Biochimica et Biophysica Acta 1790 (2009) 1463–1470

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbagen

Review

On the unique function of selenocysteine — Insights from the evolution of selenoproteins

Sergi Castellano *

Janelia Farm Research Campus, Howard Hughes Medical Institute, 19700 Helix Drive, Ashburn, VA 20147, USA

ARTICLE INFO

Article history:
Received 25 January 2009
Received in revised form 20 March 2009
Accepted 24 March 2009
Available online 1 April 2009

Keywords:
Selenium
Selenocysteine
Cysteine
Selenoprotein
Amino acid exchangeability
Natural selection

ABSTRACT

The process of natural selection leaves signatures in our genome that can be used to identify functionally important amino acid changes in proteins. In natural populations, amino acids that are better adapted to local conditions might increase in frequency, whereas moderately to severely deleterious protein mutations tend to be eliminated and do not contribute to protein differences between species. Amino acid mutations with no fitness consequences are, however, lost or fixed without regard to natural selection. Looking for evidence of natural selection is, therefore, an attractive strategy for characterizing the contribution of a residue to protein function. Because the majority of identified selenoproteins have now been found in Cys-form, the extent of exchangeability between Sec and Cys residues can be measured in proteins over long periods of time. The statistical analysis of the pattern of Sec/Cys exchanges in diversity (within species) and divergence (between species) data, provides robust inferences of the strength and mode of natural selection acting on these protein sites. Such inferences inform us not only of the long-term exchangeability between Sec and Cys residues, but also of the nature of the selective factors shaping Sec usage in proteins.

1. Selenium biology in the genomics era

As more and more genomes are sequenced, more and more biologists have turned to study the evolution of their proteins of interest. These evolutionary studies stem from the realization that patterns of amino acid conservation and substitution in proteins may provide important functional information. Through population genetics or comparative genomics, researchers identify the changes that evolution has sculpted into proteins over thousands or millions of years. Selenium biologists are no exemption in our desire to explain observed changes in proteins through evolutionary time. The more so, since the defining feature of selenoproteins, the selenocysteine residue (Sec), is known to be replaced by cysteine (Cys) in most selenoprotein families. The extent of functional exchangeability between these two residues is a long-standing question in selenium biology [1]. This question is implicitly an evolutionary question. The long-term exchangeability between Sec and Cys amino acids cannot be fully ascertained from functional studies on extant selenoproteins, as functional differences in present-day sequences are not a measure of fitness in natural populations. The extent of exchangeability between Sec and Cys residues, however, depends on the strength and mode of natural selection acting on these protein sites. Inferences of natural selection over evolutionary time, in turn, rely on population or comparative genomics data. Such data are abundant today and hold the answer to many evolutionary questions in selenium biology.

It is important to acknowledge, however, the fundamental distinction between the outcome of evolution and the events that lead to such changes. Because different evolutionary forces can result in the same observed substitutions in proteins, the documentation of protein differences among species is not proof of the role of natural selection in shaping Sec usage. There is a need of formally testing any inference about the role of natural selection on Sec sites before an adaptive hypothesis can be made. That is, to quantitatively assess natural selection with a robust evolutionary test. Without a statistical assessment it is not possible to distinguish an evolutionary hypothesis with merit from evolutionary story telling.

The purpose of this article is to review the current knowledge regarding the evolution of Sec/Cys usage in proteins in light of modern evolutionary theory. The origin and incorporation of Sec into the genetic code is beyond the scope of this review, and only the exchangeability between the two amino acids is discussed. Towards this end, it is necessary first to explicitly distinguish the description of selenoproteins and selenoproteomes among species, a very successful area of research in the last few decades [2–15], from the evolutionary analysis of these data. The application and interpretation of basic evolutionary theory to selenium studies is then discussed at length. The difficulty to infer natural selection over deep evolutionary time from functional studies, an important but sometimes unrecognized problem in these types of studies, is also discussed. The first inferences of natural selection on selenoproteins from diversity (within species) and divergence (between species) data are examined. Finally, functional interpretations of the long-term exchangeability between Sec and Cys residues in comparative analyses are reviewed.

* Tel.: +1 571 209 4000x 7160; fax: +1 571 209 4095.
E-mail address: castellanos@janelia.hhmi.org.

0304-4165/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.bbagen.2009.03.027

2. Descriptive selenium genomics

Sound evolutionary analysis must always rely on a thorough genomic description. Despite the difficulty to identify selenoproteins, the number of selenoprotein families has expanded considerably in the last four decades. First, through experimental means [2–3]. For example, the first experimentally identified protein to incorporate selenocysteine was glycine reductase in 1976 [16]. More recently, though, selenoproteins have been identified through computational and comparative genomics approaches [4–15]. To date, dozens of selenoprotein families have been identified and the selenoproteome (set of selenoproteins in a proteome) for some species is now believed to be complete [9,11,17]. Detailed accounts of prokaryotic and eukaryotic selenoproteomes can be found in other contributions in this issue.

Evolutionary studies on Sec/Cys replacements are sensitive to mispredictions and should only include verified selenoproteins. The history of selenoprotein identification is, however, no strange to controversy. The *Drosophila oaf* and *kelch* genes were once believed to encode selenoproteins [18,19], but further experimental and computational studies have not supported selenocysteine incorporation into these proteins. More recently, several new selenoproteins have been predicted *in silico* in the *Anopheles gambiae* genome [20,21]. Inspection of sequence conservation beyond the putative Sec codon in homologous alignments (Castellano, unpublished) do not, however, support these predictions. A more consequential result is the computational prediction of hundreds of selenoproteins in disrupted mRNAs in mouse [22]. This figure is an order of magnitude larger than the current estimate for the mouse selenoproteome [9], and would dramatically shift our views of the use and importance of selenium in mammals. The deceiving nature of selenoprotein identification, however, suggests caution, and it is unlikely that a large fraction of these predictions turn out to be real selenoproteins [23]. A common element to these contentious predictions is the lack of experimental validation. Because computational approaches to the prediction of selenoproteins suffer badly from false positives, the experimental test of selenoprotein candidates must be the gold standard in the field.

The identification of a novel selenoprotein is usually followed by the interrogation of sequence databases for proteins with shared ancestry. In this way, the pattern of Sec/Cys exchanges is described. Inferring accurately the homologous relationships between Sec/Cys sites is central to any evolutionary study on selenoproteins. Homology search algorithms, notably the NCBI-BLAST program [24], have been used extensively for the purpose of describing the distribution of selenoprotein families. Karlin/Altschul statistics [25], which assesses the statistical significance of homology searches, ensures the validity (as valid as *E*-values go) of the homologous relationships within selenoprotein families. Note, however, that the use of Sec in different selenoprotein families is not necessarily homologous. Sec usage is believed to have arisen independently (polyphyletic origin) in many selenoprotein families.

