The high conservation of the genetic code and its fundamental role in genome decoding suggest that its evolution is highly restricted or even frozen. However, various prokaryotic and eukaryotic genetic code alterations, several alternative tRNA-dependent amino acid biosynthesis pathways, regulation of tRNA decoding by diverse nucleoside modifications and recent in vivo incorporation of non-natural amino acids into prokaryotic and eukaryotic proteins, show that the code evolves and is surprisingly flexible. The cellular mechanisms and the proteome buffering capacity that support such evolutionary processes remain unclear. Here we explore the hypothesis that codon mis-reading and reassignment played fundamental roles in the development of the genetic code and we show how a fungal codon reassignment is enlightening its evolution.

1. Introduction

Life is based on the extraordinary capacity of cells to translate the nucleic acids information of their genomes into the amino acids information of their proteomes. The genetic code determines how gene words (codons) are translated into protein words (amino acids), highlighting the fundamental role of 20 aminoacyl-tRNA synthetases (aaRSs) in genome decoding [1]. Each aaRS binds and activates a specific amino acid and transfers it to a cognate tRNA, producing aminoacylated tRNAs (aa-tRNAs) [2,3]. The latter read mRNA codons translating the nucleic acids alphabet into the amino acids alphabet through specific ribosome dependent decoding rules [4]. The genetic code is therefore established by specific attachment of amino acids onto tRNA adaptor molecules by aaRSs and by direct reading of mRNA codons by aa-tRNA anticodons in the ribosome. This suggests that reconstruction of the evolutionary pathways that established the genetic code requires deep structural, biochemical, functional and evolutionary knowledge of aaRSs, tRNAs, mRNAs and of the ribosome. To date, many crystal structures of these molecules have been obtained, and detailed biochemical and biophysical characterization of the tRNA aminoacylation and decoding reactions [2,5–7], as well as large scale phylogenetic analysis of the various components of the genetic code have been carried out [8]. Despite these extraordinary advances, the evolution of the genetic code remains an open biological question.

The Frozen Accident Theory proposed by Crick in 1968 postulates that the code is immutable because any alteration to it would be lethal or highly detrimental to life [9]. However, a number of genetic code alterations discovered over the last 40 years indicate that the code has intrinsic flexibility and can evolve (reviewed in [10,11]). We discuss below how these genetic code alterations are enlightening the evolution of the genetic code and we raise the hypothesis that codon reassignment processes played an important role in the code development. The origin of the genetic code, i.e., the origin of tRNAs, aaRSs, the ribosome and the mechanisms of incorporation of the first 10 prebiotic amino acids into the code, which mediated the transition of life from the RNA to the protein worlds, are beyond the scope of this review and will not be addressed. We mention briefly the main theories that have been proposed to explain the origin of the genetic code in order to provide an integrated view of the code evolution.

2. Origin and early evolution of the genetic code

There are three main theories to explain the origin and structure of the genetic code, namely: (i) the Stereochemical Theory, (ii) the Adaptive Theory and (iii) the Coevolution Theory (reviewed in [12]). The Stereochemical Theory posits that codon and amino acid assignments were determined by physicochemical affinities
between amino acids and nucleic acids [13,14]. This theory is supported by experimental data arising from selection-activation of small RNAs (SELEX) which show that at least 8 of the 20 natural amino acids select RNA sequences enriched in cognate codon or anticodon binding motifs [15,16]. Indeed, RNA aptamers selected in the presence of Trp contained Trp CCA anticodons while small RNAs selected in the presence of Ile were enriched in Ile UAU anticodons [17–19], however the statistical significance and the strength of the associations between RNA aptamers and amino acids has been questioned and the Stereochimical Theory requires further validation [20].

The Adaptive Theory postulates that the evolution of the genetic code is mainly driven by the selective forces that minimize the effects of protein synthesis errors, being them from mutational origin or from mRNA misreading [21,22]. The observation that amino acids with similar chemical properties are assigned to similar codons plus statistical and computational evidence for a strong bias towards error minimization pressure in the code provide important support for this theory [12,23,24].

