDGQPTGSHSQRWLQFPLTPE
hyp.Syn   6 GLEVEIYTGKKTGE ( 97) TASIHINIGI ( 22) LSASSPFLNGQVTGYHSSRWQMFPKTPQ
hyp.Synech9 GFEVELFTGRPDGT ( 95) TASIHINLGI ( 22) LSASSPFLGGELTGHHSQRWHQFPLTPR
```

Figure 3
Blocks identified in GshA homologs by Block Maker. Sequences from groups 1, 2 and 3 are highlighted in red, green and blue, respectively. Cyanobacterial sequences are underlined. BGshA.Ec, GshA, *E. coli* (gi12517129); BGshA.Pa, GshA, *Pseudomonas aeruginosa* (gi11348607); BGshA.Ps, putative GshA, *Pseudomonas* sp. (gi6634496); BGshA.Ba, GshA, *Buchnera aphidicola* (gi11386815); BGshA.Ca, GshA (putative), *Clostridium acetobutylicum* (gi15024489); Bunk.Pm, unknown protein, *Pasteurella multocida* (gi12721383); EGshA.Ca, GshA, *Candida albicans* (gi12002873); EGshA.Hs, GshA, *Homo sapiens* (gi 4557625); EGshA.Pb, GshA, *Plasmodium berghei* (gi4713921); EGshA.Sc, GshA, *Saccharomyces cerevisiae* (gi6322360); EGshA.Sp, GshA, *Schizosaccharomyces pombe* (gi2130201); EGshA.Tc, GshA, *Trypanosoma cruzi* (gi3747103); hyp.Ce, hypothetical protein, *Caenorhabditis elegans* (gi7500706); EGshA.At, GshA, *Arabidopsis thaliana* (gi1742963); BGshA.Cc, GshA, *Caulobacter crescentus* (gi13425126); BGshA.Xf, GshA, *Xylella fastidiosa* (gi11282603); BGshA.At, GshA, *Agrobacterium tumefaciens* (gi15155610); hyp.Halo, hypothetical protein, *Halobacterium* sp. NRC-1 (gi10580902); hyp.Mt, hypothetical protein, *Mycobacterium tuberculosis* (gi6831717); hyp.Stc, hypothetical protein, *Streptomyces coelicolor*A3(2) (gi8246826); hyp.Nostoc, hypothetical protein, *Nostoc punctiforme* (DOE 63737, Contig399 revised gene10 protein); hyp.Prochl, hypothetical protein, *Prochlorococcus marinus* (DOE 59919, Contig26 gene 578); hyp.Syn, hypothetical protein, *Synechocystis* sp. (strain PCC6803) (gi7469602); hyp.Synech, hypothetical protein, *Synechococcus* sp. WH8102 (DOE 84588, Contig72 revised gene139 protein).

conifers and more primitive plants are needed to pin down the timing of this transfer.

Lateral gene transfer from alpha-proteobacteria to plants is particularly intriguing because transfer and retention of a foreign gene in a sophisticated multicellular organism is more difficult than in bacteria. Bacteria can acquire DNA from their environment in multiple ways (transformation, transduction and conjugation) [26]. Furthermore, a transferred gene can be easily transmitted to progeny after recombination into genomic or plasmid DNA. However, known mechanisms for transfer of DNA into plants are more limited. The best understood mechanism is the transfer of T-strand DNA from the Ti-plasmid of *Agrobacterium tumefaciens* into wounded plant tissues, a process resulting in the formation of tumors [30]. It is not known whether foreign genes can be transferred into plants by this mechanism in nature, but such a process is plausible. Perpetuation of a transferred gene is also not as easily achieved in seed plants as it is in bacteria, because the gene must be incorporated into genomic DNA in apical meristem cells, undifferentiated stem cells that produce new organs, including the cones or flowers that generate male and female gametes. An interesting issue is whether the group 3 alpha-proteobacterial gene displaced an ancestral group 2 eukaryotic gene, or whether the ancestral gene was first lost, allowing the alpha-proteobacterial gene to fill the functional gap.

Additional potential cases of lateral gene transfer are also suggested by our data. GshA from *Xylella fastidiosa*, a gamma-proteobacterium, clusters with group 3, rather than with the other gamma-proteobacterial sequences in group 1. This organism is a plant pathogen, and its physical association with plants has apparently provided an opportunity for gene transfer, either from a plant or, more likely, from an associated alpha-proteobacterium. Sequences from *Halobacterium* sp.

Table 2

***p*-values for blocks found in some known and putative GshAs**

Group	Organism	gi	*E*-value	*p*-value block 1	*p*-value block 2	*p*-value block 3
1	*Pseudomonas aeruginosa*	15600396	1.6e-18	6.5e-12	1.4e-09	8.4e-16
	Pseudomonas sp.	6634496	3.8e-13	4.2e-10	1.9e-08	3.3e-13
	Escherichia coli	15803207	4.7e-15	1.2e-09	3.3e-09	8.1e-15
	Salmonella typhimurium	16421367	5.9e-14	1.2e-09	3.3e-09	1.1e-13
	Buchnera aphidicola	11386815	6.2e-14	6.6e-09	3.0e-09	2.3e-14
	Vibrio cholera	15640578	7.6e-10	4.9e-07	3.5e-08	4.7e-13
	Clostridium acetobutylicum	15894817	1.5e-13	6.0e-09	1.7e-09	1.4e-13
	Yersinia pestis	16123454	3.6e-13	1.9e-08	3.3e-09	4.5e-12
2	*Candida albicans*	12002873		1.5e-11	2.3e-09	8.3e-20
	Schizosaccharomyces pombe	2130201	5.2e-19	2.2e-10	2.3e-09	2.9e-18
	Homo sapiens	4557625		1.9e-11	1.8e-09	1.3e-18
	Onchocerca volvulus	7328216	6.8e-19	2.2e-10	6.5e-12	1.2e-15
	Drosophila melanogaster	7290879	4.3e-17	8.5e-12	1.1e-6	1.0e-17
	Leishmania tarentolae	1743291	2.2e-16	1.3e-07	2.0e-09	2.1e-18
	Caenorhabditis elegans	7500706	1.3e-20	2.4e-10	6.5e-12	1.8e-17
	Trypanosoma cruzi	3747103	2.2e-16	6.0e-09	2.0e-09	4.4e-17
	Saccharomyces cerevisiae	312704		3.2e-12	2.3e-09	2.1e-19
	Plasmodium berghei	4713921	5.2e-14	5.7e-10	1.3e-07	7.7e-16
3	*Caulobacter crescentus*	16127644	3.2e-14	5.5e-09	3.1e-07	2.1e-16
	Xylella fastidiosa	15838029	3.1e-17	3.2e-11	1.3e-07	6.4e-17
	Brassica juncea	3688156	2.3e-15	3.8e-09	4.7e-08	1.4e-16
	Arabidopsis thaliana	16226411	4.0e-15	3.8e-09	4.7e-08	1.4e-16
	Agrobacterium tumefaciens	15887998	6.3e-15	3.4e-09	4.4e-08	4.2e-16
	Sinorhizobium meliloti	15964523	1.3e-14	6.0e-09	4.4e-08	4.2e-16
	Bradyrhizobium japonicum	8708927	2.7e-10	9.8e-07	9.3e-07	4.5e-15
	Lycopersicon esculentum	3913791	2.3e-14	2.9e-09	4.7e-08	1.2e-15
	Mesorhizobium loti	13475748	2.1e-14	7.1e-10	4.4e-08	6.9e-15
	Glycine max	10130004	2.4e-13	5.2e-08	4.7e-08	3.8e-15
	Pisum sativum	6651031	2.9e-13	3.3e-08	4.7e-08	1.7e-15
	Medicago truncatula	11386873	3.9e-13	3.3e-08	4.7e-08	2.1e-15
	Mycobacterium tuberculosis	15607574	1.6e-14	9.3e-09	3.9e-09	8.1e-15
	Halobacterium sp. NRC-1	15790414	1.9e-12	1.4e-09	9.7e-08	3.5e-13
Cyano-bacteria	*Synechocystis* sp.	16331213	2.0e-15	6.6e-08	4.7e-08	1.0e-17

E-values are given for the entire sequence except for cases in which the algorithm found a second occurrence of block 1 with a low and probably insignificant *p*-value and included that *p*-value in the calculation of the *E*-value.

