Minireview

Making sense of nonsense: the evolution of selenocysteine usage in proteins

Paul R Copeland

Address: Department of Molecular Genetics, Microbiology and Immunology, UMDNJ-Robert Wood Johnson Medical School, 675 Hoes Lane, Piscataway, NJ 08854, USA. E-mail: paul.copeland@umdnj.edu

Published: 27 May 2005

Genome ***Biology*** 2005, **6:**221 (doi:10.1186/gb-2005-6-6-221)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/6/221

Abstract

A recent analysis of sequences derived from organisms in the Sargasso Sea has revealed a surprisingly different set of selenium-containing proteins than that previously found in sequenced genomes and suggests that selenocysteine utilization has been lost by many groups of organisms during evolution.

As well as the 20 amino acids universally found in proteins, two other amino acids - pyrrolysine and selenocysteine - are incorporated into a small number of proteins in some groups of organisms. L-pyrrolysine is a C4-substituted pyrroline-5-carboxylate attached to the ϵ-nitrogen of lysine; L-selenocysteine is identical to cysteine but with selenium substituted for sulfur. Pyrrolysine has so far been found only in enzymes required for methanogenesis in some archaebacteria, suggesting a possible role in catalysis, but the precise role of this amino acid has not been identified. The selenium atom in selenocysteine confers a much higher reactivity than cysteine, as its lower pKa (5.2) allows it to remain ionized at physiological pH. Most selenoproteins use their higher nucleophilic activity to catalyze redox reactions, but many have no known function. The current studies of selenoprotein evolution represent one of the important tools used to completely identify and categorize selenoprotein function.

The Sargasso Sea (named for the surface-borne sargassum seaweed) is a body of water covering 2 million square miles in the middle of the North Atlantic Ocean near Bermuda. Its well defined physical and geochemical properties, including relatively low nutrient levels, made it an alluring target for a shotgun sequencing project covering a whole biome - a collection of interrelated ecosystems typical of a particular physical environment [1]. This effort, the first 'biome sequencing project', represents a novel application for shotgun genome sequencing and is an important new component of modern bioinformatics. Of the 1.2 million genes identified by this approach, however, a small subset is likely to be misannotated because of the presence of in-frame nonsense codons, either UGA or UAG, which in these cases are acting as codons for selenocysteine and pyrrolysine, respectively. In some archaea, the UAG codon is redefined as a pyrrolysine codon, apparently forcing these organisms to rely on only two redundant signals (UGA and UAA) for translation termination [2]. In many bacteria, some methanogenic archaea and most, if not all, animals, the codon UGA can be used to specify the incorporation of selenocysteine as well as for translation termination. As well as UGA, selenocysteine incorporation requires an additional *cis*-element in the gene and *trans*-acting factors.

Although selenocysteine incorporation is much more widely distributed than that of pyrrolysine, it is still an evolutionary mosaic. In fact, two kingdoms of life - plants and fungi - have eschewed the system entirely - or perhaps never acquired it (Table 1). So why does selenocysteine incorporation persist in some groups of organisms and not others? What are the forces driving the evolution of selenoproteins? In which direction is the evolution going - are animals in the process of phasing out or phasing in selenocysteine utilization? There are no answers to these questions yet, but a recent analysis of the large Sargasso Sea sequence dataset by Vadim Gladyshev and colleagues [3] at the University of Nebraska, Lincoln, is a first step toward shaping our view of selenoprotein evolution.

Table 1

Mosaic of selenoprotein evolution

Domains of life	Phyla	Selenogenomes	Total genomes
Eubacteria	Actinobacteria	2	18
	Aquificae	1	1
	Bacteroidetes/Chlorobi	0	5
	Chlamydiae/Verrucomicrobia	0	9
	Chloroflexi	1	1
	Chrysiogenetes		-
	Cyanobacteria	0	9
	Deferribacteres		
	Deinococcus-Thermus	0	3
	Dictyoglomi		-
	Fibrobacteres/Acidobacteria		-
	Firmicutes	9	58
	Fusobacteria	0	1
	Gemmatimonadetes		-
	Nitrospirae		-
	Planctomycetes	0	1
	Proteobacteria	29	95
	Spirochaetes	1	6
	Thermodesulfobacteria		-
	Thermotogae	0	1
Archaea	Crenarchaeota	0	4
	Euryarchaeota	4	16
	(Methanogens)	(4)	(6)
	Korarchaeota		-
Eukarya	Protists	1	1
	Fungi	0	3
	Plantae	0	8
	Animalia	7	7

Selenoproteins are found in a variety of phyla within all three lines of descent of life. The number of genomes encoding selenoproteins is indicated ('selenogenomes') together with the total number of sequenced genomes in the phylum. Numbers are based on data obtained in [8] except that any completed genomes entered into GenBank since 31 December 2003 were added to the total genome number and those possessing both *selB* and *selD* homologs were added to the number of selenoprotein-encoding genomes.