The Coevolution Theory postulates that the structure of the genetic code reflects directly the evolution of amino acid biosynthetic pathways [25]. This theory assumes that the number of amino acids that existed in the prebiotic earth was small (10 or so) and that the other amino acids of the genetic code were derived from the prebiotic ones through biosynthetic processes. The theory is supported by the identification of precursor-product pairs of amino acids and by the discovery of tRNA-dependent biosynthesis of Gln, Asn, Cys and Sec in various prokaryotes and eukaryotes (see below) [26].

The evolutionary scenarios described above, in particular the one proposed by the Coevolution Theory, suggest the existence of three critical moments (steps) in the development of the genetic code (Fig. 1A). An initial step (Phase-1) characterized by the incorporation of the prebiotic amino acids Gly, Ala, Ser, Asp, Glu, Val, Leu, Ile, Pro and Thr. An intermediate step (Phase-2) involving the incorporation of 7 additional amino acids derived from the prebiotic ones through biosynthetic means, namely Phe, Tyr, Arg, His, Trp, Lys and Met. And, a final step (Phase-3) where the five amino acids whose synthesis is tRNA-dependent or is mediated through non-canonical biosynthetic pathways, namely Asn, Gln, Cys, selenocysteine (Sec) and pyrrolysine (Pyl), were incorporated into the genetic code [12,25,26].

We discuss below the mechanistic and structural implications of this stratified evolution of the genetic code under the assumption of the following postulates for the Phase-1 of the code development:

1. The triplet nucleotide nature of codons and the translational machinery were largely established during the incorporation of the first 10 prebiotic amino acids into the genetic code.
2. The basic structure of the tRNA molecule and the codon–anticodon decoding principles were defined.
3. An essential proteome was synthesized with the 10 prebiotic amino acids.

The simultaneous existence of only 10 prebiotic amino acids and 64 codons suggests that some codons were initially unassigned (did not code for any amino acid) or that the 10 prebiotic amino acids were assigned to more than one codon family box (Fig. 1B), as is the case for Leu, Ser and Arg, in extant organisms (Fig. 1D). Indeed, Leu is still encoded by the CUN (N = any nucleotide) codon family plus the UUA/G codons of the UUN codon family box (Fig. 1C and D). The other two codons of the UUN codon family box (UUA/C codons) encode Phe, which was incorporated late into the genetic code [26]. Therefore, during Phase-1 of the code development Leu must have been assigned to both the CUN and the UUN
codon family boxes (Fig. 1B). Phe addition to the code required a new (mutant) tRNA$^\text{Phe}$ to capture the UUU/C codons from Leu. Complete reassignment of these codons to Phe required the loss of the ancestral tRNA$^\text{Leu}$ that decoded them (Fig. 1C). The same principle of codon capture followed by reassignment can be applied to the incorporation of the other Phase-2 amino acids (Fig. 1C and D). An alternative explanation would be that UUU/UUC, as well as the other codons of split codon families, were initially unassigned and that their late assignment to new amino acids escaped reassignment from one amino acid to another. However, tRNAs with U at the wobble position are able to decode the four codons of codon family boxes and it is likely that these rather than more sophisticated tRNAs bearing nucleoside modifications or expanded sets of tRNA isoacceptors were originally used to decode the 61 sense codons of the genetic code. Furthermore, the pairs of codons of split codon family boxes end with a purine or a pyrimidine and consequently cannot be unassigned simultaneously by genome G + C pressure alone. Therefore, it is unlikely that codon unassignment played a relevant role in the early amino acid assignments.