NRC-1 (archaeon), *Mycobacterium tuberculosis* and *Streptomyces coelicolor* (high-GC Gram-positive bacteria), and *Streptococcus pneumoniae* (low-GC Gram-positive bacterium) cluster with the plant and alpha-proteobacterial sequences in group 3, and a sequence from *Clostridium acetobutylicum*, a low-GC Gram-positive bacterium, clusters with the gamma-proteobacterial sequences in Group 1. (Note that synthesis of γ-glutamyl cysteine has been demonstrated in *Halobacterium* sp. NRC-1 [31], but not in *Mycobacterium tuberculosis*, *Streptomyces coelicolor* or *Streptococcus pneumoniae*.) The occurrence of GshA homologs in these organisms could reflect persistence of an ancestral GshA in only some genera in the Archaea and in the Gram-positive bacteria, or could be the result of lateral transfer into a limited number of organisms. As discussed above, it is difficult to distinguish between these two explanations, although the lateral gene transfer hypothesis is most appealing.

Why are GshA sequences so divergent?

The three groups of GshA sequences are so divergent that it was difficult to demonstrate an evolutionary relationship between them. This level of sequence divergence is unexpected, and warrants some thought. Three obvious factors contribute to divergence of sequence in orthologs. First, the organisms being compared may be very distant. This explanation is probably not sufficient to explain the sequence divergence in the GshAs. The gamma-proteobacteria represented in group 1 are reasonably closely related to the alpha-proteobacteria in group 3, and the crown eukaryotes in group 2 are reasonably closely related to the plants in group 3.

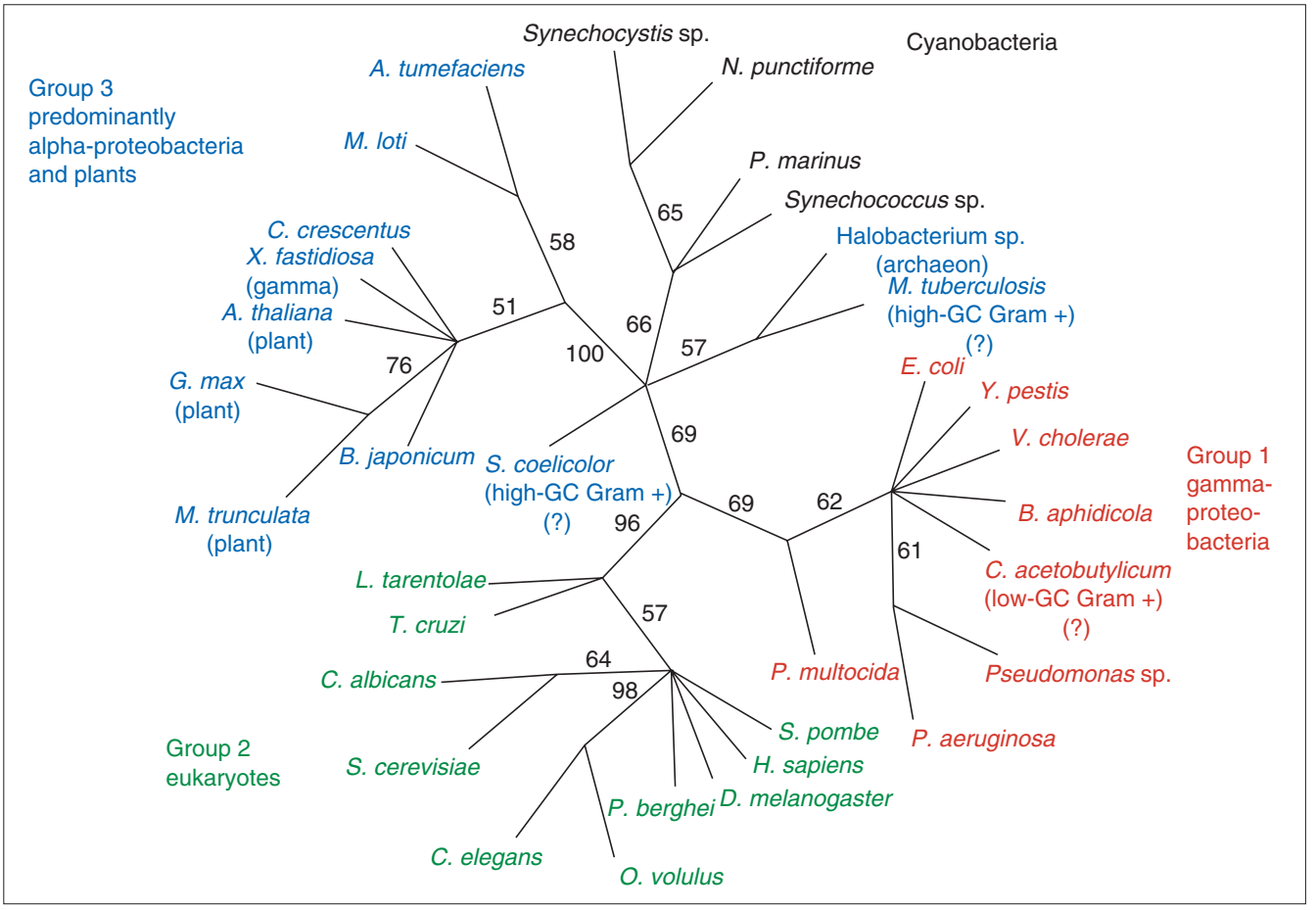

Figure 4
Phylogenetic tree constructed from the blocks identified by Block Maker using the parsimony method in PAUP 4.0b. Bootstrap values greater than 50% are indicated. Sequences from groups 1, 2 and 3 are colored red, green and blue, respectively. Accession numbers for putative GshA proteins from cyanobacteria are: *Synechocystis* sp., gi7469602; *Nostoc punctiforme*, gnl|DOE_63737|Contig399; *Prochlorococcus marinus*, DOE 59919 Contig26 gene 578; *Synechococcus* sp. , nl|DOE_84588|Contig72. (Functions for these cyanobacterial proteins have not been experimentally verified. However, cyanobacteria are known to contain GSH [59] and the function of GshB in *Synechococcus* sp. PCC 7942 has been verified [60]. Thus, it is likely that these sequences indeed encode GshAs.) Other organism names and accession numbers are given in Table 2. Question marks indicate genera for which the ability to synthesize glutathione has not been demonstrated.

Orthologous relationships between these groups can often be identified. Of the 1,923 COGs (clusters of orthologous groups) [32-34] identified in *P. aeruginosa* (a gamma-proteobacterium), 80% are also found in the combined group containing *Mesorhizobium loti* and *Caulobacter crescentus* (both alpha-proteobacteria). As a specific example, the GshB sequences from *E. coli* (a gamma-proteobacterium) and *C. crescentus* have 41% identity, and those from *Arabidopsis* and human have 43% identity. Detection of orthologs in even more distantly related organisms is also possible in many cases. We have used PSI-BLAST to find orthologs of enzymes that use glutathione (including glutaredoxins, glutathione-*S*-transferases, glutathione reductases and glutathione peroxidases) in both bacteria and eukaryotes (data not shown).