Cleaning the database

The misannotation of selenoproteins has been carefully and systematically corrected in completed genomes by Gladyshev's group. Work in this arena began just before the 'genomic era' when two groups published algorithms designed to identify eukaryotic selenoprotein genes by locating selenocysteine insertion sequence (SECIS) elements downstream of in-frame UGA codons [4,5]. SECIS elements specify a stem-loop mRNA structure that is required for selenocysteine (Sec) incorporation. Two *trans*-acting entities are also required: a specialized translation elongation factor for Sec-tRNA$^{[Ser]Sec}$ binding and delivery to the ribosome as well as a SECIS-element-binding component. In bacteria, the SECIS element is located just downstream of the Sec codon and the SECIS-binding component is a domain within the elongation factor. In eukaryotes, the SECIS element is in the 3' untranslated region of the gene and the SECIS-binding protein is encoded by a separate gene (SBP2, reviewed in [6]). Archaea appear to possess a mixture of the two systems, with SECIS elements located in untranslated regions but SECIS binding being a function of the elongation factor.

Gladyshev's group subsequently applied their algorithmic wares to the human genome to catalog a complete 'selenoproteome' consisting of 25 human genes encoding selenoproteins [7]. A similar task proved more challenging for prokaryotes because the SECIS element is not well conserved in bacteria. To tackle the prokaryotic genomes, Kryukov and Gladyshev [8] took a slightly different approach, using the assumption that all selenoproteins have orthologs in other species that have a conserved cysteine residue in place of selenocysteine. While this may seem a risky assumption, the risk is tempered by the fact that their study found that only 20% of eubacteria with completely sequenced genomes utilize selenoproteins. This suggests that a complete comparison of gene sets should yield plenty of cysteine homologs, assuming these genes to represent relatively stable gene families. In addition, the ability of an organism to utilize selenocysteine can be determined quite easily, and independent of selenoprotein analysis, because at least four genes are required for incorporation in bacteria: *selA* (selenocysteine synthase), *selB* (Sec-specific translation elongation factor), *selC* (encoding tRNA^{Sec}) and *selD* (selenophosphate synthase).

Each of the 'idiosyncracies' of the selenocysteine system was exploited in rank order and an algorithm was designed for identifying selenoprotein genes [8]. The algorithm looks something like this: first, identify bacteria containing at least one component of the selenocysteine incorporation machinery; second, identify pairs of homologous genes with cysteine codon-TGA pairs and align the regions flanking the TGA; third, make sure that the TGA positions correspond to conserved cysteine residues in cluster groups; and fourth, analyze genes individually for potential SECIS elements and for homology with known selenoproteins. Using this algorithm, ten known selenoprotein families were identified, as well as five new families (those with definitive eukaryotic selenoprotein homologs), eight strong candidates (new cysteine-selenocysteine pairs appearing at least twice in the dataset) and one weak candidate that appeared as a singleton. One class of selenoproteins that this algorithm cannot detect is that in which no cysteine-containing homolog exists. As noted above, this would seem very unlikely, but one such gene is known to exist: that for glycine reductase selenoprotein A. This is an apparently unique case, as the

recently developed bacterial SECISearch program confirmed the fact that all known bacterial selenoproteins except glycine reductase selenoprotein A have cysteine-containing homologs [9].

From an evolutionary perspective, things seemed fairly tidy on the basis of the analysis of completed and partially completed prokaryotic genomes: there was minimal overlap in the eukaryotic and prokaryotic selenoproteomes, and the prokaryotic selenoproteome was dominated by a single gene family, formate dehydrogenase α-chain (*fdhA*). The authors [8] argued that there is evidence of 'recent' cysteine-to-selenocysteine evolution for genes that are rare as well as the 'ancient' preservation of major gene families such as *fdhA*.

