Circular code motifs in the ribosome: a missing link in the evolution of translation?

Gopal Dila, Raymond Ripp, Claudine Mayer, Olivier Poch, Christian Michel, Julie D Thompson

▶ **To cite this version:**

Gopal Dila, Raymond Ripp, Claudine Mayer, Olivier Poch, Christian Michel, et al.. Circular code motifs in the ribosome: a missing link in the evolution of translation?. RNA, 2019, 25 (12), pp.1714-1730. 10.1261/rna.072074.119 . hal-02381481

HAL Id: hal-02381481
https://hal.science/hal-02381481

Submitted on 4 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Circular code motifs in the ribosome: a missing link in the evolution of translation?

Gopal Dila[1], Raymond Ripp[1], Claudine Mayer[1,2,3], Olivier Poch[1], Christian J. Michel[1,*] and Julie D. Thompson[1,*]

[1] *Department of Computer Science, ICube, CNRS, University of Strasbourg, Strasbourg, France*

[2] *Unité de Microbiologie Structurale, Institut Pasteur, CNRS, 75724 Paris Cedex 15, France*

[3] *Université Paris Diderot, Sorbonne Paris Cité, 75724 Paris Cedex 15, France*

* To whom correspondence should be addressed; Email: thompson@unistra.fr

***Corresponding authors:**

Names: Christian J. Michel, Julie D. Thompson

Address: Department of Computer Science, ICube, Strasbourg, France

Phone: (33) 0368853296

Email: c.michel@unistra.fr, thompson@unistra.fr

Running title: circular code motifs in the ribosome

Keywords: origin of life, genetic code, circular code, translation, ribosome evolution

Abstract

The origin of the genetic code remains enigmatic five decades after it was elucidated, although there is growing evidence that the code co-evolved progressively with the ribosome. A number of primordial codes were proposed as ancestors of the modern genetic code, including comma-free codes such as the *RRY, RNY* or *GNC* codes (*R* = G or A, *Y* = C or T, *N* = any nucleotide), and the *X* circular code, an error-correcting code that also allows identification and maintenance of the reading frame. It was demonstrated previously that motifs of the *X* circular code are significantly enriched in the protein-coding genes of most organisms, from bacteria to eukaryotes. Here, we show that imprints of this code also exist in the ribosomal RNA (rRNA). In a large-scale study involving 133 organisms representative of the three domains of life, we identified 32 universal *X* motifs that are conserved in the rRNA of >90% of the organisms. Intriguingly, most of the universal *X* motifs are located in rRNA regions involved in important ribosome functions, notably in the peptidyl transferase center and the decoding center that form the original 'proto-ribosome'. Building on the existing accretion models for ribosome evolution, we propose that error-correcting circular codes represented an important step in the emergence of the modern genetic code. Thus, circular codes would have allowed the simultaneous coding of amino acids and synchronization of the reading frame in primitive translation systems, prior to the emergence of more sophisticated start codon recognition and translation initiation mechanisms.

INTRODUCTION

Unraveling the emergence and evolution of the genetic code remains an elusive challenge (Koonin and Novozhilov 2017). It has been estimated that the events shaping the genetic code took place 3.7–4.1 billion years ago (Nutman et al. 2016) and led to the formation of the Last Universal Common Ancestor (LUCA) as a primordial ancestor of all life on Earth today. Since LUCA, the same standard genetic code has been used to translate nucleotides into amino acids in (quasi-) all organisms. The universality of the code is a hindrance with regard to studying its formation, since no organisms exist containing a primitive or intermediate genetic code for comparison. Nevertheless, different scenarios have been proposed that attempt to explain how the genetic code could have emerged from the primordial soup. Until recently, the text-book scenario has been an initial RNA world, in which RNA polymers acted both as a carrier of genetic information and as a catalyst for translation (Gilbert 1986). However, there is growing evidence supporting an early peptide-RNA world (e.g. Van der Gulik and Speijer 2015; Carter 2015; Bowman et al. 2015; Kunnev and Gospodinov 2018; Chatterjee and Jadav 2019), where the first RNA polymers coexisted and interacted with short peptides. Irrespective of these scenarios, a key question is how the modern standard genetic code came into being.

The contemporary genetic code represents a nearly universal assignment of 64 triplets of nucleotides (codons) to 20 amino acids. Many alternative hypotheses for the origins of this assignment have been put forward (reviewed in Grosjean and Westhof 2016; Koonin 2017). For example, the stereochemical hypothesis (Woese et al. 1966) postulates that the code developed from interactions between codons, anticodons and amino acids. The coevolution theory posits that the code coevolved with amino acid biosynthesis pathways, while the error minimization theory assumes that the adverse effect of point mutations and translation errors was the principal factor of the code's evolution. These theories are not mutually exclusive, and they may all have contributed to create the contemporary code. Initial amino acids may have

been defined by stereochemical affinities, but extension of such initial assignments *via* co-evolution and adaptation were probably essential to complete the modern coding table (Chatterjee and Yadav 2019).