The Phase-3 amino acids (Asn, Gln, Cys, Sec, Pyl, fMet) are particularly interesting because their alternative biosynthesis suggests that they were incorporated rather late into the genetic code [27,28]. In various bacterial and archaeal species, Asn is still synthesized on a tRNA$^\text{Asn}$ which is charged by Asp by a non-discriminating Asp-tRNA$^\text{Asn}$ synthetase. A similar mechanism is used in archaea, in most bacteria and in chloroplasts for the synthesis of Gln. In this case, a tRNA$^\text{Gln}$ is charged
with Glu by a non-discriminating GluRS generating a mischarged Glu-tRNA^{Glu} [30]. Downstream reactions catalysed by the amido transferases Asp-AdT and Glu-AdT convert Asp and Glu into Asn and Gln, respectively [31]. A slightly different mechanism is used in methanogenic archaea to synthesize cysteine (Cys). In this case, a tRNA^{ASL} is initially charged with O-phosphoserine (Sep) by the enzyme O-phosphoseryl-tRNA synthase (SepRS), generating the mischarged Sep-tRNA^{ASL}, a Sep-tRNA: Cys-tRNA synthase (SepCysS) then transforms Sep-tRNA^{Cys} into Cys-tRNA^{Cys} [32]. On the other hand, selenocysteine is synthesized on a tRNA^{SecSec} which is first aminocylated with serine by a SerRS. In bacteria, a Sec synthetase (SecS) converts Ser into Sec [33], while in archaea and in eukaryotes the seryl moiety is O-phosphorylated by the O-phosphoseryl-tRNA kinase (PSTK) and the phosphate group is then converted into selenocysteinyl-tRNA^{Sec} (Sec-tRNA^{Sec}) by the Sep-tRNA:Sec-tRNA synthase [34–36]. Sec is inserted at specific UGA codons which are recoded by a mRNA cis-element named SECIS [33,37]. Finally, pyrrolysine is synthesized in Methanosarcinales using a metabolic pathway involving the PylB, PylC and PylD genes and α-ornithine as precursor. Pyl is charged directly on the tRNA^{Pyl} by a PylRS and is then incorporated into the genetic code in response to specific UAG codons [38,39].

The late incorporation of Asn, Gln, Cys, Sec and Pyl into the genetic code and their assignment to codons that belong to split codon families (Fig. 1D) suggest that their codons were reassigned during Phase-3 of the code development. If so, incorporation of Gln into the genetic code should have required the reassignment of the CAG/G codons which were originally assigned to His (Phase-2 amino acid) (Fig. 1C). Similarly, Asn incorporation into the code should have required capture of the AAU/C codons from the respective decoding tRNA due to lack of selective pressure to maintain its gene. Unassigned codons are unassigned in the G + C rich genome of Bacillus subtilis (75% GC), and the CGG (Arg) codon is unassigned in the A + T rich genome of Mycoplasma spp. whose genome is A + T rich (75% AT). (B) Alternatively, mutant or wild-type tRNAs may misread codons and in bacteria and also in the nuclear genome of unicellular eukaryotes are helping to clarify this question.

### 3. Genetic code alterations

To date, a number of alterations to the standard genetic code have been discovered in various organisms. Stop codons have been reassigned to Trp, Glu, Gln, Cys and Pyl and have also been used to expand the genetic code to Sec and Pyl (reviewed in [10]). The AUA codon has been reassigned from Ile to Met, the AGA/G (Arg) codons have been reassigned to Ser, Gly or to Stop and the AAA (Lys) codon has been reassigned to Trp in Saccharomyces cerevisiae (Leon) (green colour) and Batophora cesthedi (blue colour). The UAA and UAG stop codons are also involved in the model Tetrahymena thermophila, in the green algae Acetabularia spp. and Batophora cesthedi the UAA and UAG stop codons have been reassigned to Gln (reviewed in [10]).

The evolution of genetic code alterations can be explained by the Codon Capture and the Ambiguous Intermediate theories [44,45]. The Codon Capture theory postulates that G + C pressure minimizes through gradual reduction of the codon usage. This decreases fitness and should be eliminated by the Codon Capture theory. The Codon Capture theory postulates that G + C pressure
plays a major role in the evolution of genetic code alterations via its biased effects on codon usage [44,46] (Fig. 3A). The theory posits that codons can disappear from genomes due to strong G + C or A + T replication pressure, and is supported by the unassignment of the AGA, AUA codons in Micrococcus spp. (75% GC) and the CGG codon in Mycoplasma spp. (25% GC). The theory also postulates that such unassigned codons promote reassignment if they reappear in the genome, due to alteration in the DNA replication bias. Their reassignment is mediated by non-cognate tRNAs that misread them [47]. However, if such misreading tRNAs do not exist, the re-emerged codons block mRNA decoding and can be toxic [48], but the theory does not provide a mechanism to circumvent such toxicity. Also, this theory cannot explain reassignment of codons in the absence of DNA replication biases or in cases where the usage of the reassigned codon is favoured by such bias. Examples of such exceptions are the reassignment of the UGA stop codon to Trp, the UAA from Stop to Tyr, the UAU from Ile to Met, the AAA from Lys to Asn and the AGA from Arg to Ser or Stop in A + T rich mitochondria (Fig. 2). The reassignment of the entire Leu CUN codon family to Thr in fungal mitochondria or the reassignment of the Leu CUG codon to Ser in some fungal species also escape the Codon Capture theory [10,11]. These codon reassignments are better explained by the Ambiguous Intermediate theory, which postulates that misreading tRNAs can capture codons from their cognate tRNAs through a selection-driven process involving gradual increase of misreading efficiency and subsequent disappearance of cognate tRNAs (Fig. 3B) [45,49]. In this process, the codons become ambiguous and their reassignment introduces significant proteome disruption. The theory does not explain how codon ambiguity is selected, but it is strongly supported by CUG reassignment from Leu to Ser in fungi (see below).