This comfortable scenario for prokaryotic selenoprotein evolution lasted precisely a year. The Sargasso Sea database analysis [3] now provides two new pieces of information that shatter previous assumptions: three selenoprotein families that were thought to be of eukaryotic origin are found among the bacteria in the Sargasso Sea (deiodinase, glutathione peroxidase and SelW), and *fdhA* was found to be a minority selenoprotein gene in this dataset (around 3% of the selenoprotein genes). In the Sargasso Sea data, a total of 310 known and new selenoproteins (clustered from a total of 2,131 unique TGA-containing open reading frames) were identified from the pool of 811,372 sequences with 88% of the selenoprotein genes falling into one of three families - SelW-like, peroxiredoxin or proline reductase. The remaining 12% of genes were spread over 22 families.

Because the Sargasso Sea database is reported to represent at least 1,800 species with variable coverage, it is difficult to assess what percentage of the species possess selenoproteins. But searches in this database for highly conserved genes defined anywhere from 341 to 569 species [1], suggesting that the most common selenoprotein gene (*selW* with 48 unique sequences), if universally conserved among marine bacteria utilizing selenocysteine, would correspond to the presence of selenoproteins in approximately 8-14% of bacterial species. Despite the vast number of assumptions made in arriving at those percentages, they are not too far from the 20% of species found to utilize selenocysteine among those with at least partially sequenced genomes. Yet the Sargasso Sea yielded entirely different sets of selenoproteins from the fully sequenced genomes. Of the multitude of possible explanations for this phenomenon, two stand out. First, as Zhang *et al*. [3] suggest, the relatively constant supply of selenium in seawater would mean less need for flexibility in the use of selenoproteins than is experienced by terrestrial organisms that must deal with dramatic differences in local selenium concentrations depending on location. Alternatively, it is tempting to speculate that laboratory culture conditions have selected for a subset of bacteria that require selenoFdhA, thus dramatically increasing the representation of that gene among the well-studied bacteria. As most microbes cannot be cultured in the laboratory, the Sargasso Sea dataset may simply more accurately reflect the gene distributions in nature, thus bearing out the main advantage of biome sequencing.

The forces driving the evolution of selenocysteine utilization

The discovery of new prokaryotic selenoprotein families in the Sargasso Sea data revealed phylogenetic information clearly demonstrating independent evolution of all three gene families common to both prokaryotes and eukaryotes (glutathione peroxidase, deiodinase and SelW). In addition, the hallmarks of the selenocysteine utilization system also show evidence of a common ancestor. That is, all three systems share three major features: selenocysteine is always encoded by UGA, incorporation always requires a stem-loop specificity sequence (SECIS element), and there is always a dedicated translation elongation factor plus an RNA-binding component. Nevertheless, the present distribution of selenocysteine utilization among the major phyla clearly illustrates an evolutionary mosaic for selenoproteins (Table 1). If the assumption is made that all life began with the opportunity to utilize selenocysteine, then one is forced to conclude that some groups lost their incorporation machinery, most probably as a result of limiting selenium. The persistence of selenocysteine utilization makes it clear that maintaining the system provides selective advantage, but that the advantage quickly becomes a serious (or perhaps fatal) disadvantage if selenium supply is inadequate.

Interestingly, if the system had usurped a cysteine codon instead of a stop codon the situation might have turned out differently, allowing an organism to switch between cysteine- and selenocysteine-containing enzymes when selenium supply allowed. The fact that the system did not evolve this way may suggest that there is something more to the loss of selenocysteine than a simple conversion to cysteine-containing enzymes. Because selenoenzymes substituted with cysteine are generally considered significantly less active, it seems quite likely that cysteine-containing redox enzymes must have adapted to the loss of selenium by co-evolving active-site contexts that improve the efficiency of cysteine's redox power. One can therefore imagine that a biome-sequencing project comparing selenium-rich and selenium-poor environments would yield significant insight into the forces behind selenoprotein evolution.