All these theories are compatible with the idea that the universal genetic code gradually evolved from a simpler primordial form that encoded fewer amino acids, first postulated by (Crick et al. 1957). Crick's original proposal that the genetic code was a comma-free code explained how a sequence of trinucleotides could code for 20 amino acids, and at the same time how the correct reading frame could be retrieved and maintained. The main idea of comma-free codes is that coding trinucleotides are found only in one frame, known as the reading frame, i.e. trinucleotides in the reading frame make sense, while trinucleotides in the shifted frames 1 and 2 make nonsense. In coding theory, such a comma-free code is also known as a self-synchronizing code, since no external synchronization is required. It was later proved that the modern genetic code could not be a comma-free code (Nirenberg and Matthaei 1961), when it was discovered that TTT, a trinucleotide that cannot belong to a comma-free code, codes for phenylalanine. Although the standard genetic code used by nearly all modern organisms is not a comma-free code, other comma-free codes have been proposed that may have represented primeval codes, notably the *RRY* code (*R* = G or A, *Y* = C or T) with 8 trinucleotides and 4 amino acids (Crick et al. 1976), the *RNY* code (*N* = any nucleotide) with 16 trinucleotides and 8 amino acids (Eigen and Schuster 1978; Shepherd 1981), or the *GNC* code with 4 trinucleotides and 4 amino acids (Ikehara 2002).

A weaker version of comma-free codes, the so-called circular codes, has also been proposed (Arquès and Michel 1996). Circular codes are less restrictive than comma-free codes, as a frameshift of 1 or 2 nucleotides in a sequence entirely consisting of trinucleotides from a circular code will not be detected immediately but after the reading of a certain number of nucleotides (reviewed in Michel 2008; Fimmel and Strüngmann 2018). Circular codes possess

the circular property, i.e., any word written on a circle (the last letter becoming the first in the circle) has a unique decomposition into trinucleotides of the circular code (Fig. 1A). A circular code naturally excludes the homopolymer trinucleotides {AAA, CCC, GGG, TTT}. It also excludes trinucleotides related by circular permutation, e.g. AAC and ACA, since the concatenation of AAC with itself …AACAAC…, for example, can be decomposed in two ways: …AAC, AAC… or …A, ACA, AC… (Michel 2008). By excluding the homopolymer trinucleotides and dividing the 60 remaining trinucleotides into three disjoint classes, a circular code of trinucleotides has at most 20 trinucleotides (called a maximal circular code). There exist 12,964,440 maximal circular codes, although it has been shown that there is no maximal circular code that can code 20 or 19 amino acids and only 10 can code for 18 amino acids (Michel and Pirillo 2013). Remarkably, one of the maximal circular codes, called the *X* circular code (Fig. 1B), was found to be overrepresented in the reading frame of protein coding genes from eukaryotes and prokaryotes (Arquès and Michel 1996; Michel 2017). Other circular codes, and notably variations of the common *X* circular code, are hypothesized to exist in different organisms (Frey and Michel 2003, 2006; Ahmed et al. 2010; Michel 2015, 2017).

The *X* circular code has additional symmetry properties, in particular it is self-complementary, meaning that if a trinucleotide belongs to *X* then its complementary trinucleotide also belongs to *X* (Fig. 1C). Moreover, the +1 and +2/-1 circular permutations of *X*, denoted X_1 and X_2 respectively, are also maximal circular codes and are complementary to each other (Fig. 1D). The class of circular codes, like comma-free codes, also have the property of synchronizability, i.e. they have the ability to retrieve the correct reading frame by using an appropriate window of nucleotides. In any sequence generated by a trinucleotide comma-free code, the reading frame can be determined in a window length of at most 3 nucleotides, while for the *X* circular code, at most 13 consecutive nucleotides are enough to always retrieve the

reading frame (Fig. 1E). In other words, any sequence 'motif' containing 4 consecutive *X* trinucleotides is sufficient to determine the correct reading frame.

The hypothesis of circular codes, and in particular the *X* circular code, is supported by evidence from several statistical analyses of modern genomes. For example, it was shown in a large scale study of 138 eukaryotic genomes (El Soufi and Michel 2016) that *X* motifs (in the case of protein-coding genes, an *X* motif was defined as a run of at least 4 trinucleotides from the *X* circular code) are found preferentially in protein-coding genes compared to non-coding regions with a ratio of ~8 times more *X* motifs located in genes. More detailed studies of the complete gene sets of yeast and mammal genomes (Michel et al. 2017; Dila et al. 2019) confirmed the strong enrichment of *X* motifs in genes and further demonstrated a statistically significant enrichment in the reading frame compared to frames 1 and 2 (p-value$<10^{-10}$). In addition, it was shown that most of the mRNA sequences from these organisms (e.g. 98% of experimentally verified genes in *S. cerevisiae*) contain *X* motifs. Intriguingly, conserved *X* motifs have also been found in many tRNA genes (Michel 2013), as well as near the decoding center of 16S/18S ribosomal RNA from bacteria, archaea and eukaryotes (El Soufi and Michel 2015), which suggest their involvement in universal gene translation mechanisms.