A bigger limit to progress in the study of selenoprotein evolution is poor annotation rather than lack of sequence. Homology searches usually result in truncated proteins, and selenoprotein gene structures need to be annotated in genomic sequences. Genome annotation compared with sequence generation is more difficult, time-consuming and costly. This is particularly true for selenoproteins. The dual role of the UGA codon confounds gene prediction programs and human curators alike, and results in the misannotation of selenoprotein genes in genome projects and databases. Some ongoing efforts to systematically annotate selenoprotein gene structures are a step in the right direction [26]. The rapid pace of genome sequencing, however, all but ensures a growing gap between selenoprotein identification and the correct annotation of selenoprotein gene structures. The new 454 and Solexa sequencing technologies will only exacerbate this problem.

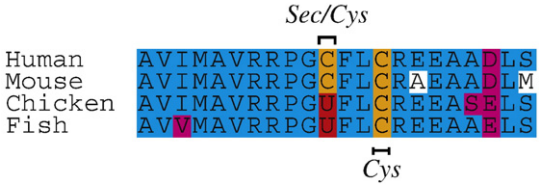

Fig. 1. A typical alignment of selenoproteins (SelUa) and their Cys-homologs. The sequence for *Takifugu rubripes*, a puffer fish, is shown. Note the symmetric conservation around the Sec codon. The observed Sec/Cys changes, however, cannot be interpreted as direct proof of the role of natural selection in shaping Sec usage in proteins.

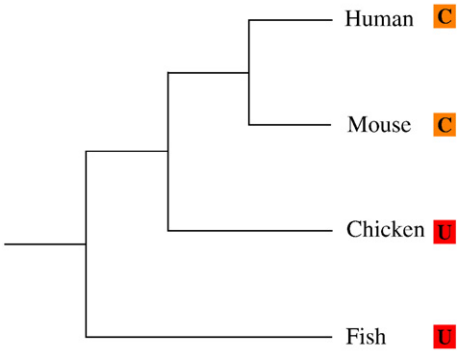

Fig. 2. Typical representation of Sec/Cys exchanges (SelUa) onto a phylogenetic tree. Note that this is a topological tree and branch lengths are not proportional to the rate of protein evolution. Assuming an ancestral Sec state, a Sec to Cys replacement has occurred in mammals. As in Fig. 1, however, this pattern of Sec/Cys replacements cannot be taken as proof of selection.

Comparative genomics thrives on well-annotated data. Homologous protein sites can be compared in a multiple sequence alignment, and the pattern of Sec/Cys exchanges revealed (Fig. 1). It is important to recognize, however, that amino acid exchanges are no proof of natural selection, and the evolutionary forces driving Sec/Cys replacements cannot be inferred straight off alignment data (Fig. 1). The mapping of Sec/Cys replacements onto the leaves of a species phylogeny, a common presentation of Sec/Cys data in the literature, provides the necessary phylogenetic context (Fig. 2). Phylogenetic trees, though, describe the outcome of millions of years of selenoprotein evolution but not the series of events that lead to what we observe today. Put differently, to interpret the leaves of a tree (present) we need to analyze the evolutionary process in its branches (past). To understand the evolutionary forces behind such process, it is necessary first to discuss some guiding evolutionary principles.

3. Basic evolutionary biology

In the same way that enzyme kinetics theory is the basis to understand enzyme catalysis, evolutionary theory underpins the study of sequence evolution. Basic evolutionary principles describe the patterns of sequence variation within species (population genetics) and variation between species (more generally known as molecular evolution). An in-depth introduction to major aspects in evolutionary biology can be found elsewhere [27,28].

3.1. Evolutionary forces

Evolution is a population level process governed by four fundamental forces. Natural selection is one of them. Natural selection acts on changes with fitness consequences, that is, changes that affect the capacity of an organism to survive and reproduce. Selection acting upon deleterious mutations is known as negative (or purifying)

selection. Similarly, selection acting upon advantageous mutations is known as positive selection. Variants that increase the fitness of an individual in its environment might increase rapidly in frequency as a result of positive selection. The identification of molecular changes subject to positive selection provides the basis to understand evolutionary adaptations at the molecular level.

The remaining three evolutionary forces, however, are nonadaptive. In consequence, they are not a function of the fitness properties of individuals. While details regarding these forces differ between species (e.g. due to different reproductive styles in bacteria and mammals), their significance remains constant in nature. First, mutation (broadly including insertions, deletions and duplications) is the fundamental source of variation on which natural selection acts. Second, recombination (including crossing-over and gene conversion) assorts variation within and among chromosomes. Finally, random genetic drift ensures that gene frequencies will deviate a bit from generation to generation independently of other forces. Random genetic drift is due to the sampling process that is inherent to reproduction. Neutral or nearly neutral mutations may then become fixed differences between species. The importance of genetic drift as an evolutionary force was recognized by the neutral theory of molecular evolution [29]. This theory, in brief, argues that most polymorphism and fixed differences between species are selectively neutral. The neutral theory, however, is not incompatible with the idea of an important role for natural selection in shaping genetic variation. Assessing the relative contribution of adaptive and nonadaptive forces to patterns of sequence diversity and divergence is a central topic in evolutionary biology [30]. Finding whether a particular amino acid difference between two species was deleterious, neutral or adaptive is the focus of many evolutionary studies. Indeed, this question, as it relates to selenium, is crucial to understand the extent of functional exchangeability between Sec and Cys amino acids in proteins.