The prevalence of genetic code alterations in mitochondria highlights yet another important feature of the evolution of the genetic code, namely that proteome size imposes strong negative pressure on codon reassignment. This has been demonstrated by a large scale comparative genomics study showing a negative correlation between the number of genetic code alterations and the number of genes encoded by mitochondrial genomes [50]. This principle is nicely illustrated in human mitochondria where only 13 of the 900 or so proteins of its proteome are encoded by its genome [51]. Since nuclear encoded proteins are synthesized in the cytoplasm using the standard genetic code and are transported into the mitochondria using a signal peptide translocation system their synthesis escapes the disruption caused by mitochondrial codon reassignments. This is in line with the Genome Minimization hypothesis which postulates that replication speed imposes a strong negative pressure on the mitochondrial genome, leading to selection of small size genomes [52]. In other words, changes in codon usage that relax the pressure to maintain certain tRNA and release factor genes lead to their disappearance and favour codon unassignments and/or reassignments [49]. Therefore, the data available indicate that low codon usage, codon unassignment, genome G + C pressure, genome minimization, small proteome size and tRNA disappearance, are critical players in the evolution of the genetic code. Interestingly, plant mitochondria escape somehow the effects of these evolutionary forces and maintain the standard genetic code.

4. A fungal genetic code alteration

In several species of the genus Candida and Debaryomyces, the so-called CTG clade, Leu CUG codons are decoded as Ser by a novel seryl-tRNA<sub>CAG</sub> (tRNA<sub>Ser<sup>CAG</sup></sub>) (Fig. 4A) [11,53–57]. Since these species have sophisticated genomes encoding thousands of genes (7000 genes on average) [56] and Leu and Ser are chemically distinct – Leu is hydrophobic and is located in the hydrophobic core of proteins while Ser is polar and is located on the surface of proteins in direct contact with the solvent – the reassignment of CUG codons should have caused maximal protein structural disruption [58]. Moreover, the ancestor of the CTG clade species reassigned between 26 000 and 30 000 CUG codons which existed in approximately 50% of its 7000 or so genes at an average frequency of 1–6 CUGs per gene, indicating that more than half of the proteins of the fungal ancestor had their structure affected [56,59].

The reassignment of those CUG codons was initiated 275 ± 25 million years (My) ago by a mutant Ser tRNA that acquired a 5'-CAG-3' Leu anticodon (tRNA<sup>Ser</sup><sub>CAG</sub>) [59] (Fig. 4A). During the early stages of CUG reassignment, the tRNA<sub>Ser</sub><sup>CAG</sup> competed with a
CUG codons was reconstructed in S. cerevisiae. Could CUG misreading have driven evolution of the CTG clade? Interestingly, recent comparative genomics analysis showed that key components of the mating and meiosis pathways are missing in species of the Candida clade. The human pathogen Candida albicans uses a parasexual mating system that does not involve nuclear fusion or meiosis, other species apparently do not mate, while a subgroup of species use homothallic and others use heterothallic sexual cycles. Such mating diversity is related to high genetic diversity of the mating locus (MTL) which controls mating isoenzymes, and the resistance to several stress agents, namely proteases, and had a strong effect on cell adhesion on solid surfaces. Furthermore, A35 and G37 allowed for charging of the tRNAGlu with Leu by its cognate LeuRS and did not interfere with serylation of the tRNA Ser.