Here, we investigate whether the overrepresentation of *X* motifs in genes might reflect traces of an ancestral coding system based on circular codes, one that used a smaller number of trinucleotides than the modern genetic code but that had the specific capacity to identify or maintain the reading frame. If the *X* circular code represents a predecessor of the genetic code, then we should be able to find imprints or traces of the code in the evolution of the translation machinery, and in particular in the ribosome, a highly conserved ribonucleoprotein complex.

Since the ribosome is universal in all extant organisms (Melnikov et al. 2012), it can be deduced that it was largely formed at the time of the LUCA, and its earliest origins likely lie in the prebiotic world. It is widely accepted that in the primordial soup, increased chemical

complexity led to RNA or RNA-like oligomers. Interactions between these RNA conformations and prebiotic amino acids or short oligopeptides could have stabilized the structures and provided catalytic functions (Szathmáry 1999; Plankensteiner et al. 2005; Van der Gulik and Speijer 2015). Several mechanisms establishing correspondence between anticodons/codons and their cognate amino acids have been suggested, possibly representing a 'proto-translation machine' (Yarus et al. 2009; Ma 2010; Noller 2012; Carter 2016). Thus, an early ribosome may have consisted of rRNAs stabilized by a few small peptides containing glycine, alanine, aspartic acid and/or valine, essential for the structure of the nucleoprotein particle (Fournier et al. 2010; Maier et al. 2013). According to this theory, RNA and protein-based molecules would then have evolved concurrently and interactively, giving rise to the first system capable of translating genetic information (Kunnev and Gospodinov 2018) and self-replicating (Banwell et al. 2018). Thus, the original translation machinery would have been RNA-based, and this RNA translation template would have evolved to form the tRNA, mRNA and rRNA established at the time of the LUCA (Chatterjee and Jadav 2019; Root-Bernstein and Root-Bernstein 2019). Most likely the initial specificity of translation would have been very low. The question remains of how such a system could have evolved to a more specific mapping between the genetic sequence and the peptide sequence, either by direct rRNA/amino acid interactions or indirectly via tRNA, in order to produce longer peptides that could fold into the first functional proteins (Lupas and Alva 2017). The coevolution theory suggests the idea of a growing coding repertoire interacting with a simultaneously growing repertoire of biosynthetic products. Although it is impossible to recreate the entire path along which the very complex process of translation evolved, it is possible to propose, and provide supporting evidence for, certain theoretical solutions.

To test our hypothesis that the *X* circular code represents an intermediate coding system between the primordial, non-specific RNA-peptide interactions and the modern ribosome-based translation machinery (Fig. 2), we performed a large-scale study of extant rRNA sequences

from 133 representative organisms covering the three domains of life, in order to identify *X* motifs that have been conserved since the LUCA. In a comprehensive analysis of ribosome structural data, we show that most of these universally conserved *X* motifs, denoted *uX* motifs, are located in important functional sites including the decoding center and the peptidyl transferase center (PTC). Furthermore, these functional sites are widely accepted to be essential building blocks of the primeval 'proto-ribosome' that was already present in the LUCA (Smith et al. 2008; Bokov and Steinberg 2009; Hsiao et al. 2009, 2013; Petrov et al. 2015; Agmon 2017; Agmon 2018). Building on the previously described accretion models of ribosome growth (Hsiao et al. 2009; Petrov et al. 2015), we propose that error-correcting circular codes represent an important step in the co-evolution of the genetic code and the ribosome, where a single code allowed the simultaneous coding of amino acids and synchronization of the reading frame. To our knowledge, this is the first study to propose an ancestral mechanism for reading frame maintenance, prior to the emergence of more sophisticated start codon recognition and translation initiation systems.

RESULTS AND DISCUSSION

Universal *X* motifs in rRNA of extant organisms

Modern ribosomes are highly sophisticated molecular machines, consisting of two subunits that come together during the initiation of protein synthesis, remain together as individual amino acids are added to a growing peptide according to information encoded on the mRNA, and finally separate again in conjunction with the release of the finished protein. Each subunit is a large nucleoprotein complex. In bacteria and archaea, the large subunit (LSU) contains a 23S rRNA and a 5S rRNA, whereas the small subunit (SSU) contains the 16S rRNA. In eukaryotes, the LSU contains a 28S rRNA, a 5S rRNA and a 5.8S rRNA, whereas the SSU

contains the 18S rRNA. By comparing 3D ribosome structures from different organisms, a common core of rRNA was identified that is conserved over the entire phylogenetic tree, especially in terms of secondary/tertiary structures (Hsiao et al. 2009; Petrov et al. 2015; Opron and Burton 2018).