So, as mutations go their fitness effects can be deleterious, neutral or advantageous. Although we tend to divide mutations into simple categories, there is, in reality, a continuum of fitness effects. From those strongly deleterious, through weakly deleterious mutations, to neutral, mildly or highly adaptive mutations. Fitness is an elusive quantity [31], because direct (experimental) measurement of the fitness effect of a single mutation is only possible when it has a large effect on fitness. Most mutations, even if they are deleterious, have such small effects that their fitness consequences cannot be measured. Fitness experiments involve mutating, following and measuring the survival and/or fertility of many individuals in a population, usually from a virus or unicellular species [32,33], and are difficult and time-consuming. Furthermore, whether these experiments represent the complex environments where most organisms live is unclear. For example, deletion experiments show that a majority of genes in yeast are not essential under laboratory conditions [32]. It follows that a certain fraction of genes will have marginal fitness contributions under different conditions, including natural ones. Glutathione peroxidase 1 knockout experiments are instructive in this regard. Mice deficient in this gene are healthy and fertile, even under many situations of oxidative stress or dietary deficiencies [34]. Other oxidative challenges (e.g. paraquat), however, make laboratory mice more susceptible [34]. It is this difficulty to extrapolate laboratory conditions to the natural world that hinders the estimation of fitness effects. In general, whether the deletion of selenoprotein genes, one at a time, or the substitution of Cys for Sec, a common process in nature, has a substantial effect on organismal fitness is not known (but see below). In any case, true fitness experiments have yet to be performed for Sec/Cys mutations in selenoproteins.

Short from fitness experiments, one may be tempted to use the impressive wealth of *in vitro* data on selenocysteine function to infer Sec/Cys exchangeability. Mutational data on Sec sites are important to understand the precise role of selenium in individual selenoproteins. In particular, the role of selenium in catalysis. Such mutational data, however, are far from conclusive. While increased catalytic activity is provided by Se in some enzymes [35–39], evidence for similar enzymatic efficiency between Sec and Cys residues has also been reported [39–41]. In any case, while *in vitro* data are invaluable to the study of present-day selenoprotein functions, they are not a measure of Sec/Cys organismal fitness. And, thus, they are not a measure of the strength and direction of natural selection. This is true even for catalytic sites in proteins. Functional differences alone do not demonstrate the past or present action of selection [42].

So, what is the role of functional data in evolutionary biology? In particular, how should functional information on selenoproteins and Sec/Cys sites be interpreted in evolutionary studies? In essence, functional data are necessary to the interpretation of statistical inferences of selection. To suggest possible functional reasons why selection may be acting. Despite these limitations, molecular biology plays an increasing role in evolutionary studies. More and more inferences of selection are accompanied by the reconstruction (and synthesis) of ancestral proteins in the laboratory and subsequent functional or fitness evaluation of mutations [43]. While the inference of ancestral protein states is not without statistical and evolutionary uncertainties, the rigorous study of ancestral genes can contribute to the overall interpretation of signatures of selection in sequence data. While none of these approaches have yet been applied to selenoproteins, this is a promising venue of research in selenium biology.

An additional challenge to the inference of fitness effects is that the fitness for the same amino acid substitution varies between species. Natural selection is less efficient in species with small population sizes (e.g. humans) and, therefore, the same mutation is effectively more neutral than in another species with a larger population (e.g. mouse). For example, a mutation in the same homologous Sec site may have a different effect on fitness in these two species. This is an important result in evolutionary biology, and it has been suggested as an alternative to natural selection to many aspects of genome evolution [44].

An alternative to experimental approaches is to directly infer the evolutionary forces action on the exchange of Sec/Cys residues in proteins from polymorphism or comparative data. An important benefit from this approach is that weakly selected mutations (those with small fitness effects) and other organismal effects (e.g. epistatic interactions between genes) are taken into account. This alternative is reviewed below.

3.2. Patterns of sequence divergence and diversity

Evolutionary forces leave signatures at the molecular level that can be detected using statistical tests [42,45,46]. One of the main effects of selection is to modify the levels of variability within and between species. Negative selection removes new deleterious mutations, which reduces both intraspecific and interspecific variability. Positive selection, on the other hand, decreases intraspecific variability but may increase or decrease variability between species. Such evolutionary patterns are better understood with a simple example of divergence evolution. In Fig. 3, I simulate the divergence of Sec/Cys codons in proteins along a phylogenetic tree under three different scenarios: 1) neutrality; 2) strong negative selection in some proteins and/or lineages; and 3) strong positive selection in some proteins and/or lineages. An important lesson from Fig. 3 is that conservation (or lack thereof) is a relative concept. We cannot infer the direction or strength of selection in Sec/Cys sites in Fig. 3 without comparison to the neutral standard (the expected amount of evolution in Sec/Cys sites under neutrality). This simple hypothesis, the lack of selective forces, is used to test whether the differences we observe under constraint or adaptation are statistically significant. Such tests are called neutrality tests.

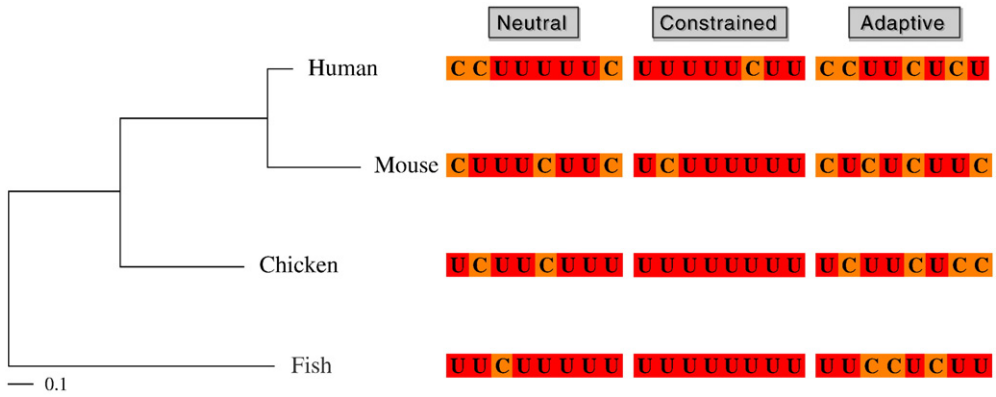

Fig. 3. Patterns of Sec/Cys divergence. Simulation of the divergence of 8 Sec codons along a phylogenetic tree under three different scenarios: 1) neutrality; 2) strong negative selection in some proteins and/or lineages, which results in less divergence between lineages; and 3) strong positive selection in some proteins and/or lineages, which, in this case, results in greater overall divergence. Branches are scaled at the expected number of substitutions per site under neutrality. The overall evolutionary forces acting on Sec residues in proteins can be assessed from comparison to the neutral standard.

3.3. Neutrality tests

One of the main interests of evolutionary biology is to distinguish molecular variation that is neutral (only affected by random genetic drift) from variation that is subject to selection, particularly positive selection. A neutrality test is a statistical test of a model in which all observed mutations are neutral. Today, dozens of neutrality tests exist, all based on neutral evolution as the null model. Rejection of the null hypothesis is usually interpreted as support for the action of natural selection. I say usually because inferences of selection are challenged by several confounding factors, especially the complex demographic history (e.g. changes in population size or structure) of natural species [42,45]. But after controlling for these confounding factors, neutrality tests provide robust inferences of natural selection.