A second reason that homologous sequences may be very divergent is that little selective pressure has been required to maintain function at the level required for the organism to succeed. This situation may occur if the reaction being catalyzed is not very demanding, or if the product of the reaction does not contribute to the fitness of the organism in an important way. An example of the first scenario is *o*-succinylbenzoate synthase, which catalyzes the dehydration of 2-hydroxy-6-succinyl-2,4-cyclohexadiene carboxylate to form *o*-succinylbenzoate synthase in the biosynthetic pathway for menaquinone synthesis. *o*-Succinylbenzoate synthases from various organisms have very low pairwise sequence identities compared to those seen for other members of the enolase superfamily. The low sequence identities in these enzymes have been interpreted as reflecting relatively low constraints upon the sequence because the reaction is quite facile even in the absence of the enzyme (because it forms an aromatic product), and thus the enzyme is not required to provide a great deal of assistance [35]. This

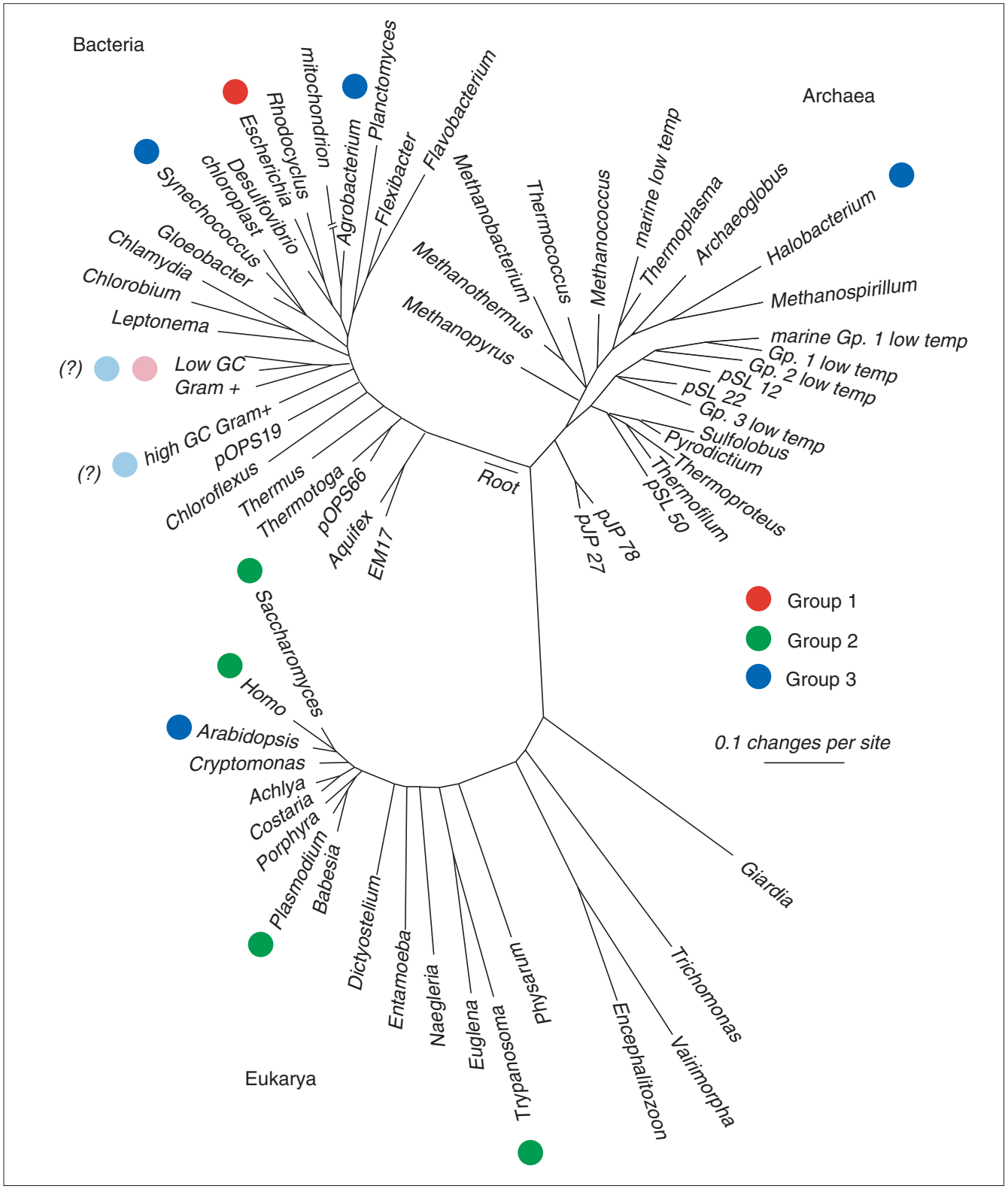

Figure 5
Mapping of GshA genes and homologs onto a universal tree of life generated from 16S rRNA sequences. Group 1, 2 and 3 sequences are designated by red, green and blue dots, respectively. Light blue or light red dots indicate that GshA homologs are found in only some genera in the indicated group. Question marks indicate that some of the Gram-positive bacteria are not known to synthesize GSH, so the functions of the GshA homologs in these cases are uncertain.

Table 3

Pairwise sequence identities between GshA sequences from plants and alpha-proteobacteria

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. *M. loti*	100												
2. *C. crescentus*	50.4	100											
3. *X. fastidiosa*	54.7	56.3	100										
4. *Z. mobilis*	44.4	51.2	53.3	100									
5. *B. japonicum*	53.3	53.4	52.9	48.4	100								
6. *A. tumefaciens*	71.1	50.7	50.1	45.5	53.0	100							
7. *S. melliloti*	72.2	51.9	51.3	45.0	54.2	79.6	100						
8. *L. esculentum*	**55.1**	**58.1**	**56.2**	**48.5**	**55.6**	**55.5**	**53.5**	100					
9. *M. truncatula*	**55.1**	**56.5**	**56.2**	**49.2**	**55.6**	**55.3**	**53.7**	78.9	100				
10. *P. sativum*	**55.3**	**56.3**	**56.4**	**49.4**	**55.8**	**55.7**	**54.2**	81.0	93.4	100			
11. *B. juncea*	**55.5**	**57.2**	**57.5**	**49.0**	**54.4**	**55.7**	**53.5**	79.3	80.2	82.8	100		
12. *P. vulgaris*	**56.0**	**56.8**	**56.6**	**49.4**	**55.4**	**55.1**	**53.8**	78.7	84.7	86.5	81.1	100	
13. *A. thaliana*	**56.0**	**58.3**	**57.3**	**49.2**	**55.6**	**56.8**	**53.7**	78.4	80.1	82.6	92.6	80.3	100

1-7 are bacteria and 8-13 are plants. The bold numbers highlight the pairwise sequence identities between bacterial and plant sequences.

scenario is unlikely to account for the divergence of the GshA genes, as formation of a peptide bond is a quite difficult reaction. In fact, members of the ATP-grasp superfamily that catalyze comparable reactions (that is, GshBs (see further below), ribosomal S6 modification enzymes, D-Ala-D-Ala ligases) are sufficiently well conserved to be easily detected by PSI-BLAST, even though they utilize different substrates [36]. With respect to the second scenario, it is clear that the ability to synthesize GSH provides a significant advantage. Bacteria and yeast that lack functional GshA are viable, but are hypersensitive to oxidative damage [37,38]. GshA-deficient strains of *A. thaliana* are viable, but are hypersensitive to cadmium [39]. Mice in which γ-Glu-Cys ligase has been knocked out die before gestational day 13 [40]. Thus, low selection pressure cannot account for the high levels of divergence among the GshA sequences.

Finally, sequences of homologs may diverge if they are subject to different selective pressures in different lineages. This might occur if the protein has a second function (a 'moonlighting' function) in some lineages and is subject to selective pressure that alters regions of the protein involved in that function. Alternatively, if a protein interacts with other proteins, then differences in those partner proteins will drive changes in the regions of the protein involved in the interaction. The possibility that the high level of divergence in GshA proteins is due to one of these factors is intriguing and worth experimental exploration.