To investigate the presence of *X* motifs, i.e. motifs composed of trinucleotides from the circular code *X*, in this common core of rRNA, we identified universal *X* motifs (denoted *uX* motifs) in multiple sequence alignments of the LSU rRNAs (23S/28S and 5S) and SSU rRNAs (16S/18S) for 133 representative species covering all three domains of life. *X* motifs are defined as universal (denoted *uX* motifs) if they are present in at least 90% of the aligned sequences and have a length of at least 6 consecutive nucleotides. It is important to note that *uX* motifs are not necessarily conserved in terms of the nucleotide sequence. An example is the SSU trinucleotide 1505-1507, which is highly conserved in bacteria and archaea as GUA and conserved in eukaryotes as GUU, thus affecting the sequence conservation but not the universality of the *X* trinucleotide. In the SSU, 13 *uX* motifs were present in more than 90% of the sequences (Table 1 and Fig. 3A), in the LSU 19 *uX* motifs were identified (Table 2 and Figure 3B), while no *uX* motifs were found in the 5S alignment. The *uX* motifs are labeled according to the accretion model of Petrov et al. 2015, and using capital letters for LSU motifs and small letters for SSU motifs (see below). The mean sequence conservation across the full length of the SSU and LSU is 65% and 62% respectively, while the *uX* motifs are 81% conserved. A more detailed comparison of nucleotide sequence conservation and the universality of *uX* motifs is provided in Supplemental Fig. S1 and Supplemental Table S1. Within the *uX* motifs, no significant correlation (Pearson correlation coefficient: $p<10^{-4}$, Spearman correlation coefficient: $p=0.006$, Kendall coefficient $p=0.007$) was observed between the *X* universality and the sequence conservation (Supplemental Tables S2, S3). In fact, >28% of the rRNA alignments covered by *uX* motifs are not conserved in terms of sequence (Supplemental Table S1). Taken

together, these results suggest that in certain regions of the ribosome, the *X* circular code property exists in addition to sequence level constraints in the ribosome.

The overall coverage of nucleotides in *uX* motifs in the SSU and LSU rRNAs are similar (7.8% and 6.0% respectively), however coverage is not homogeneous across the different structural domains of both subunits (Table 3 and Fig. 4). It is interesting to note that the SSU 3'm domain, containing the central pseudoknot (CPK) and the decoding center, has the highest coverage with 19% of nucleotides in *uX* motifs. The SSU 3'M domain corresponding to the 'head' region and the LSU V domain containing the PTC are also enriched with ~12% coverage, in contrast to the SSU central domain and the LSU 0, I, III and VI domains which have only ~3% coverage. In order to evaluate the significance of the observed coverage, we chose an approach that involved comparing the results obtained for the *uX* motifs with those obtained for universal random motifs (*uR* motifs) generated by random sampling of 100 different codes *R* with properties similar to the *X* code, except for the circularity property (defined in detail in Materials and Methods). The distributions of the number and total length of *uR* motifs (Fig. 5) thus provide an estimate of the expected values for the *uX* motifs. As shown in Supplemental Fig. S2, the observed number of *uX* motifs in the SSU (13) and in the LSU (19) are significantly higher than expected (mean values for *uR* motifs are 10 and 13 respectively). We also determined how many of the *uR* motifs display the same level of occurrence and coverage as the *uX* motifs (Fig. 5). None of the *R* codes had the same number of observed uX motifs (=32), while 2% of the *R* codes had the same number of motifs. 3% of the *R* codes had a longer total length than the uX motifs. These findings reveal an overrepresentation of *uX* motifs in the LSU (23S/28S) and SSU rRNAs (16S/18S) conserved in the three domains of life.

We then asked whether this overrepresentation might be linked to a compositional bias of the rRNA sequences. In terms of nucleotide composition, some bias is observed in the rRNA sequences (Supplemental Table S4) where G is the most frequent (31.1%) and T is the least

frequent (20.5%). However, the *X* circular code shows no bias with equal frequencies of the four bases A, C, G and T (Supplemental Table S4), and therefore the nucleotide bias cannot explain the observed enrichment. The nucleotide composition of the 13 *uX* motifs in the SSU and the 19 *uX* motifs in the LSU are provided in Supplemental Tables S5, S6. Concerning the trinucleotide composition of the rRNA sequences (Supplemental Table S4), no significant enrichment of *X* trinucleotides is observed, according to a Mann-Whitney U Test (z-score=-0.51419). We conclude that the enrichment concerns *X* trinucleotides located within motifs specifically. The trinucleotide composition of the 13 *uX* motifs in the SSU and the 19 *uX* motifs in the LSU are provided in Supplemental Tables S7, S8.

Finally, we investigated whether the observed enrichment of uX motifs might be associated with the fact that rRNA sequences co-vary in order to preserve their 3D structure. To do this, we used an Infernal covariance model (CM), a probabilistic model which captures many important features of structured RNA sequence variation (Nawrocki and Eddy 2013). We constructed two CMs for each ribosomal subunit, one where each position in the sequences is treated independently, and one where base paired positions are dependent on one another. However, no significant difference was observed between the two CMs (data not shown), and we conclude that the co-variation constraints in the rRNA do not impose an enrichment of *uX* motifs.

uX motifs map to functional centers of modern ribosomes

In this section, we investigate the location of the 32 *uX* motifs identified in modern ribosomes and how they relate to known functional regions. Although some variation exists, modern translation mechanisms are generally similar in archaeal, bacterial, and eukaryotic systems and the main functions of the ribosome are conserved in the three domains of life (Opron and Burton 2018). The SSU binds messenger RNA (mRNA) and, together with the

transfer RNA (tRNA), is responsible for translational fidelity by ensuring base pairing between the codon and anticodon in the decoding center. The LSU binds the acceptor ends of the A-site and P-site tRNAs and catalyzes peptide bond formation at the peptidyl transferase center (PTC). As the nascent protein is synthesized it passes through an exit tunnel that begins at the PTC and exits from the back of the LSU. Both subunits are actively involved in translocating the mRNA by one trinucleotide in each cycle, and conformational dynamics are crucial (Jenner et al. 2010; Belardinelli et al. 2016). Large-scale rearrangements include rotation of the SSU and LSU relative to one another (also known as ratcheting), swiveling of the SSU head in relation to the body, and stepwise translocation of the tRNAs together with the mRNA through the ribosome.