Another argument against organisms acquiring selenocysteine utilization *de novo* is the fact that in *Escherichia coli*, for example, only two of the four genes for the selenocysteine incorporation machinery are physically linked in an operon [10]. If organisms had acquired the system from lateral gene transfer, then one might expect to see a much closer physical relationship among the genes. In addition, there have been no reports that these genes have ever been

found on a plasmid or phage. Interestingly, a search of the GenBank plasmid database does yield one plasmid hit in *Sinorhizobium* (*Rhizobium*) *meliloti*, the nitrogen-fixing plant symbiont [11]. This 1.35-Mbp plasmid, called pSymA, is actually genomic in scale, but it is interesting to note that all four selenocysteine incorporation genes are located within an approximately 20 kb region with a transposon between the *selA*/*B* and *selC*/*D* (Figure 1). Perhaps a vector for selenoprotein acquisition does exist - only time and a lot more sequencing and gene mapping will tell whether a subset of organisms can be classified as having obtained the selenocysteine-utilization system from a pSymA-like arrangement.

Molecular archeology

If the majority of microbes lack the selenocysteine utilization system, then those that might have 'recently' lost access to selenium could still contain relics of the system. In addition, as the genes are not all linked, it seems likely that gene loss would proceed at variable rates, leaving an imbalance in the components of the selenocysteine system. Indeed, using the *Salmonella enterica* sequences for the four components in a search of the nonredundant GenBank bacterial sequence database, with a stringent significance cutoff (10^{-14}) to eliminate annotation errors, yields 65 hits for *selA*, 31 hits for *selB*, 31 hits for *selC* and 99 hits for *selD*. While this is a crude method, it clearly suggests that *selD* and perhaps *selA* persist in organisms that lack selenoproteins, thus increasing the likelihood that they are remnants of the selenocysteine utilization system that have probably been retained for use in other processes. This latter point may be borne out by the fact that *selD* shows some sequence similarity to thymidine monophosphate nucleotide kinase and, perhaps not surprisingly, *selA* is similar to selenocysteine β-lyase, the enzyme that catalyzes the back-reaction of selenocysteine synthesis.

Perhaps the most interesting evolutionary question for selenoprotein biology is why archaea and animals evolved an incorporation system different from that of bacteria, in that it uses a distal SECIS element and, in the case of animals, a separate SECIS-binding component. Perhaps it is a question of efficiency. Selenocysteine incorporation is routinely reported as being inefficient (around 10% at best) in both bacteria and mammalian cells [12,13]. Unfortunately, efficiency has never been measured for an endogenous selenoprotein, probably because it is a daunting task on account of the differential stabilities of full-length selenoproteins and truncated versions (the result of termination instead of selenocysteine incorporation). It is known, however, that at least one mammalian selenoprotein (glutathione peroxidase 4) is expressed in very large quantities in the testis, and it seems unlikely that this overexpression would come from an inherently inefficient system. In addition, because the bacterial system is extremely well defined, it is likely that the low efficiency values reported are accurate.

Thus, one might argue that the main difference between bacterial and eukaryotic selenocysteine incorporation is efficiency. But if primordial selenocysteine utilization was inefficient, then it seems surprising that 'efficiency elements' were not simply laid on top of the already functioning bacterium-like system. New evidence suggests that this may indeed be the case. Recent work from John Atkins' laboratory at the University of Utah [14] has identified in-frame stem-loop structures in several mammalian selenoprotein genes that can account for a significant portion of total selenocysteine incorporation activity. In fact, they are able to support selenocysteine incorporation in the absence of a SECIS element. This similarity to bacterial SECIS elements is too attractive to ignore and begs the question of whether there are primordial eukaryotic SECIS elements in bacterial mRNAs. One current hypothesis is that the mammalian system has strong links to ribosome structure and function