GshB: a different story

A set of GshBs was assembled by searching the NCBI protein database using either GshB or glutathione synthetase as query words, and from the outputs of BLAST searches with the *E. coli* and *S. cerevisiae* proteins as query sequences. The sequences in the GshB set fall into two distinct groups, corresponding to bacteria and eukaryotes, that have no significant relationship to each other on the basis of pairwise sequence identities. As seen with GshA, PSI-BLAST searches with GshB sequences from either group did not find GshB sequences from the other group. A PSI-BLAST search with the human GshB converged after two iterations. The output contained only eukaryotic GshBs, a few putative homoglutathione synthestases, and a few eukaryotic proteins of unknown function. No bacterial GshBs were found in the output. After multiple iterations, a PSI-BLAST search using the *E. coli* GshB (gi121663) as an initial query sequence found 509 sequences of enzymes in the ATP-grasp superfamily, to which the bacterial enzyme is known to belong, but no eukaryotic GshB sequences. The ATP-grasp superfamily [41] includes at least 15 families of enzymes that catalyze formation of a bond between a carboxylate group of one substrate and an amino, imino or thiol group of a second substrate. The bacterial GshBs are most closely related to ribosomal S6 modification enzymes, which catalyze the addition of glutamate to the carboxyl terminus of ribosomal protein S6. Other members of the superfamily include carbamoyl phosphate synthases, cyanophycin synthetases and D-Ala-D-Ala ligases. Notably, many members of the ATP-grasp superfamily were found in eukaryotes. For example, *A. thaliana* has at least three superfamily members in the PSI-BLAST output, including carbamoyl phosphate synthetase, 3-methylcrotonyl-CoA carboxylase, and acetyl CoA carboxylase. However, the *Arabidopsis* GshB is not found in the output. Thus, ATP-grasp superfamily members in eukaryotes are more closely related to bacterial GshBs than to eukaryotic GshBs.

Crystal structures of GshBs from *E. coli* [42] and human [43] are available, so we are able to consider the evolutionary

relationship between the two groups of sequences in the context of the structural information (Figure 6). The *E. coli* and human enzymes have the same overall structural fold, which is typical of the ATP-grasp superfamily [36]. The human enzyme has a number of insertions, and in addition, has a circular permutation of the sequence that results in the movement of part of the carboxy-terminal domain to the amino-terminal domain. The sequence identity between the *E. coli* and human enzymes is strikingly low. Even in the regions that can be structurally superimposed, it is only 10%. Furthermore, only 2 out of 11 residues involved in binding GSH are conserved between the two enzymes [43].

The structural similarity between the *E. coli* and human GshBs indicates that the bacterial and eukaryotic enzymes could be related to each other, but is not sufficient to prove that they are related to each other because the common structure might have arisen by convergent evolution from different progenitors. This consideration is especially worth noting in light of the very low conservation of residues in the active site. Consequently, we looked for evidence of an evolutionary relationship by attempting to find proteins that might be distantly related to the bacterial and eukaryotic GshB proteins and therefore might bridge the sequence gap between them using the Shotgun algorithm [44], which was designed to facilitate searches for distant relations between proteins. The algorithm performs a BLAST search with each of a set of query sequences. It then sorts the hits found by all of the proteins, and for each hit, identifies the query sequences that found that hit. A hit that is found by multiple members of two distinct groups of proteins, even with low BLAST scores, can be examined closely to determine whether it provides a link between the groups.

We ran Shotgun with a query set containing six eukaryotic GshBs, eight bacterial GshBs, eight S6 modification enzymes (four from bacteria and four from archaea), and four proteins of unknown function (three from bacteria and one from the archaeon *Methanococcus jannaschii*). Sample outputs are shown in Figure 7. Generally, sequences were found by either bacterial GshBs and the S6 modification enzymes, or by eukaryotic GshB proteins, but not by both groups. There was only one exception, which was a protein (a putative GshB from *Medicago truncatula*, gi 4808539) found by all six of the eukaryotic GshBs with BLAST scores above 194 (BLAST probabilities less than 5.2e-31), and one bacterial GshB, with a BLAST score of 63 (BLAST probability of 1.0). Given the number of bacterial GshB sequences in the query set, the finding of one bacterial sequence with an insignificant *p*-value can be regarded as a chance occurrence. Thus, we were unable to establish any relationship between the two groups of GshB sequences using Shotgun.

Demonstration of conserved sequence motifs would provide supporting evidence for an evolutionary relationship between eukaryotic and bacterial GshBs. In this case, the Block Maker algorithm used to analyze the GshA proteins did not perform well. This algorithm begins by constructing a ClustalW [45] multiple sequence alignment. The circular permutation of the eukaryotic sequences with respect to the bacterial sequences,

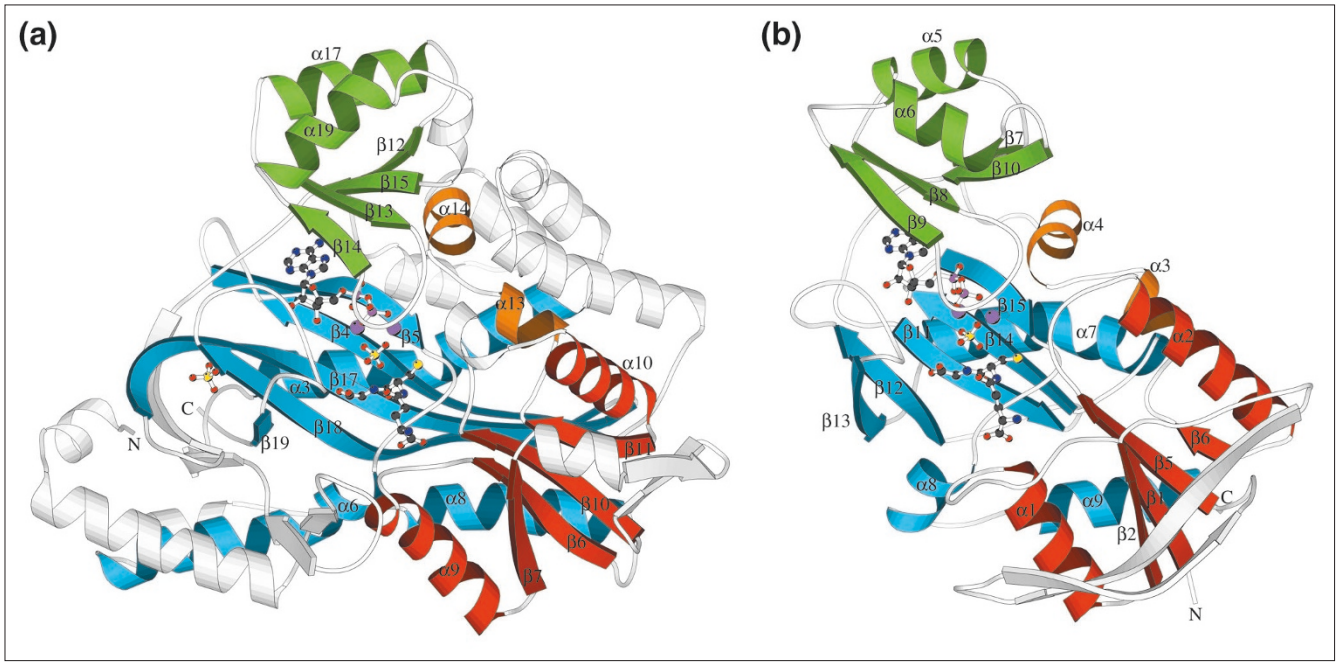

Figure 6
Comparison of GshB structures from a bacterium and a eukaryote. **(a)** Human; **(b)** *E. coli*. Reproduced with permission from [43].