We based our study on a representative 3D structure of the ribosome from the bacteria *T. thermophilus,* since it contains mRNA nucleotides and three deacylated tRNAs in the A, P and E sites. Fig. 6 shows the positions of the 19 *uX* motifs in the LSU rRNA (Fig. 6A) and the 13 *uX* motifs in the SSU rRNA (Fig. 6B) and Tables 4,5 summarize the interactions of *uX* motifs with different molecules, including mRNA, tRNA and ribosomal proteins.

In the LSU, the most conserved functional site is the PTC, where amino acids are polymerized onto the growing nascent chain. The majority of the *uX* motifs are clustered around the PTC (Fig. 6C) with 3 motifs within a radius of 10 Å (*B, D, F*), 6 motifs within a radius of 30 Å (*B, C, D, E, F, P*) and 13 out of the 19 motifs within a radius of 50 Å (*A, B, C, D, E, F, G, H, I, L, K, M, P*). Thus, 105 (60%) of the 175 nucleotides covered by *uX* motifs are found within 50 Å of the PTC. Several *uX* motifs are in direct contact with tRNA: nucleotides G2553, U2555 (motif *F*) and G2583, U2585 (motif *D*) are in contact with the A-site tRNA; U2585 (motif *D*) and U2506 (motif *B*) are in contact with the P-site tRNA; and G1850-A1853 (motif *N*) are in contact with the E-site tRNA. One motif (*A*) is found in helix H89, which is known to be involved in the accommodation of the A-site tRNA in the PTC (Jenner et al. 2010). Another important structure in the LSU is the polypeptide exit tunnel that extends from the PTC to the

surface of the ribosome. The tunnel shape is more conserved in the upper part close to the PTC, while in the lower part, it is substantially narrower in eukaryotes than in bacteria (Dao Duc et al. 2019). Fig. 6D shows the eight *uX* motifs that are close to the exit tunnel: (*B, D, E, F, H, G, L, S*). Finally, two *uX* motifs are found in regions involved in interactions with GTPase proteins during translation initiation and elongation: motif *Q* is in the GTP Associated Center (GAC) and motif *O* is in the sarcin-ricin loop. The four remaining *uX* motifs (*I, J, M, R*) in the LSU are not associated with known functions to our knowledge.

In the SSU, 7 of the 13 *uX* motifs (*a, b, c, d, e, h, i*) are in contact with the mRNA (at a distance of <5 Å) (Fig 6E). Remarkably, only 3 of the 25 rRNA nucleotides in contact with the mRNA are not found in *uX* motifs. The *uX* motifs also include many of the rRNA contacts with tRNAs, such as the A-site conserved nucleotides A1492-A1493 (motif *b*) and G530 (motif *h*); the P-site G926 (motif *d*), A790 (motif *e*), U1498 (motif *b*), and C1400 (motif *a*); and the E-site C795 (motif *e*) (Khade and Joseph 2010).

An important feature of the SSU is the dynamic swiveling of the SSU head (3'M domain) relative to the body (5' domain) during translation elongation. The movement originates from flexing at two hinge points, one in the middle of helix h28 at G926, and one in the linker between h34 and h35. Both of these hinges are found in *uX* motifs (*d, l* respectively). Rotation of the SSU head has also been linked to the opening and closing of a 13 Å constriction or 'gate' between the head and body domains between the P and E sites, presenting a steric block to the movement of the P-site tRNA. The gate involves G1338 (motif *j*) situated in the stable ridge that sterically separates the P and E sites, and A790 (motif *e*) located on the opposite side of the constriction (Achenbach and Nierhaus 2015). The C1397 (motif *a*) and A1503 (motif *c*) have also been considered to be 'ratchet pawls' that intercalate with mRNA bases during reverse rotation of the head (Achenbach and Nierhaus 2015). Three *uX* motifs (*f, g, k*) in the SSU are not associated with known functions to our knowledge.

Many of the *uX* motifs identified in this study are also in contact with ribosomal proteins (11 out of 13 *uX* motifs in the SSU and 16 out of 19 *uX* motifs in the LSU). Among the 102 known ribosomal protein families, 34 (15 in the SSU, 19 in the LSU) are represented in all three domains of life (Smith et al. 2008 and Supplemental Table S9). Many of these universal proteins have been shown to be crucial for ribosome assembly, formation of inter-subunit bridges, and interactions with the tRNAs or the polypeptide exit channel (Lecompte et al. 2002). Interestingly, nearly all the proteins in contact with *uX* motifs are universal ribosomal proteins (in *T. thermophilus*, all 10 proteins in contact with the SSU *uX* motifs are universal, and 10 out of 14 proteins in contact with the LSU *uX* motifs are universal).

uX motifs were present in the primordial proto-ribosome

It is generally assumed that the large and small subunits of the ribosome initially existed independently, although there is some debate as to whether the LSU or the SSU emerged first (Kunnev and Gospodinov 2018; Opron and Burton 2018). Based on comparative structural analyses, proto-LSU (Smith et al. 2008; Bokov and Steinberg 2009; Hsiao et al. 2009, 2013; Petrov et al. 2015; Agmon 2017) and proto-SSU (Petrov et al. 2015; Agmon 2018) models have been proposed (Fig. 7).