4. Selenium biology meets evolutionary biology

Given the many challenges evolutionary studies present, a critical review of the current knowledge on the evolution of Sec/Cys usage is timely. The literature on this topic is recent, still small and somewhat detached from mainstream evolutionary biology. A common thread to many of these studies is the uncritical assumption of natural selection as an explanation to Sec/Cys exchanges. Here, I review the literature following the evolutionary principles outlined above.

4.1. Phylogenetic distribution of selenoproteins

The distribution of selenoprotein families among species has received considerable attention in the last few years. On one hand, selenoprotein families can have widely different phylogenetic distributions. For example, some selenoprotein families are present only in prokaryotic genomes, while others exist only in eukaryotes [but see [13]]. In addition, many prokaryotic and eukaryotic species have no selenoproteins but Cys-containing homologs. Furthermore, some selenoprotein families have an extremely restricted phylogenetic distribution and are not present, not even in Cys-form, in other genomes. Indeed, our view of the use and distribution of Sec-containing proteins in nature has changed over the years. It was suggested early that selenoproteins accumulated during the evolution of eukaryotes culminating in vertebrates [47]. While vertebrates have a large number of selenoproteins, non-vertebrate species can also have large selenoproteomes, for example, some spiders do (Castellano, unpublished). In addition, it has been found that selenoproteins have a more scattered distribution in eukaryotes than previously thought [8,10,12,15], and that mammals do not recapitulate the eukaryotic selenoproteome. Prokaryotic selenoproteins are even more diverse and scattered [11,13]. The pattern of Sec/Cys exchanges described by this mosaic-like distribution of selenoproteins in nature is in need of evolutionary interpretation.

4.2. Inference of constraint (purifying selection) of selenoprotein genes

Much of the natural selection acting on genomes may be negative selection acting to remove new deleterious mutations. Some studies suggest that 70–75% of amino acid altering mutations are affected by moderate or strong negative selection [48]. Strongly deleterious mutations do not segregate (polymorphism) in the population and do not contribute to differences between species. This is why the constrained pattern observed in Fig. 3 is less divergent than the pattern under neutrality.

The extent of negative selection acting on Sec/Cys exchanges is important to understand the functional equivalence of the two residues. A recent analysis of vertebrate selenoproteomes is the first assessment of the evolutionary forces acting specifically on Sec/Cys sites [49]. Fifteen selenoproteomes, encompassing 450 Myr of vertebrate evolution, were studied. A neutrality test was carried out comparing the observed divergence in the complete sets of enzymatic Sec and homologous Cys codons in these genomes, to the expected divergence under a neutral model. Such null model was obtained through neutral simulations of the evolution of ancestral Sec or Cys codons along the phylogeny. The results of this test are consistent not only with strong purifying selection acting on both Sec and Cys sites, but also with a low level of functional exchangeability between the two residues over half a billion years of vertebrate evolution. These results underscore the unique role of Sec in protein activity. Furthermore, no evidence of variation in the use of Sec and Cys residues among human populations worldwide was found. Although neutrality cannot be rejected as an explanation for the absence of variants, the absence of polymorphism observed suggests that natural variation in these sites is rare, if at all present, in human populations. This is consistent with low Sec/Cys exchangeability in human selenoproteins. A better understanding of the selenoproteomes and neutral evolutionary patterns involved in other taxa (e.g. insects) will be necessary to fully assess the generality of this conclusion.

These results are important to the ongoing discussion in the literature regarding the selective pressures acting on Sec/Cys sites [50–54]. Because the interpretation of signatures of selection in terms of their ultimate biological cause is complicated, a separate, much detailed review is provided below.

4.3. Inference of adaptive evolution (positive selection) of selenoprotein genes

While there is agreement on the importance of natural purifying selection to genome evolution, much less is known about the contribution of adaptive mutations. In general, relatively few of the mutations that are not effectively neutral are believed to be advantageous. However, although advantageous mutations are rare, they can contribute substantially to evolutionary change. For example, some studies suggest that most amino acid substitutions in *Drosophila* are subject of positive selection [55].

As a positively selected mutation increases in frequency, it leaves a distinct signature on the pattern of genomic variation. If the favored allele goes to fixation, and recurrent rounds of advantageous fixations occur in the same gene, positive selection can be detected as an increase rate of amino acid substitutions in proteins. For example, recurrent positive selection due to heterogeneous selective pressures (e.g. uneven distribution of selenium) can produce the increased overall divergence in the adaptive scenario in Fig. 3. The strength of divergence due to positive selection is, again, dependent on the neutral standard in the same figure. At the population level, strongly selected advantageous mutations that have recently become fixed (all individuals in the population carry the mutation) can leave a distinct pattern of sequence variation. As these mutations increase in frequency, they tend to reduce variation in the neighboring region where neutral variants are segregating. That is, physically linked alleles also become either fixed or lost. This process is known as a selective sweep and has been proposed to be important in the evolution of Glutathione peroxidase 1 in human populations (see below).

Recently, genome-wide analyses of human polymorphism have led to the identification of genes that seem to have been targeted by natural positive selection [56]. Because of the difficulty to interpret these patterns of variation and the complex demographic history of human population, agreement between these studies is variable. The broad functional classes of genes identified in these studies are, however, remarkably similar. Immunity and defense genes, for example, are usually inferred as targets of selection by most methods. Note, however, that the underlying cause of selection for the selected alleles is not usually clear. To my knowledge, though, no selenoprotein has been identified with confidence in these genome-wide scans of selection. An example relevant to selenium biologists for its nutritional implication is the lactase gene (*LCT*), which is necessary for lactose digestion [56]. Lactase persistence has independently evolved at least twice in geographically distinct populations. For example, the *LCT* region appears to have undergone a selective sweep 2000–20 000 years ago in Northern Europeans, coinciding with the domestication of cattle. Multiple neutrality tests support this conclusion. The European allele, however, is absent or at low frequency in African populations that are also lactose persistence. A different variant, which strongly correlates with pastoral population in Africa, is believed to provide lactase persistence in this continent.

For those of us interested in selenium studies, it is then natural to wonder whether Sec/Cys exchanges in selenoproteins (or any other amino acid replacement in these proteins) are or have been adaptive. Few studies have rigorously addressed this question. Indeed, the assumption that interspecific differences at the molecular level reveal the mechanism of evolution, and that those changes are adaptive, has so far dominated the selenium field [50–54]. Recently, however, evolutionary inferences of selection on selenoproteins and Sec/Cys sites, based on the theory above, have been carried out.