gb|AAG55228.1|AE005266_7 (AE005266) ribosomal protein S6
Shotgun Score: 20

Query File	BLAST score	BLAST Prob.
bgshb_pa	89	0.44
bunk_sc	96	0.094
bgshb_ec	101	0.025
aS6mod_af	112	0.0012
bgshb_ac	112	0.0014
bgshb_cc	115	0.00067
bunk.bh	140	4.7e-07
bgsh_ss	141	7.0e-07
as6mod_ha	186	1.3e12
bS6mod_dr	192	2.6e-13
bgshb_an	192	4.9e-13
bgshb_no	192	4.9e-13
aunk_me	202	7.7e-15
bunk_me	215	7.0e-17
aS6mod_ap	227	1.3e-18
aS6mod_ss	237	1.1e-19
bS6mod_Mx	282	1.9e-24
bS6mod_st	355	3.5e-32
brimk_pm	620	2.9e-60

gb|AAG38537.1|AF309805_2 (AF309805) glutathione synthetase
Shotgun Score: 6

Query File	BLAST score	BLAST Prob.
egshb_lm	131	1.2e-16
egshb_ce	276	2.9e-23
egshb_hs	335	4.7e-30
egshb_at	354	8.0e-32
egshb_sp	362	2.1e38
egshb_pc	1220	7.7e-124

Figure 7
Sample Shotgun output. Bacterial, archaeal and eukaryotic sequences are shown in blue, red, and green, respectively. The query set consisted of the following sequences. bgshb_pa, GshB, *P. aeruginosa* (gi11348620); bunk_sc, unknown protein, *Streptomyces coelicolor* (gi 6752313); bgshb_ec, GshB, *E. coli* (gi121663); aS6mod_af, ribosomal S6 protein modification enzyme, *Archaeglobus fulgidus* (gi 11499884); bgshb_ac, GshB, *Anaplasma centrale* (gi544429); bgshb_cc, GshB, *Caulobacter crescentus* (gi13421252); Bunk_Bh, unknown protein, *Bacillus halodurans*, (gi10175219); bgsh_ss, GshB, *Synechocystis* (gi8134484); as6mod_ha, ribosomal S6 protein modification enzyme, *Halobacterium* sp. NRC-1 (gi10579747); bS6mod_dr, ribosomal S6 protein modification enzyme, *Deinococcus radiodurans* (strain R1) (gi7473852); bgshb_an, GshB, *Anabaena* sp. (gi1364079); bgshb_no, GshB, *Nostoc* sp. PCC 7120 (gi17231351); aunk_me, unknown protein, *Methanococcus jannaschii* Y620 (gi2496080); bunk_me, unknown protein, *Methylobacterium extorquens* (gi1193062); aS6mod_ap, S6 modification enzyme, *Aeropyrum pernix* (gi14601423); aS6mod_ss, ribosomal protein S6 modification protein, *Sulfolobus solfataricus* (gi6015830); bS6mod_Mx, ribosomal protein S6 modification protein, *Myxococcus xanthus* (gi2625142); bS6mod_st, ribosomal protein S6 modification protein, *Salmonella typhimurium* (gi16764237); brimk_pm, rimK protein, *Pasteurella multocida* (gi12721133); egshb_lm, GshB, *Leishmania major* (gi7940268); egshb_ce, unknown protein, *C. elegans*, (gi7506074); egshb_hs, GshB, *Homo sapiens* (gi4504169); egshb_at, GshB, *Arabidopsis thaliana* (gi5531229); egshb_sp, GshB, *Schizosaccharomyces pombe* (gi1708058); egshb_pc, GshB, *Pneumocystis carinii* (gi11596248).

as well as the extremely low sequence identities, make alignment difficult and preclude the identification of blocks in the permuted region. Therefore, a different motif-finding algorithm, MEME 3.0 [46,47], was used to analyze a divergent set of GshBs, along with several of the proteins related to the bacterial GshBs. Twelve strongly conserved motifs were identified (Table 4). The pattern of distribution of these motifs in eukaryotic GshBs, bacterial GshBs and S6 modification enzymes (the enzymes most closely related to the bacterial GshBs) is illustrated in Figure 8. Motifs 11 and 12 are peculiar to the bacterial GshB proteins and map to the substrate-binding region of the protein. These motifs are replaced by others in the S6 modification enzymes and in D-Ala-D-Ala ligases (data not shown), while the motifs involved in the ATP-binding site (1, 4, 6 and 5) are conserved. This protein scaffold clearly provides a modular structure that is easily adapted to bind different substrates in proximity to bound ATP. Two motifs (1 and 5) that contribute to the ATP-binding pocket were found in both eukaryotic and bacterial GshBs, as well as other members of the ATP-grasp superfamily. These motifs were present in different orders in the bacterial and eukaryotic GshBs (see Figure 8) because motif 5 is involved in the circular permutation that moves part of the carboxy-terminal domain to the amino-terminal domain of the eukaryotic enzymes. The lack of conserved motifs in the substrate-binding region is consistent with the comparison of the *E. coli* and human GshB structures, which shows remarkably little similarity in this region of the protein.

In a case such as this one, it can be difficult to determine whether common sequence motifs have arisen by divergence from a common progenitor, or by convergent evolution driven by a common function. The two common motifs found in the bacterial and eukaryotic GshBs correspond to the ATP-binding pocket. Thus, we examined the possibility that motifs 1 and 5 arose by convergent evolution driven by the need to bind ATP by looking for these motifs in other proteins that bind ATP. We searched the non-redundant database with motifs 1 and 5 using the MAST algorithm. Searches with motif 1 and 5 retrieved 145 and 40 sequences with *E*-values less than 10, respectively. Nearly all of the proteins with known functions in the output were GshBs or other members of the ATP-grasp superfamily. All of the 145 proteins found by motif 1 were members of the ATP-grasp superfamily except for two transcriptional regulators with *E*-values of 4.8 (gi15613287) and 7.8 (gi15224768), dihydroorotate dehydrogenases with *E*-values greater than 7 from several organisms, and a proline/betaine transporter (gi1589297) with an *E*-value of 9.9. All of the proteins with known functions found by motif 5 were members of the ATP-grasp superfamily except the LIM-containing protein kinase 2t (gi3273207), which had an *E*-value of 6.2. Thus, motifs 1 and 5 are characteristic of the ATP-binding region of the ATP-grasp superfamily enzymes.

The question of whether eukaryotic GshBs are members of the ATP-grasp superfamily is difficult to answer with certainty because the eukaryotic GshBs are so dramatically different from the other superfamily members. The two conserved sequence motifs involved in ATP binding do provide a link between eukaryotic GshBs and the ATP-grasp superfamily, but it is rather tenuous, as it is possible that these sequences provide the best way to bind ATP within the

Table 4

Motifs found in GshBs and related proteins

Motif	Found in	Best possible match	*E*-value
1	**All**	**HFPFVLKPQFGSWGNGVFK**	2.5e-107
2	Eukaryotic GshBs	WEARLLIEESHAIKCPSIAYHLAGSKKIQQVL	5.3e-49
3	S6 modification enzymes	NDPHAIERCCDKWWTKQLLAKHGIPVPDT	2.2e-106
4	ATP-grasp superfamily	RDWRVFVVGGEVVGA	8.5e-52
5	**All**	**GYWVIEVNTTP**	3.4e-44
6	ATP-grasp superfamily	GDWRTNCHQGGTAEPCSLTE	3.6e-34
7	Eukaryotic GshBs	NKQAGYLCRTKPKDTNEGGVAAGYAVLDSCYL	1.2e-17
8	S6 modification enzymes	EWLAVKAAKCMGLDYCGVDIL	2.6e-20
9	Eukaryotic GshBs	QEVAVVYFRSGYSPDHYPS	4.0e-16
10	Eukaryotic GshBs	TLFPSPFPHNVFEQACDVQMLFNELYDRISQDFEFLRDSLKSTVKYDDFT	8.0e-15
11	Bacterial GshBs	TLVVNNPQGLRDAPEKLYTQWFPKIIPPT	1.5e-10
12	Bacterial GshBs	FMRQDPPFDMQYIYATYILE	1.9e-8

Motifs found in the entire set of proteins are highlighted in bold (as shown in Figure 8)

context of this structural fold and have evolved by convergent evolution in eukaryotic GshBs and the ATP-grasp superfamily members. The lack of conservation in the glutathione-binding region of bacterial and eukaryotic GshBs is also an important consideration. We feel that the evidence for a true evolutionary relationship between the eukaryotic GshBs and the ATP-grasp superfamily is rather weak given the evidence available at this time. For our purposes here, it is sufficient to conclude, in agreement with Polekhina *et al.* [43], that eukaryotic GshBs did not evolve directly from bacterial GshBs, but rather that both evolved from ancestors that had the characteristic fold of the ATP-grasp superfamily.