The proto-LSU corresponds to the PTC, a symmetrical region deep within the large rRNA, where new amino acids are incorporated into the growing peptide chain (Agmon 2009). This region has generally been modeled using the contemporary *E. coli* sequence to represent the ancestral system (Fig. 7). It consists of approximately 120 nucleotides, forming a pocket-like structure that could have accommodated two random amino acids, and would have provided positional catalysis, producing short peptides with random composition. We mapped the *uX* motifs to the 2D model and found a total of 40 nucleotides (30%) in *uX* motifs. The motifs are almost exclusively located in the A-monomer corresponding to the modern A-tRNA site, with

35 (58%) of the 60 A-monomer nucleotides in *uX* motifs. In addition to the universal regions, many of the nucleotides that constitute the two halves of the PTC cavity are composed of *X* trinucleotides and these trinucleotides have been shown to have a high level of complementarity in different ancient bacteria (Agmon 2017), reflecting the self-complementary property of the *X* circular code. This complementarity has been suggested to indicate a simple and efficient mode of replication, i.e. the proto-LSU may have been a self-replicating ribozyme (Agmon 2017).

The ancestor of the SSU is more controversial, but it may have worked simply as a location to bind RNAs in an open structure configuration (de Farias et al. 2019). The proposed models correspond to the contemporary central pseudoknot (CPK) in the decoding center (Noller 2012). However, in contrast to what is observed in the LSU, there is no single self-folding segment in the modern 16S RNA that encompasses the majority of the decoding site rRNA. A number of disjoint short segments of total length of about 150 nucleotides have been considered ancestral (Petrov et al. 2015; Agmon 2018). Of these, 40 nucleotides (27%) are found in *uX* motifs, notably including the future A-site (A1492-A1493) and P-site (C1402-C1403, U1498-A1499) tRNA binding sites.

It is worth noting that the combined models of the proto-ribosome, incorporating the active sites of both ribosomal subunits, cover less than 6% of the modern prokaryotic rRNA, yet they integrate 80 (27%) of the 296 rRNA nucleotides found in *uX* motifs.

Accretion of *uX* motifs in the transition from the proto-ribosome to the modern ribosome

Given the complexity of the modern ribosome, it is unlikely that it appeared spontaneously (Hsiao et al. 2009; Petrov et al. 2015; Opron and Burton 2018). According to the RNA-peptide world theory, RNA and protein-based molecules would have evolved concurrently and interactively, giving rise to the first system capable of translating genetic information (Kunnev

and Gospodinov 2018) and self-replicating (Banwell et al. 2018). For example, Petrov et al. 2015 suggested that the proto-ribosome evolved to the modern rRNA core by recursive accumulation of ancestral expansion segments (AES) and proposed an accretion model of rRNA evolution divided into six major phases representing successive steps in the complexification of the ribosome. Fig. 4 shows the location of the *uX* motifs with respect to this accretion model, where the *uX* motifs are labeled (*a-m* for the SSU *uX* motifs and *A-S* for the LSU *uX* motifs) according to their presumed ancestry. We can differentiate two subsets of the *uX* motifs: those already present in the proto-ribosome described above (phases 1 and 2 of the accretion model) and those gained in the subsequent phases of ribosome evolution (phases 3-6). Thus, 4 motifs (*B-E*) of the 19 *uX* motifs were already present in the proto-LSU, 2 additional motifs (*A,F*) are located close to the slightly extended ancestral region defined by Petrov et al. 2015, and 4 motifs (*a-d*) of the 13 *uX* motifs were present in the proto-SSU. In phase 3, *uX* motifs *G-L* are incorporated near the extended exit tunnel and motifs *K,M* in the LSU–SSU interface. In phase 4, motif *e* is included in the SSU-LSU interactions, and motifs *f,g* in the A-site and P-site tRNA binding pockets respectively. In phase 5, motif *O* is incorporated near the binding sites for elongation factors G and Tu, and motifs *P,Q* in the L11 stalk. In the SSU, motifs *i,j* are included in the P-site tRNA pocket and motif *h* in the central pseudoknot. In phase 6, the remaining motifs *R-S, k-m* are introduced in AES that serve mainly as binding sites for the globular domains of ribosomal proteins.