The first work to explicitly infer selection on selenoproteins studied the glutathione peroxidase 1 to 4, thioredoxin reductase 1 and selenoprotein P genes [57]. This resequencing project, 102 individuals of 4 major ethnic groups in the United States, explored sequence variation in the coding and untranslated region (including the SECIS element) of selenoprotein genes. The studied selenoproteins have antioxidant properties and it is therefore possible that population differences in selenoprotein activity and expression influence risk for a range of complex diseases (e.g. cancer). Disease genes should be under negative selection when the disease phenotype leads to a reduction of fitness. Classic neutrality tests were carried out and the observed pattern of genetic variation was found consistent with neutrality for 5 genes. The GPx1 gene, however, showed signatures of natural selection. In particular, the data showed a possible selective sweep in the Asian population. The causal interpretation of the inferred selective sweep in this gene is difficult to ascertain, and whether adaptation is in response to environmental, infectious or other pressures is not yet known. The nature of the evolutionary forces acting specifically on the Sec codon in these selenoprotein genes was, however, not pursued in this study.

More recently, a measure of the evolutionary forces acting on Sec/Cys exchanges in complete vertebrate selenoproteomes was inferred [49]. Limited evidence for a role of positive natural selection in selenoproteome evolution was found. The neutrality test applied in this work, however, may not be powerful enough to detect adaptive events in single selenoprotein in a single lineage. Therefore, whether nonneutral evolutionary processes may be responsible for some of the Sec/Cys replacements in vertebrates is not settled.

With regard to Sec/Cys exchanges, a recent exciting finding is the identification of animals with no selenoproteins. This discovery was first reported in the analysis of 12 *Drosophila* genomes [58]. While most *Drosophila* species have the previously identified selenoproteins in *D. melanogaster* [6,7], *D. willistoni* has either Cys-containing homologs or lost these genes altogether. In addition, many of the genes involved in selenoprotein synthesis have been lost in this species. Other insect genomes also appear to lack selenoproteins [59,60], and losses of Sec-containing genes in insects are likely to be independent (polyphyletic). A common interpretation to this Sec/Cys pattern is a relaxation of selective constraints on selenoprotein genes. This is an interesting hypothesis because, in contrast with vertebrates, implies high exchangeability between Sec and Cys residues in proteins. The neutrality of these replacements (lack of fitness consequences) would not support a distinct role of Sec in proteins in this particular lineage. An alternative explanation to these replacements is, however, natural positive selection favoring Cys mutations in *D. willistoni* selenoproteins. In the case of selective pressures related to protein function, as opposed to environmental factors, Sec and Cys residues in this lineage would have low exchangeability. The answer to this controversy lies in the observed pattern of Sec/Cys divergence. Whether this pattern is consistent with a neutral explanation or with functional adaptations, however, remains untested.

In conclusion, little evidence exists today for an adaptive role of Sec/Cys exchanges in proteins. While it is reasonable to expect the use of selenium to be adaptive in some selenoproteins, it is challenging to prove a direct role of natural positive selection in any single Sec residue. This is, however, a very exciting question in selenium biology and more sophisticated evolutionary tests should provide an answer to this question.

4.4. Inference of selective pressures in selenoprotein evolution

A second equally difficult question is what selective pressures account for natural selection in selenoprotein evolution. This is a particularly challenging evolutionary problem, in which knowledge of selenium and selenoprotein biology is combined with evolutionary and ecological approaches to the study of natural variation. Over the years, environmental, metabolic and biochemical selective pressures have been suggested to shape Sec use in proteins. Some classic factors are, i) the wide differences in Se status among populations due to the worldwide variability of Se content in soils and waters [61–63], which may lead to disease due to excess or deficit of Se [62]; ii) the different

Sec sensitivities to oxidation among selenoproteins and selenoproteomes due to variable O_2 levels over geologic time [64–67]; iii) the higher anabolic cost and lower translational efficiency of Sec [35,68–70]; and iv) the increased reactivity provided by Se in some selenoenzymes [35–38]. It is not immediate, though, how to assess the relative importance of such widely different selective factors in natural populations. Indeed, how to understand the biological importance and adaptive significance of inferences of selection is an area of active research in evolutionary biology. It is clear, however, that the discussion of selective pressures without regard to the statistical inference of selection is not a productive approach [30].

Here, the selective pressures acting on Sec/Cys sites are understood as a testable evolutionary hypothesis and reviewed accordingly. The extent of constraint inferred in vertebrate selenoproteomes (see above) can be interpreted in this way. We know that heterogeneous selection causes local adaptation. Most species are not distributed homogeneously throughout their geographical range, but are instead subdivided into populations that experience local conditions. We also know that nutrition is a prominent selective force in humans and other species [71] and that selenium is an unevenly distributed trace element worldwide. It is therefore not unreasonable to hypothesize that dietary adaptations due to changes in nutrient (selenium) availability have arisen in vertebrate evolution. Environmental changes and range expansions in populations may have resulted in different nutritional pressures regarding Se dietary intake, leading to high patterns of selenoproteome divergence. Nevertheless, strong conservation is observed in vertebrate selenoproteomes. Such conservation is consistent with low functional exchangeability between Sec and Cys amino acids, and a minor role for environmental Se in driving the use of Sec in vertebrate enzymes. Furthermore, despite a considerable range of variation in dietary Se intake among human populations, no evidence of variation in the use of Sec and Cys residues among populations worldwide is found [49]. Similarly, the observation that vertebrate selenoproteomes have remained similar in size, virtually unchanged in mammals, for hundreds of millions of years despite levels of atmospheric O_2 exhibiting the greatest variability of any geological period is a strong evidence of a minor role for O_2 concentrations in driving Sec use in vertebrates [49]. In addition, the constraint in vertebrate Sec sites suggests no major detrimental effect on fitness of Sec larger metabolic cost. Had Sec metabolic cost a nonnegligible fitness effect, an adaptive pattern of Sec/Cys replacements would have been observed.