A different twist? A fused GSHA-ATP-grasp superfamily homolog in an odd collection of bacteria

In most organisms, GshA and GshB are encoded by separate genes that are not in close proximity. An interesting variation on this theme may occur in a small number of bacteria. *Clostridium perfringens*, *Listeria monocytogenes*, *Listeria innocuans* and *Pasteurella multocidans* have an open reading frame (ORF) that could encode a GshA homolog fused to an ATP-grasp superfamily member, raising the possibility that this protein might combine the two activities required for synthesis of GSH. *C. perfringens* and *L. monocytogenes* contain low levels of GSH (0.25 μmol/g residual dry weight) that are about 20-fold lower than those found in *E. coli*, but it is not known whether they synthesize GSH or simply import it from the medium [4]. Experimental determination of the function of these fused proteins is clearly needed.

The amino-terminal part of the fusion proteins is most closely related to group 1 GshAs, although the sequence identities are not high (Table 5). The carboxy-terminal parts of these proteins are most closely related to cyanophycin synthetase, which catalyzes the synthesis of a polypeptide storage polymer in cyanobacteria, and more distantly related to D-Ala-D-Ala ligase, which is involved in peptidoglycan

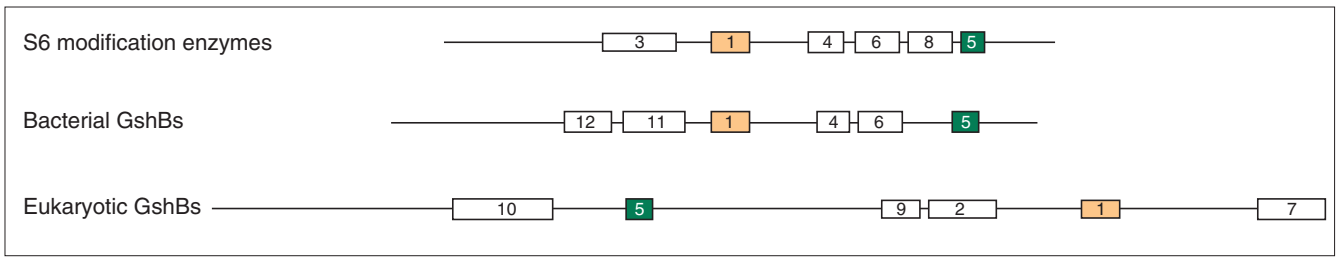

Figure 8
Motifs found by MEME 3.0 in a divergent set of eukaryotic GshBs, bacterial GshBs, and members of the ATP-grasp superfamily, mapped onto bacterial and eukaryotic GshBs and ribosomal protein S6 modification enzymes. (The colors of the motifs bear no relationship to the colors used to designate domains in Figure 6.)

Table 5

Percentage identities between fused GshA-ATP-grasp superfamily homologs and GshA, GshB and cyanophycin synthetase

Organism	Gene ID for fused GshA-ATP-grasp superfamily homolog	Percent identity between amino-terminal region and *E. coli* GshA	Percent identity between carboxy-terminal region and *E. coli* GshB	Percent identity between carboxy-terminal region and *Anabaena variabilis* cyanophycin synthetase
Clostridium perfringens	18310555	28	< 15	46
Listeria monocytogenes	16804807	26	< 15	44
Listeria innocua	16801972	26	< 15	44
Pasteurella multocida	15602913	24	< 15	43

biosynthesis in many bacteria. These enzymes are members of the ATP-grasp superfamily, to which GshB also belongs. However, there is no significant relationship between the carboxy-terminal region of these putative fusion proteins and known GshBs (see Table 5). If this part of the protein does indeed function as a glutathione synthetase, then this would be another example of independent recruitment of an ATP-grasp superfamily member to provide this function.

The occurrence of an ORF for the fusion protein in this cluster of bacteria is curious because *C. perfringens*, *L. monocytogenes* and *L. innocuans* are low-GC Gram-positive bacteria, while *P. multocida* is a gamma-proteobacterium. *C. perfringens* is found in soil and sewage and is often part of the normal intestinal flora of animals and humans. It causes gangrene and food poisoning in humans [48]. *L. monocytogenes* and *L. innocuans* are ubiquitous contaminants of soil and water, and *L. monocytogenes* causes listeriosis, a serious food-borne illness [48]. *P. multocida* colonizes the nasopharynx and gastrointestinal tract of many animals and birds, and causes a wide range of illnesses [49]. Human infections are most often caused by dog or cat bites. Thus, the association of these bacteria with animals as either commensal or pathogenic organisms has apparently provided an opportunity for lateral transfer of the gene encoding the fusion protein.

The ORF for the putative fusion protein in *P. multocida* is particularly intriguing because gamma-proteobacteria typically have a group 1 GshA and a typical bacterial GshB. *P. multocida* has neither of these, and neither does its close relative, *Haemophilus influenzae*. The lineage leading to *P. multocida* and *H. influenzae* diverged from other gamma-proteobacteria approximately 270 million years ago [50]. *P. multocida* and *H. influenzae* have considerably smaller genomes (2,014 and 1,743 predicted coding regions, respectively [50,51]) than *E. coli* (4,288 predicted coding regions [52]), suggesting that this lineage has undergone substantial genome reduction. It is possible that the lineage leading to *P. multocida* and *H. influenzae* lost GshA, and *P. multocida* subsequently acquired the fused GshA-ATP-grasp superfamily homolog in its place.

Putting together the pieces: thoughts on the evolution of the pathway

The pathway for GSH biosynthesis involves two enzymes, and it is of interest to consider which of these evolved first. Horowitz has postulated that biosynthetic pathways evolve in a retrograde fashion, beginning with the last enzyme in the pathway [53]. This hypothesis rests on the assumption that organisms had, at one time, access to a supply of precursors for biological polymers such as DNA, RNA, proteins and polysaccharides. As the supply of a given precursor dwindled, the most successful organisms would be those that 'invented' an enzyme with which to catalyze formation of that precursor from compounds present in the environment. Thus, there would be continuous selective pressure to add enzymes in the retrograde direction to catalyze synthesis of precursors from ever more simple constituents. Evolution of a biosynthetic pathway in the forward direction was deemed unlikely, as there would be no selective pressure for evolution of enzymes to produce intermediates of no further use to the organism. Horowitz's proposal is logical and appealing. There are cases, however, in which forward evolution of a pathway seems more likely. For example, many organisms make complex natural products whose roles generally involve killing or manipulating other organisms. The pathways for building these complex structures have probably evolved in a forward direction by addition of enzymes capable of adding to the complexity of a pre-existing molecule and thereby contributing to its biological potency.