Another important consequence of CUG misreading was a drastic reduction in the usage of this codon in the ancestor of the CTG clade species. Comparative genomics of orthologous genes of CTG and non-CTG clade species demonstrated that the original CUG codons present in the genome of the ancestor of the CTG clade species were eliminated (Fig. 4B) and that the CUGs present in the genomes of extant CTG clade species evolve recently from Ser rather than Leu codons. The biological implications of this finding are profound. Firstly, they confirm that codon decoding fidelity is a major selective force in the evolution of codon usage, which is compatible with the Adaptive Theory. Secondly, it demonstrates that codons can disappear from genomes due to tRNA misreading rather than biased genome G+C pressure. Thirdly, it shows unequivocally that mutant tRNAs with novel decoding properties can capture codons from unrelated codon families (different amino acids) and that codons can be rapidly reintroduced in the genome from mutation of other codons, even in cases where 2 or 3 simultaneous mutations are required (Fig. 4B).

6. Genetic code ambiguity as a phenotypic diversity generator

The study of CUG reassignment in C. albicans has also shown that CUG misreading is an important phenotypic diversity generator. This is in line with the phenotypes observed in S. cerevisiae (see above), however the phenotypic variation observed in C. albicans was far more extensive and relevant to adaptation than that observed in S. cerevisiae. In the former, CUG misreading activated morphogenetic pathways that led to expression of an array of highly variable cell and colony morphological phenotypes. The study of CUG reassignment in C. albicans also suggested that CUG misreading was a major force in the evolution of codon usage, which is compatible with the Adaptive Theory. Secondly, it demonstrates that codons can disappear from genomes due to tRNA misreading rather than biased genome G+C pressure. Thirdly, it shows unequivocally that mutant tRNAs with novel decoding properties can capture codons from unrelated codon families (different amino acids) and that codons can be rapidly reintroduced in the genome from mutation of other codons, even in cases where 2 or 3 simultaneous mutations are required (Fig. 4B).
which may have allowed for selection of the genetic code alteration (Fig. 6).

7. Conclusions

Understanding the evolution of the genetic code is fundamental to elucidate the origin of life. Over the last 60 years remarkable progress has been made on the structural and functional analysis of the various components of the genetic code, namely tRNAs, aaRSs, the ribosome and the overall mRNA translation process. However, the development of the genetic code remains unclear. The hypothesis that codon reassignment played an important role during the early evolution of the code is supported by the gradual incorporation of amino acids into it and by its expansion from 20 to 22 amino acids. The existence of genetic code alterations in extant organisms which evolved from the standard code explains how codons can be reassigned.

Considering that most codon reassignments cannot be explained by neutral mechanisms, it is likely that codon ambiguity was a prevalent mechanism in the early evolution of the genetic code, raising the question of how did the code evolve under negative selection. The phenotypic diversity generated in S. cerevisiae and C. albicans by codon misreading shows that codon ambiguity may have positive outcomes. Also, the incorporation of selenocysteine and pyrrolysine into the genetic code shows how incorporation of new amino acids can expand the proteome and create new functional classes of proteins which bring about selective advantages. Therefore, codon misreading is not necessarily disadvantageous. Since its effects can be largely overcome by proteome novelty and also by indirect advantages such as phenotypic diversity and adaptation to new environmental conditions.

Extrapolation of the cellular consequences of codon misreading and reassignment in living organisms, such as fungi, to the early evolution of the genetic code requires a degree of caution. Clearly the differences between modern cells and the primordial forms of life that existed before the last universal common ancestor (LUCA) are paramount. The variety of extant genetic code alterations also suggest that the forces and mechanisms that mediate the evolution of genetic code alterations are complex and diverse, thus preventing generalizations. However, genome minimization and the role of small proteome size in mitochondrial codon reassignments, suggest that the small size of the pre-LUCA proteomes may have facilitated codon reassignments during Phase-2 and Phase-3 of the code development. In this scenario, the proteome novelty arising from incorporation of new amino acids into the genetic code should have been the major driving force for the expansion of the genetic code up to 22 amino acids, as is amply demonstrated from the advantages generated from the introduction of selenocysteine and pyrrolysine into the genetic code.

Acknowledgements

The Santos group is supported by the Portuguese Foundation for Science and Technology (FCT) through Projects PTDC/BIA-BMC/
64745/2006 and 72251/2006 and also by the European FP7 project MEPHITIS.

References


and evolutionary implications of genes for tRNA(Ser)CAG responsible for translation of a non-universal genetic code. Nucleic Acids Res. 22, 115–123.