The universal ribosomal proteins mentioned above have also been incorporated into this accretion model (Kovacs et al. 2017), based on the assumption that the age of a given segment of protein is the same as that of the rRNA with which it interacts. In phases 1 and 2, it is assumed that only short random peptides are present in the proto-ribosome system. In phases 3 and 4, *uX* motifs (*A-M, a-g*) interact with 7 of the 19 universal proteins in the LSU (Table 5) and 7 of the 15 universal proteins in the SSU (Table 4). Many of these proteins are known to interact

with the PTC (L2, L3, L4, L14) or have contacts to the tRNA binding site and/or the mRNA (S7, S9, S11, S12) mainly *via* their non-globular extensions (Smith et al. 2008). In phase 5, *uX* motifs (*O-Q, h-j*) contact globular domain proteins, including L6, L13, L36, and S3. In phase 6, most of the newly incorporated proteins are on the surface of the ribosome, and the *uX* motifs (*R-S, k-m*) contact only a few of them: L23, S2 and S17.

Model of coevolution of genetic code and translation system

Based on our analyses of *uX* motifs in the proto-ribosome and the accretion model of ribosome evolution, we suggest that comma-free codes and circular codes represented ancestors of the modern genetic code and were used to map the first trinucleotides to amino acids. We thus propose a model for the coevolution of the genetic code and the translation system in four stages, shown in Fig. 8 and discussed in the following paragraphs.

Recent evidence suggests that RNA and peptides co-evolved from the beginning, or at least that the proto-ribosome building blocks gained the ability to bind amino acids or small peptides very early (Lupas and Alva 2017; Kunnev and Gospodinov 2018). The first peptides were most probably of abiotic origin, most likely including glycine and alanine, and binding would have been non-specific. However, natural selection would soon have favored forms encoded and synthesized by nucleic acids. We propose that the first encoding system was based on a comma-free code, such as {GGC, GCC}, which would have allowed encoding of the amino acids and the reading frame within a single code. At this time, the LSU and SSU would have evolved separately, with the proto-LSU having a PTC function and the proto-SSU binding proto-mRNA.

Assembly of the two subunits with the intermediate tRNA would have given rise to the first ribosomes capable of coding longer and more specific peptides. From this time, the ribosome and genetic code would have co-evolved (Vitas and Dobovišek 2018). With the addition of new

amino acids, comma-free codes were no longer viable and the genetic code would have evolved towards the circular codes, possibly with a smaller number of amino acids initially. For example, we have shown previously (Michel et al. 2017) that an *X'* circular code exists with 10 trinucleotides capable of coding 8 of the 10 hypothesized 'early amino acids' (Koonin 2017). Only two universally conserved motifs from this *X'* circular code can be observed in the modern ribosome, at positions 1396-1404 in the SSU and 2500-2511 in the LSU. It is interesting to note that these two 'primitive' circular code motifs correspond to the *X* motifs a and B (in the SSU and LSU respectively), which are predicted to be the earliest *X* motifs in the ribosome according to the accretion model. The peptides synthesized by the early ribosomes may have functioned as primordial ribosome co-factors, possibly to increase rRNA stability (Lupas and Alva 2017).

At the early/intermediate stages, in addition to their function of amino acid assignment, circular codes would have allowed reading frame detection and/or maintenance before the emergence of complex start codon recognition systems, allowing to code the first simple proteins. The *X* circular code may thus have been the first error detection/correction system, avoiding reading the mRNA in the wrong frame.

Finally, no circular codes can include more than 20 trinucleotides, so the circular code property was not sufficient when more amino acids were needed. The standard genetic code requires a specific start codon that initiates translation, and sophisticated ratchet mechanisms for maintaining the reading frame during translation elongation. Intriguingly, *uX* motifs are found in modern ribosomes in many of the ratchet pawls, as well as in the PTC and the decoding center.

Conclusion

The genetic code is too complex to have emerged spontaneously and it is hypothesized that the coding process started with a set of primitive amino acids and that others were added until

the total of 20 was reached (Chatterjee and Yadav 2019). Most studies of the origin and evolution of the genetic code have focused on the mapping between codons and amino acids (e.g. Ikehara 2002; Hartman and Smith 2014; Koonin 2017), and the origin of reading frame maintenance has not been addressed before. Here, we have investigated the hypothesis that the contemporary genetic code arose from simpler comma-free codes *via* circular codes. In addition to encoding the amino acids, comma-free codes and circular codes present the important synchronization property that would have allowed detection and maintenance of the reading frame in primordial and less sophisticated translation systems. Should our hypothesis be true, the contemporary translation system may still contain vestiges of such codes. To test this, we used the X circular code as it has the most "universal" occurrence in genes and also strong mathematical properties, in particular it is self-complementary and C^3. We compared rRNA sequences from the three domains of life and identified 32 motifs from the X circular code that are universal, even though they occur in sequences that are not conserved in terms of nucleotides. The enrichment of the rRNA in *uX* motifs is statistically significant and most of the motifs are clustered around important functional sites, including the PTC and the exit tunnel in the LSU and the decoding center and ratchet mechanisms in the SSU. We propose that they represent the observable remnants of a primordial code used during the emergence of the RNA- or RNA-peptide world.