The GSH biosynthesis pathway is most likely to have evolved in a forward direction. If the pathway had evolved in a retrograde direction, the Horowitz theory would postulate that GshB arose to take advantage of γ-Glu-Cys present in the environment. It is unlikely that γ-Glu-Cys would have been available because formation of the high-energy amide bond would be unlikely to occur abiotically. Furthermore, this molecule would be unstable to oxidation in aerobic environments. However, evolution of the GSH biosynthesis pathway in the forward direction makes considerable sense. γ-Glu-Cys can serve some of the functions of GSH, and therefore could be advantageous to an organism even in the absence of GshB. Indeed, halobacteria contain millimolar levels of

γ-Glu-Cys, but do not convert it further to GSH [2]. However, γ-Glu-Cys is not an ideal solution, as it is more easily oxidized than GSH [31]. Furthermore, the reactivity of a thiol depends upon its pKa, as thiolates are orders of magnitude more nucleophilic than thiols [54]. The nucleophilicity of the thiol in γ-Glu-Cys should be diminished by the proximity of the negatively charged carboxylate. Further reaction of γ-Glu-Cys with Gly to form GSH would improve its properties with respect to both oxidation and nucleophilicity, thus, providing selective pressure for evolution of a GSH synthetase (GshB).

One possible source for an enzyme to catalyze the next step in a pathway evolving in the forward direction is the enzyme that catalyzed the last step, as this enzyme has a binding site that accommodates the product of the last reaction, and that product is the substrate for the next reaction. A similar situation occurs for pathways evolving in a retrograde direction. This type of enzyme recruitment, which takes advantage of already existing substrate-specificity determinants, but requires changes in catalytic groups, appears to occur rather infrequently [55]. For example, among 510 proteins involved in the small-molecule metabolic pathways in *E. coli*, homology between consecutive enzymes in a pathway occurs only six times [56]. Most often, enzymes are recruited to catalyze new reactions by virtue of the catalytic abilities of their active sites, and interactions required for substrate binding are then optimized. The GSH biosynthesis genes, however, appear to be an optimal case for recruitment of one enzyme to catalyze a subsequent reaction. As GshA has a binding site that accommodates γ-Glu-Cys, it would appear to be an ideal progenitor of GshB, which uses γ-Glu-Cys as a substrate and also catalyzes the ATP-dependent formation of an amide bond. However, GshA and GshB appear to be structurally distinct. There are no experimental structures for GshAs, but recent work suggests that GshAs are homologs of glutamine synthetases [57]. GshBs have a different structural fold, characteristic of the ATP-grasp superfamily. Thus, the data support a scenario in which emergence of GshA was followed, in most organisms, by the recruitment of a different protein to serve as the progenitor of GshB. It is particularly interesting that, in both the bacterial and eukaryotic lineages, the ATP-grasp structural fold provided the starting point for the evolution of GshB.

Conclusions

Our analysis of the sequences of GshAs and GshBs suggests that the evolutionary history of these proteins is more complex than expected on the basis of the distribution of GSH in extant organisms. Our results, as well as the observation that GshA and GshB genes are generally not found in proximity in microbial genomes, suggest that these genes did not evolve together. Therefore, we must consider the evolutionary history of the two genes separately. Although the origin of the GshA gene cannot be unequivocally determined, it is most plausible to suppose that it arose in cyanobacteria, which would have been the first cells to require the protection conferred by γ-Glu-Cys against reactive oxygen species. If this hypothesis is correct, then subsequent lateral gene transfers must have occurred to spread the gene to the proteobacteria and eukaryotes, as well as to at least one archaeon and possibly to some Gram-positive bacteria. Because of the high level of sequence divergence, there is no clear indication in the sequence data as to whether eukaryotes acquired a GshA gene from a cyanobacterium or a proteobacterium. After the acquisition of GshA, a further improvement in protection against reactive oxygen species was obtained in most organisms by recruitment of an enzyme to convert γ-Glu-Cys to GSH. This recruitment apparently took place independently in the bacterial and eukaryotic lineages, since the sequence of the eukaryotic GshB is remarkably different from that of the bacterial GshBs, despite the structural similarities between these two proteins. At least for GshB, therefore, we can eliminate the possibility of transfer from the mitochondrial progenitor into an early eukaryote. The emerging picture of the evolution of the glutathione biosynthesis pathway is significant because it suggests that the pathway evolved in a forward direction, in contradiction to the Horowitz hypothesis.

Materials and methods

BLAST [16] and PSI-BLAST [16] searches were carried out at the NCBI website [15]. Multiple sequence alignment was performed using ClustalW [45] at the Pittsburgh Supercomputing Center. Pairwise sequence identities were determined using the Distances algorithm in the GCG package at the Pittsburgh Supercomputing Center. Motif analyses were carried out using MEME [46] at the San Diego Supercomputing Center [47] and Block Maker [19] at the Fred Hutchinson Cancer Center [20]. Phylogenetic analyses were carried out using PAUP 4.0b [21].

Acknowledgements

We thank Norman Pace, William Friedman and Scott Kelley for helpful discussions, and Patricia Babbitt for carrying out the Shotgun analysis. Financial support was provided by the NASA Astrobiology Program and a research computing grant from the Pittsburgh Supercomputing Center.

References

1. Carnegie PR: **Structure and properties of a homolog of glutathione.** *Biochem J* 1963, **89:**471-478.
2. Newton GL, Javor B: **gamma-Glutamylcysteine and thiosulfate are the major low-molecular-weight thiols in halobacteria.** *J Bacteriol* 1985, **161:**438-441.
3. Klapheck S, Chrost B, Starke J, Zimmermann H: **γ-Glutamylcysteinylserine - a new homolog of glutathione in plants of the family Poaceae.** *Bot Acta* 1992, **105:**174-179.
4. Newton GL, Arnold K, Price MS, Sherrill C, Delcardayre SB, Aharonowitz Y, Cohen G, Davies J, Fahey RC, Davis C:. **Distribution of thiols in microorganisms: mycothiol is a major thiol in most actinomycetes.** *J Bacteriol* 1996, **178:**1990-1995.
5. Holmgren A: **Thioredoxin and glutaredoxin systems.** *J Biol Chem* 1989, **264:**13963-13966.

6. Russel M, Model P, Holmgren A: **Thioredoxin or glutaredoxin in *Esherichia coli* is essential for sulfate reduction but not for deoxyribonucleotide synthesis.** *J Bacteriol* 1990, **172:**1923-1929.
7. Gladysheva TB, Oden KL, Rosen BP: **Properties of the arsenate reductase of plasmid R773.** *Biochemistry* 1994, **33:**7288-7293.
8. Fahey RC, Newton GL, Arrick B, Overdank-Bogart T, Aley SB: ***Entamoeba histolytica*: a eukaryote without glutathione metabolism.** *Science* 1984, **224:**70-72.
9. Brown DM, Upcroft JA, Upcroft P: **Cysteine is the major low-molecular weight thiol in *Giardia duodenalis*.** *Mol Biochem Parasitol* 1993, **61:**155-158.
10. Ellis JE, Yarlett N, Cole D, Humphreys MJ, Lloyd D: **Antioxidant defenses in the microaerophilic protozoan *Trichomonas vaginalis*: comparison of metronidazole-resistant and sensitive strains.** *Microbiology* 1994, **140:**2489-2494.
11. Fahey RC, Sundquist AR: **Evolution of glutathione metabolism.** In *Adv Enzymol Rel Areas Mol Biol* Edited by A. Meister. New York: John Wiley and Sons, 1991, 1-53.