The low exchangeability between Sec and Cys residues is, then, better explained by strong purifying selection due to Sec/Cys functional differences and, at best, a moderate role of environmental and metabolic forces. This result suggests caution in the interpretation of evolutionary trends in Sec usage as ecological adaptations. The functional differences responsible for the inferred constraint are, as usual, difficult to precise. It is, however, possible that the higher catalytic activity usually attributed to Sec-containing enzymes only justifies a fraction of the extensive conservation in Sec and Cys sites during vertebrate evolution. Indeed, Similar catalytic activity between homologous Sec- and Cys-containing enzymes, most likely due to additional compensatory substitutions in the active site of Cys-enzymes, has been recently reported [39–41]. Functional studies on present-day selenoproteins suggest that a broader range of substrates and pH in which selenoenzyme activity is possible [40], or other properties derived from the different catalytic mechanisms between Sec- and Cys-enzymes [41], may account for the constraint and the deleterious effect of Sec/Cys replacements in vertebrates. A more complex view of Sec in protein activity is emerging, and other biochemical and functional differences with fitness consequences may apply to the majority of uncharacterized selenoenzymes. The functional characterization of selenoproteins will be particularly relevant to evolutionary studies, in those lineages where a successful inference of natural positive selection can be made.

On the other hand, the selenium literature is rich in alternative claims regarding the role of selection in maintaining Sec/Cys residues in proteins [50–54]. These adaptive hypotheses usually include one or more *ad hoc* selective factors, which tend to become more complex as more protein changes (e.g. due to additional species sequenced) have to be explained. Such adaptive stories are very difficult to formally test, because they provide no statistical criteria to prefer one author's adaptive interpretation over another. Nevertheless, some general evolutionary principles can be of use in their discussion. For example, a large number of environmental factors have been suggested to influence the use of Sec in marine microbes [72]. This is an extremely complex selective hypothesis that includes temperature, salinity, organism density, ecosystem complexity, light for phototrophs and fixed carbon/energy for chemotrophs. These claims should be examined with caution as little statistical support exists for most of these factors. First, the claim of any selective pressure is an inference of selection and one needs to statistically reject a more simple neutral explanation. Second, because the ecological causes of geographical patterns of variation are difficult to establish, they require at least a strong statistical correlation between the molecular basis of local adaptation and an ecological factor. This is of particular importance because any number of environmental differences can always be found between ecosystems. In this context, the spurious correlation between any single selenoprotein family and an environmental factor is expected. Therefore, the test of any ecological hypothesis is its ability to explain the overall pattern of Sec/Cys exchanges. While a few selenoprotein families seem to have a different frequency in marine or nonmarine ecosystems, most selenoprotein families are equally distributed [72]. The evolutionary interpretation of such trend is that, salinity, does not seem to be a major factor in driving the overall pattern of Sec/Cys evolution. Similarly, the recent claim that large selenoproteomes associate with aquatic life and small with terrestrial life [53] is difficult to support with current data, as mammals have some of the largest selenoproteomes. To date, little support exists for an environmental role in selenoprotein evolution.

The discussion above exposes an important but sometimes unrecognized problem in evolutionary analysis, namely the difficulty to assign particular selective factors to significant inferences of natural selection. We can nevertheless rule out those environmental pressures inconsistent with Sec/Cys evolutionary patterns. It is also important to note that a growing number of Sec/Cys selective pressures in the literature are, to a large extent, contradictory. The majority of environmental factors claimed to shape Sec/Cys exchanges in nature implicitly assume a high exchangeability between Sec and Cys residues, while biochemical differences suggest low long-term exchangeability. While a few ecological factors may be shown to be important in some lineages, it seems unreasonable that each Sec/Cys exchange is not only adaptive but driven by environmental differences. In this regard, it may be comforting to many selenium researchers that the functional uniqueness of selenocysteine rivals that of the more standard amino acids in vertebrates.

5. Open evolutionary questions in selenium biology

Many other important questions about the role of selenium and selenoproteins remain open. While this review has focused only on amino acid substitutions in proteins, similar evolutionary principles can be applied to the study of other types of genomic variation. Indeed, whether differences in gene transcription, translation, number (e.g. due to gene duplication) among species are neutral or adaptive are valid evolutionary questions. For example, whether adaptation to local Se levels or other selective factors have driven the evolution of selenoprotein expression, Se intake, metabolism or transport has not been addressed. In particular, whether changes in selenoprotein family sizes between mammals and fishes have an

adaptive explanation is not known. Better theoretical and experimental approaches are needed to gain insight into these questions.

6. Concluding remark

The selenium field is ripe for the study of selenoprotein evolution. On one hand, four decades of experimental and computational efforts to identify, describe and annotate selenoproteins have provided a fairly complete characterization of many selenoproteomes. On the other, the underlying forces behind the exchange of Sec and Cys residues in most lineages remain unknown, and constitute an important evolutionary question. More so, since selenium-containing proteins may be of ecological importance in some species. For example, adaptations to local environmental conditions may prove significant in non-vertebrate lineages. The inference of such forces provides a measure of the long-term exchangeability between Sec and Cys residues in proteins. The exchangeability between these two residues, in turn, reflects the contribution of Sec to protein function. In vertebrates, Sec is a functionally unique amino acid. As we move beyond the descriptive phase of selenoprotein studies, evolutionary analyses (theoretical and experimental) will help answer many exciting questions about the twenty-first amino acid in nature.

Acknowledgement

SC thanks AM Andres for insightful comments and suggestions on the application of evolutionary theory to selenium biology.