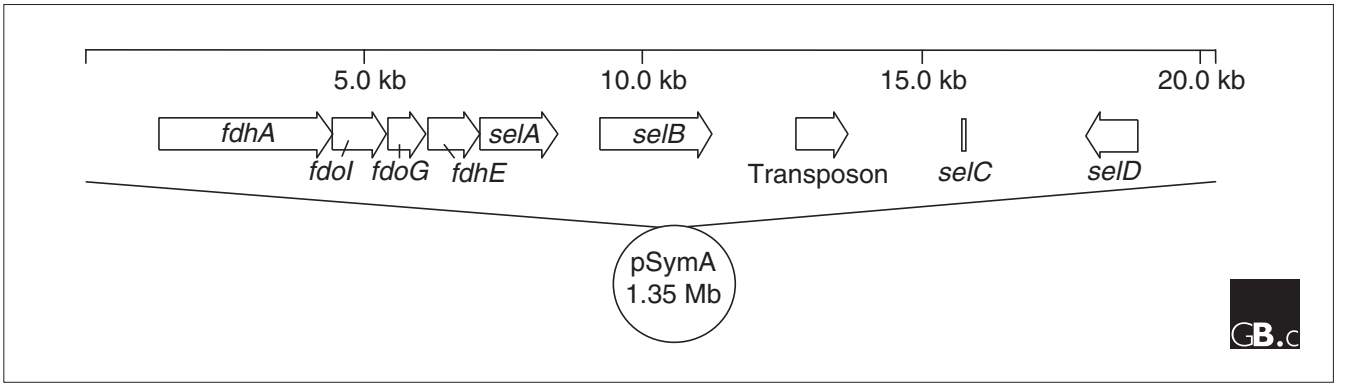

Figure 1
Diagram of the pSymA megaplasmid in *Sinorhizobium (Rhizobium) meliloti*, illustrating the physical relationships among genes of the selenocysteine utilization system (*selA*, *selB*, *selC* and *selD*) and the only known selenoprotein gene in this organism, the α-subunit of formate dehydrogenase (*fdhA*). Also noted is the location of a putative transposon between *selA*/B and *selC*/D [11].

[6], but only further forays into the world of biome sequence analysis will uncover the 'missing links' in prokaryotic selenoprotein evolution that got us to the current state of the art in mammalian cells.

References

1. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, *et al.*: **Environmental genome shotgun sequencing of the Sargasso Sea.** *Science* 2004, **304:**66-74.
2. Zhang Y, Baranov PV, Atkins JF, Gladyshev VN: **Pyrrolysine and selenocysteine use dissimilar decoding strategies.** *J Biol Chem* 2005, **288:**20740-20751.
3. Zhang Y, Fomenko DE, Gladyshev VN: **The microbial selenoproteome of the Sargasso Sea.** *Genome Biol* 2005, **6:**R37.
4. Lescure A, Gautheret D, Carbon P, Krol A: **Novel selenoproteins identified *in silico* and *in vivo* by using a conserved RNA structural motif.** *J Biol Chem* 1999, **274:**38147-38154.
5. Kryukov GV, Kryukov VM, Gladyshev VN: **New mammalian selenocysteine-containing proteins identified with an algorithm that searches for selenocysteine insertion sequence elements.** *J Biol Chem* 1999, **274:**33888-33897.
6. Driscoll DM, Copeland PR: **Mechanism and regulation of selenoprotein synthesis.** *Annu Rev Nutr* 2003, **23:**17-40.
7. Kryukov GV, Castellano S, Novoselov SV, Lobanov AV, Zehtab O, Guigo R, Gladyshev VN: **Characterization of mammalian selenoproteomes.** *Science* 2003, **300:**1439-1443.
8. Kryukov GV, Gladyshev VN: **The prokaryotic selenoproteome.** *EMBO Rep* 2004, **5:**538-543.
9. Zhang Y, Gladyshev VN: **An algorithm for identification of bacterial selenocysteine insertion sequence elements and selenoprotein genes.** *Bioinformatics* 2005, **21:**2580-2589.
10. Sawers G, Heider J, Zehelein E, Bock A: **Expression and operon structure of the *sel* genes of *Escherichia coli* and identification of a third selenium-containing formate dehydrogenase isoenzyme.** *J Bacteriol* 1991, **173:**4983-4993.
11. Barnett MJ, Fisher RF, Jones T, Komp C, Abola AP, Barloy-Hubler F, Bowser L, Capela D, Galibert F, Gouzy J, *et al.*: **Nucleotide sequence and predicted functions of the entire *Sinorhizobium meliloti* pSymA megaplasmid.** *Proc Natl Acad Sci USA* 2001, **98:**9883-9888.
12. Suppmann S, Persson BC, Bock A: **Dynamics and efficiency *in vivo* of UGA-directed selenocysteine insertion at the ribosome.** *EMBO J* 1999, **18:**2284-2293.
13. Mehta A, Rebsch CM, Kinzy SA, Fletcher JE, Copeland PR: **Efficiency of mammalian selenocysteine incorporation.** *J Biol Chem* 2004, **279:**37852-37859.
14. Howard MT, Aggarwal G, Anderson CB, Khatri S, Flanigan KM, Atkins JF: **Recoding elements located adjacent to a subset of eukaryal selenocysteine-specifying UGA codons.** *EMBO J* 2005, **24:**1596-1607.