12. Fahey RC: **Novel thiols of prokaryotes.** *Annu Rev Microbiol* 2001, **55:**333-356.
13. Yang D, Oyaizu H, Olsen GJ, Woese CR: **Mitochondrial origins.** *Proc Natl Acad Sci USA* 1985, **82:**4443-4447.
14. Gray MW: **Origin and evolution of organelle genomes.** *Curr Opin Genet Dev* 1993, **3:**884-890.
15. **NCBI protein database** [http://www.ncbi.nlm.nih.gov/]
16. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
17. Newton GL, Fahey RC, Cohen G, Aharonowitz Y: **Low-molecular-weight thiols in Streptomycetes and their potential role as antioxidants.** *J Bacteriol* 1993, **175:**2734-2742.
18. **Pfam database** [http://pfam.wustl.edu/]
19. Henikoff S, Henikoff JG, Alford WJ, Pietrokovski S: **Automated construction and graphical presentation of protein blocks from unaligned sequences.** *Gene* 1995, **163:**GC17-GC26.
20. **Blocks WWW Server** [http://blocks.fhcrc.org/blocks/]
21. Swofford DL: PAUP*. *Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Version 4. Sinauer Associates: Sunderland, MA, 2001.
22. DesMarais DJ: **When did photosynthesis emerge on earth?** *Science* 2000, **289:**1703-1705.
23. Doolittle WF: **You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes.** *Trends Genet* 1998, **14:**307-311.
24. Garcia-Vallvé S, Romeu A, Palau J: **Horizontal gene transfer in bacterial and archaeal complete genomes.** *Genome Res* 2000, **10:**1719-1725.
25. Jain R, Rivera MC, Lake JA: **Horizontal gene transfer among genomes: the complexity hypothesis.** *Proc Natl Acad Sci USA* 1999, **96:**3801-3806.
26. Ochman H, Lawrence JG, Groisman EA: **Lateral gene transfer and the nature of bacterial innovation.** *Nature* 2000, **405:**299-304.
27. Katz LA: **Transkingdom transfer of the phosphoglucose isomerase gene.** *J Mol Evol* 1996, **43:**453-459.
28. Wolf YI, Aravind L, Grishin NV, Koonin EV: **Evolution of aminoacyl-tRNA synthetases - analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events.** *Genome Res* 1999, **9:**689-710.
29. Stewart WN, Rothwell GW: *Paleobotany and the Evolution of Plants,* 2nd edn. Cambridge University Press: Cambridge, 1993.
30. Schell J, Koncz C: **The Ti-plasmid and plant molecular biology.** *Discoveries Plant Biol* 2000, **3:**393-409.
31. Sundquist AR, Fahey RC: **The function of γ-glutamylcysteine and bis-γ-glutamylcysteine reductase in *Halobacterium halobium*.** *J Biol Chem* 1989, **264:**719-725.
32. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278:**631-637.
33. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes.** *Nucleic Acids Res* 2001, **29:**22-28.
34. **Clusters of Orthologous Groups (COG) Database** [http://www.ncbi.nlm.nih.gov/COG/]
35. Palmer DR, Garrett JB, Sharma V, Meganathan R, Babbitt PC, Gerlt JA: **Unexpected divergence of enzyme function and sequence: "*N*-acylamino acid racemase" is *o*-succinylbenzoate synthase.** *Biochemistry* 1999, **38:**4252-4258.
36. Galperin MY, Koonin EV: **A diverse superfamily of enzymes with ATP-dependent carboxylate-amine/thiol ligase activity.** *Protein Sci* 1997, **6:**2639-2643.
37. Harrop HA, Held KD, Michael BD: **The oxygen effect: variation of the K-value and lifetimes of oxygen-dependent damage in some glutathione-deficient mutants of *Escherichia coli*.** *Int J Radiat Biol* 1991, **59:**1237-1251.
38. Stephen DWS, Jamieson DJ: **Glutathione is an important antioxidant molecule in the yeast *Saccharomyces cerevisiae*.** *FEMS Microbiol Lett* 1996, **141:**207-212.
39. Cobbett CS, May MJ, Howden R, Rolls B: **The glutathione-deficient, cadmium-sensitive mutant, *cad-1*, of *Arabidopsis thaliana* is deficient in γ-glutamylcysteine synthetase.** *Plant J* 1998, **16:**73-78.
40. Dalton PD, Dieter MZ, Yang Y, Shertzer HG, Nebert DW: **Knockout of the mouse glutamate cysteine ligase catalytic subunit (*Gclc*) gene: embryonic lethal when homozygous, and proposed model for moderate glutathione deficiency when heterozygous.** *Biochem Biophys Res Commun* 2000, **279:**324-329.
41. Galperin MY, Koonin EV: **A diverse superfamily of enzymes with ATP-dependent carboxylate-amine/thiol ligase activity.** *Protein Sci* 1997, **6:**2639-2643.
42. Hara T, Kato H, Katsube Y, Oda J: **A pseudo-Michaelis quaternary complex in the reverse reaction of a ligase: structure of *Escherichia coli* B glutathione synthetase complexed with ADP, glutathione, and sulfate at 2.0 Å resolution.** *Biochemistry* 1996, **35:**11967-11974.
43. Polekhina G, Board PG, Gali Rr, Rossjohn J, Parker MW: **Molecular basis of glutathione synthetase deficiency and a rare gene permutation event.** *EMBO J* 1999, **18:**3204-3213.
44. Pegg SC-H, Babbitt PC: **Shotgun: getting more from sequence similarity searches.** *Bioinformatics* 1999, **15:**729-740.
45. Thompson JD, Higgins DG, Gibson TJ: **ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.
46. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** In *Proc Int Conf Intell Syst Mol Biol* 1994, **2:**28-36.
47. **MEME 3.0** [http://meme.sdsc.edu/meme/website/]
48. Madigan MT, Martinko JM, Parker J: *Brock Biology of Microorganisms* 9th edn. Upper Saddle River, NJ: Prentice-Hall, 2000.
49. Klein NC, Cunha BA: ***Pasteurella multocida* pneumonia.** *Semin Resp Infect* 1997, **12:**54-56.
50. May BJ, Zhang Q, Li LL, Paustian ML, Whittam TS, Kapur V: **Complete genome sequence of *Pasteurella multocida*, Pm70.** *Proc Natl Acad Sci USA* 2001, **98:**3460-3465.
51. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, *et al.*: **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.** *Science* 1995, **269:**496-511.
52. Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, *et al.*: **The complete genome sequence of *Escherichia coli* K-12.** *Science* 1997, **277:**1453-1462.
53. Horowitz NH: **On the evolution of biochemical syntheses.** *Proc Natl Acad Sci USA* 1945, **31:**153-157.
54. Roberts DD, Lewis SD, Ballou DP, Olson ST, Shafer JA: **Reactivity of small thiolate anions and cysteine-25 in papain toward methyl methanethiosulfonate.** *Biochemistry* 1986, **25:**5595-5601.
55. Gerlt JA, Babbitt PC: **Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies.** *Annu Rev Biochem* 2001, **70:**209-246.
56. Teichmann SA, Rison SCG, Thornton JM, Riley M, Gough J, Chothia C: **The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*.** *J Mol Biol* 2001, **311:**693-708.
57. Abbott JJ, Pei J, Ford JL, Qi Y, Grishin YN, Pitcher LA, Phillips MA, Grishin NV: **Structure prediction and active site analysis of the metal binding determinants in γ-glutamylcysteine synthetase.** *J Biol Chem* 2001, **276:**42099-42107.
58. Sherrill C, Fahey RC: **Import and metabolism of glutathione by *Streptococcus mutans*.** *J Bacteriol* 1998, **180:**1454-1459.
59. Fahey RC, Buschbacher RM, Newton GL: **The evolution of glutathione metabolism in phototrophic microorganisms.** *J Mol Evol* 1987, **25:**81-88.
60. Okumura N, Masamoto K, Wada H: **The gshB gene in the cyanobacterium *Synechococcus* sp. PCC 7942 encodes a functional glutathione synthetase.** *Microbiology* 1997, **143:**2883-2890.