References

[1] L. Johansson, G. Gafvelin, E.S. Arner, Selenocysteine in proteins — Properties and biotechnological use, Biochim. Biophys. Acta. 1726 (2005) 1–13.
[2] T.C. Stadtman, Selenium-dependent enzymes, Annu. Rev. Biochem. 49 (1980) 93–110.
[3] M.J. Axley, T.C. Stadtman, Selenium metabolism and selenium-dependent enzymes in microorganisms, Annu. Rev. Nutr. 9 (1989) 127–137.
[4] G.V. Kryukov, V.M. Kryukov, V.N. Gladyshev, New mammalian selenocysteine-containing proteins identified with an algorithm that searches for selenocysteine insertion sequence elements, J. Biol. Chem. 274 (1999) 33888–33897.
[5] A. Lescure, D. Gautheret, P. Carbon, A. Krol, Novel selenoproteins identified *in silico* and *in vivo* by using a conserved RNA structural motif, J. Biol. Chem. 274 (1999) 38147–38154.
[6] S. Castellano, N. Morozova, M. Morey, et al., *in silico* identification of novel selenoproteins in the *Drosophila melanogaster* genome, EMBO rep. 2 (2001) 697–702.
[7] F.J. Martin-Romero, G.V. Kryukov, A.V. Lobanov, et al., Selenium metabolism in *Drosophila*: selenoproteins, selenoprotein mRNA expression, fertility, and mortality, J. Biol. Chem. 276 (2001) 29798–29804.
[8] S.V. Novoselov, M. Rao, N.V. Onoshko, et al., Selenoproteins and selenocysteine insertion system in the model plant cell system, *Chlamydomonas reinhardtii*, EMBO J. 21 (2002) 3681–3693.
[9] G.V. Kryukov, S. Castellano, S.V. Novoselov, et al., Characterization of mammalian selenoproteomes, Science 300 (2003) 1439–1443.
[10] S. Castellano, S.V. Novoselov, G.V. Kryukov, et al., Reconsidering the evolution of eukaryotic selenoproteins: a novel nonmammalian family with scattered phylogenetic distribution, EMBO rep. 5 (2004) 71–77.
[11] G.V. Kryukov, V.N. Gladyshev, The prokaryotic selenoproteome, EMBO Rep. 5 (2004) 538–543.
[12] S. Castellano, A.V. Lobanov, C. Chapple, et al., Diversity and functional plasticity of eukaryotic selenoproteins: identification and characterization of the SelJ family, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 16188–16193.
[13] Y. Zhang, D.E. Fomenko, V.N. Gladyshev, The microbial selenoproteome of the Sargasso Sea, Genome Biol. 6 (2005) R37.
[14] A.V. Lobanov, C. Delgado, S. Rahlfs, et al., The Plasmodium selenoproteome, Nucl. Acids Res. 234 (2006) 496–505.
[15] V.A. Shchedrina, S.V. Novoselov, M.Y. Malinouski, V.N. Gladyshev, Identification and characterization of a selenoprotein family containing a diselenide bond in a redox motif, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 13919–13924.
[16] J.E. Cone, R.M. Del Río, J.N. Davis, T.C. Stadtman, Chemical characterization of the selenoprotein component of clostridial glycine reductase: identification of selenocysteine as the organoselenium moiety, Proc. Natl. Acad. Sci. U. S. A. 73 (1976) 2659–2663.
[17] K. Taskov, C. Chapple, G.V. Kryukov, et al., Nematode selenoproteome: the use of selenocysteine insertion system to decode one codon in an animal genome? Nucl. Acids Res. 33 (2005) 2227–2238.
[18] D.E. Bergstrom, C.A. Merli, J.A. Cygan, R. Shelby, R.K. Blackman, Regulatory autonomy and molecular characterization of the *Drosophila out at first* gene, Genetics 139 (1995) 1331–1346.
[19] D.N. Robinson, L. Cooley, Examination of the function of two kelch proteins generated by stop codon suppression, Development 124 (1997) 1405–1417.
[20] L. Jiang, Q. Liu, P. Chen, Z. Gao, H. Xu, New selenoproteins identified *in silico* from the genome of *Anopheles gambiae*, Sci. China C. Life Sci. 50 (2007) 251–257.
[21] L. Jiang, Q. Liu, X. Wang, An Improved Method for the Prediction of Selenoproteins from the Genome of *Anopheles gambiae*, Bioinformatics and Biomedical Engineering, ICBBE (2008) 592–595.
[22] M.C. Frith, L.G. Wilming, A. Forrest, et al., Pseudo-messenger RNA: phantoms of the transcriptome, PLoS Genet. 2 (2006) e23.
[23] S. Castellano, Little biological and statistical support for hundreds of selenoproteins in mouse pseudo-messenger RNAs, PLoS Genet. eLetters. In response to M.C. Frith, L.G. Wilming, A. Forrest et al., (2006) Pseudo-messenger RNA: phantoms of the transcriptome, PLoS Genet. 2 (2006) e23.
[24] S.F. Altschul, T.L. Madden, A.A. Schäffer, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucl. Acids Res. 25 (1997) 3389–3402.
[25] S. Karlin, S.F. Altschul, Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, Proc. Natl. Acad. Sci. U. S. A. 87 (1990) 2264–2268.
[26] S. Castellano, V.N. Gladyshev, R. Guigó, M.J. Berry, SelenoDB 1.0: a database of selenoprotein genes, proteins and SECIS elements, Nucl. Acids Res. 36 (2008) D339–D343.
[27] W.H. Li, Molecular Evolution, Sinauer Associates, Inc., Sunderland Massachusetts, 1997.
[28] D.L. Hartl, A.G. Clark, Principles of Population Genetics, Fourth Ed.Sinauer Associates, Inc., Sunderland Massachusetts, 2007.
[29] M. Kimura, Evolutionary rate at the molecular level, Nature 217 (1968) 624–626.
[30] T. Mitchell-Olds, J.H. Willis, D.B. Goldstein, Which evolutionary processes influence natural genetic variation for phenotypic traits? Nat. Rev. Genet. 8 (2007) 845–856.
[31] A. Eyre-Walker, P.D. Keightley, The distribution of fitness effects of new mutations, Nat. Rev. Genet. 8 (2007) 610–618.
[32] J.W. Thatcher, J.M. Shaw, W.J. Dickinson, Marginal fitness contributions of nonessential genes in yeast, Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 253–257.
[33] R. Sanjuán, A. Moya, S.F. Elen, The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 8396–8401.
[34] J.B. de Haan, C. Bladier, P. Griffiths, et al., Mice with a homozygous null mutation for the most abundant glutathione peroxidase, Gpx1, show increased susceptibility to the oxidative stress-inducing agents paraquat and hydrogen peroxide, J. Biol. Chem. 273 (1998) 22528–22536.
[35] M.J. Berry, A.L. Mai, J. Kieffer, J.W. Harney, P. Larsen, Substitution of cysteine for selenocysteine in type I iodothyronine deiodinase reduces the catalytic efficiency of the protein but enhances its translation, Endocrinology 131 (1992) 1448–1852.
[36] C. Rocher, J.L. Lalanne, J. Chaudiere, Purification and properties of a recombinant sulfur analog of murine selenium-glutathione peroxidase, Eur. J. Biochem. 205 (1992) 955–960.
[37] M. Maiorino, K.D. Aumann, R. Brigelius-Flohé, et al., Probing the presumed catalytic triad of selenium-containing peroxidases by mutational analysis of phospholipid hydroperoxide glutathione peroxidase (PHGPx), Biol. Chem. Hoppe Seyler 376 (1995) 651–660.
[38] L. Zhong, A. Holmgren, Essential role of selenium in the catalytic activities of mammalian thioredoxin reductase revealed by characterization of recombinant enzymes with selenocysteine mutations, J. Biol. Chem. 275 (2000) 18121–18128.
[39] S.M. Kanzok, A. Fechner, H. Bauer, et al., Substitution of the thioredoxin system for glutathione reductase in *Drosophila melanogaster*, Science 291 (2001) 643–646.
[40] S. Gromer, L. Johansson, H. Bauer, et al., Active sites of thioredoxin reductases: why selenoproteins? Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 12618–12623.
[41] H.Y. Kim, V.N. Gladyshev, Different catalytic mechanisms in mammalian selenocysteine- and cysteine-containing methionine-R-sulfoxide reductases, PLoS Biol. 3 (2005) e375.
[42] R. Nielsen, I. Hellmann, M. Hubisz, C. Bustamante, A.G. Clark, Recent and ongoing selection in the human genome, Nat. Rev. Genet. 8 (2007) 857–868.
[43] J.W. Thornton, Resurrecting ancient genes: experimental analysis of extinct molecules, Nat. Rev. Genet. 5 (2004) 366–375.
[44] M. Lynch, J.S. Conery, The origins of genome complexity, Science 302 (2003) 1401–1404.
[45] M. Bamshad, S.P. Wooding, Signatures of natural selection in the human genome, Nat. Rev. Genet. 4 (2003) 99–111.
[46] R. Nielsen, Molecular signatures of natural selection, Annu. Rev. Genet. 39 (2005) 197–218.
[47] V.N. Gladyshev, G.V. Kryukov, Evolution of selenocysteine-containing proteins: significance of identification and functional characterization of selenoproteins, BioFactors 14 (2001) 87–92.
[48] A. Eyre-Walker, P.D. Keightley, High genomic deleterious mutation rates in hominids, Nature 397 (1999) 344–347.
[49] S.. Castellano, A.M. Andrés, E. Bosch, M. Bayes, R. Guigó, A.G Clark, Low exchangeability of selenocysteine, the 21st amino acid, in vertebrate proteins, Mol. Bio. Evo. 26 (2009) 2031–2040.
[50] H. Romero, Y. Zhang, V.N. Gladyshev, G. Salinas, Evolution of selenium utilization traits, Genome Biol. 6 (2005) R66.
[51] Y. Zhang, H. Romero, G. Salinas, V.N. Gladyshev, Dynamic evolution of selenocysteine utilization in bacteria: a balance between selenoprotein loss and

evolution of selenocysteine from redox active cysteine residues, Genome Biol. 7 (2006) R94.
[52] D.E. Fomenko, W. Xing, B.M. Adair, D.J. Thomas, V.N. Gladyshev, High-throughput identification of catalytic redox-active cysteine residues, Science 315 (2007) 387–389.
[53] A.V. Lobanov, D.E. Fomenko, Y. Zhang, A. Sengupta, D.L. Hatfield, V.N. Gladyshev, Evolutionary dynamics of eukaryotic selenoproteomes: large selenoproteomes may associate with aquatic life and small with terrestrial life, Genome Biol. 8 (2007) R198.
[54] A.V. Lobanov, D.L. Hatfield, V.N. Gladyshev, Reduced reliance on the trace element selenium during evolution of mammals, Genome Biol. 9 (2008) R62.
[55] S.A. Sawyer, R.J. Kulathinal, C.D. Bustamante, D.L. Hartl, Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection, J. Mol. Evol. 57 (2003) S154–164.
[56] J.L. Kelley, W.J. Swanson, Positive selection in the human genome: from genome scans to biological significance, Annu. Rev. Genomics Hum. Genet. 9 (2008) 143–160.
[57] C.B. Foster, K. Aswath, S.J. Chanock, H.F. McKay, U. Peters, Polymorphism analysis of six selenoprotein genes: support for a selective sweep at the glutathione peroxidase 1 locus (3p21) in Asian populations, BMC Genet. 7 (2006) 56.
[58] *Drosophila* 12 Genomes Consortium, Evolution of genes and genomes on the *Drosophila* phylogeny, Nature 450 (2007) 203–218.
[59] C.E. Chapple, R. Guigó, Relaxation of selective constraints causes independent selenoprotein extinction in insect genomes, PLoS ONE 3 (2008) e2968.
[60] A.V. Lobanov, D.L. Hatfield, V.N. Gladyshev, Selenoproteinless animals: selenophosphate synthetase SPS1 functions in a pathway unrelated to selenocysteine biosynthesis, Protein Sci. 17 (2008) 176–182.
[61] R.J. Shamberger, Selenium in the environment, Sci. Total Environ. 17 (1981) 59–74.
[62] O.A. Levander, A global view of human selenium nutrition, Ann. Rev. Nutr. 7 (1987) 227–250.
[63] J.L. Valentine, Environmental occurrence of selenium in waters and related health significance, Biomed. Environ. Sci. 10 (1997) 292–299.
[64] W. Leinfelder, E. Zehelein, M. MandrandBerthelot, A. Bock, Gene for a novel tRNA species that accepts L-serine and cotranslationally inserts selenocysteine, Nature 331 (1988) 723–725.
[65] T.H. Jukes, Genetic code 1990, Experientia 46 (1990) 1149–1157.
[66] R.A. Berner, GEOCARBSULF: a combined model for Phanerozoic atmospheric O_2 and CO_2, Geochim. Cosmochim. Acta 70 (2006) 5653.
[67] R.A. Berner, J.M. VandenBrooks, P.D. Ward, Oxygen and evolution, Science 316 (2007) 557–558.
[68] D.M. Driscoll, D.R. Copeland, Mechanism and regulation of selenoprotein synthesis, Annu. Rev. Nutr. 23 (2003) 17–40.
[69] A. Mehta, C.M. Rebsch, S.A. Kinzy, J.E. Fletcher, P.R. Copeland, Efficiency of mammalian selenocysteine incorporation, J. Biol. Chem. 279 (2004) 37852–37859.
[70] X.M. Xu, et al., Biosynthesis of selenocysteine on its tRNA in eukaryotes, PLoS Biol. 5 (2007) e4.
[71] R. Haygood, O. Fedrigo, B. Hanson, K.D. Yokoyama, G.A. Wray, Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution, Nat. Genet. 39 (2007) 1140–1144.
[72] Y. Zhang, V.N. Gladyshev, Trends in selenium utilization in marine microbial world revealed through the analysis of the global ocean sampling (GOS) project, PLoS Genet. 4 (2008) e1000095.