The emergence of the translation system is a chicken-and-egg problem: the ribosome is needed to code proteins, but the ribosome needs proteins to function. It has been suggested that an RNA molecule with a peptidyl transferase activity existed before the full sequential three-base decoding (Polacek and Mankin 2005). This early non-coded proto-ribosome could have catalyzed the association of arbitrary amino acids, producing short peptides of random sequences. Here, we showed that the models of both proto-LSU and proto-SSU are enriched in *uX* motifs, with 30% of the nucleotides found in *uX* motifs. Concerning the LSU, we observed

more *uX* motifs in the A-monomer than in the P-monomer, based on the *E. coli* sequence that was used in the model (Fig. 7). This may reflect an inherent asymmetry of the proto-LSU, or it may be due to a stronger conservation of the A-site in evolution.

In the RNA-peptide world scenario, the RNA polymers of the proto-ribosome served as templates to directly bind amino acids or short peptides. Cognate RNA triplets could have then evolved to act as anticodons in tRNAs and codons in mRNAs (Yarus 2017). It has been observed previously that the early prebiotic amino acids are coded by G/C-rich codons, whereas engagement of new amino acids required more of A and U to be included in the codons (Polyansky et al. 2013). We propose here that the comma-free code {GGC, GCC} was used initially to code Ala and Gly, and that this code quickly expanded to an ancestral circular code, such as the *X'* circular code containing 10 codons with a composition of 66% G/C and 33% A/T, and coding for 8 out of the 10 identified early amino acids (Koonin 2017).

The increase of the amino acid repertoire and the transition from the production of random peptides to the coding of specific protein sequences require more sophisticated mechanisms for codon recognition, but also identification of the reading frame. Circular codes represent an efficient means to synchronize the reading frame within a short window, before the evolution of a start codon and the modern translation initiation system. In support of this hypothesis, here we have identified *uX* motifs in the early rRNA. *X* motifs have also been discovered in modern mRNA sequences (Michel et al. 2017; Dila et al. 2019), as well as in many tRNA (Michel 2013). It is therefore tempting to suggest that base-pairing between the *X* motifs of the mRNA and those of the tRNA and the rRNA would have given rise to the first coded ribosome apparatus. Traces of such interactions remain in the 3D structures of modern ribosomes, where we have shown that most of the *uX* motifs in the rRNA are in contact with the mRNA or the A, P and E-site tRNAs.

Universally conserved *X* circular code motifs are present at each evolutionary stage up to the common core of the modern ribosome and are coherent with the proposed model for coevolution of the genetic code and the translation system. The question of whether the *X* motifs retain a function in modern translation systems, possibly by participating in reading frame retrieval, can only be answered by experimental studies.

Materials and Methods

Ribosomal RNA multiple sequence alignments

Multiple sequence alignments for LSU rRNAs (23S/28S and 5S) and SSU rRNAs (16S/18S) were obtained from the Center for Ribosomal Origins and Evolution's RiboVision web server at http://apollo.chemistry.gatech.edu/RibosomeGallery/Read Me/alignments/index.html. The alignments contain complete sequences for rRNAs from 133 distinct species, representing a broad but sparse sampling of the phylogenetic tree of life, including all three domains of life. The sequences for the 32 eukaryotes, 65 bacteria, and 36 archaea were originally extracted from the SILVA database at https://www.arb-silva.de. A list of the organisms present in the alignments is provided in Supplementary Table S10.

Identification of universal *X* motifs (*uX* motifs) in rRNA alignments

The trinucleotide set *X* is a maximal C^3 self-complementary circular code (Arquès and Michel 1996). A circular code is a set of words over an alphabet such that any sequence written on a circle has a unique decomposition (factorization) into words of the circular code. Any motif from the circular code *X*, called *X* motif, has the ability to retrieve the reading frame of the sequence. Formal and classical definitions related to circular codes that are not explicitly necessary to understand the results obtained in this work, are not recalled here. They are

available in (Arquès and Michel 1996; Michel 2008; Fimmel et al. 2016; Michel et al. 2017; Fimmel and Strüngmann 2018; Dila et al. 2019).

As in Michel et al. 2017, an *X* motif is defined as a consecutive sequence of trinucleotides from the *X* circular code. For each rRNA sequence in the above alignments, the *X* motifs were localized using a program developed in the Java language (El Soufi and Michel 2016). The program takes optional parameters that define the minimum length *l* (in nucleotides) of the *X* motifs searched. As in previous work, we used $l \geq 8$ nucleotides (i.e. at least 2 trinucleotides, and either prefixes or suffixes of trinucleotides) which implies that the reading frame can be retrieved with a probability of 99.6% (Michel 2012).

For each position in each of the LSU rRNA (23S/28S and 5S) and SSU rRNA (16S/18S) alignments, we then calculated the 'universality' of the *X* motifs, defined as the number of sequences having an *X* motif at that position. A universal *X* motif (denoted *uX* motif) was defined as a region in the alignment with two constraints: at least 6 consecutive positions and >=90% *X* universality (i.e. positions covered by *X* motifs in >=90% of the sequences in the alignment).

It is important to note that, in the case of the rRNA, since the notion of 'reading frame' is not relevant, we searched for X motifs starting at any position in the sequences. Thus, the trinucleotides of the X motifs in the different organisms are not necessarily in the same 'frames'. For example, one of the uX motifs in the SSU cover the sequences AG,GTA,ACC in *E. coli* and A,GGT,TTC,G in *H. sapiens*.

Identification of universal random motifs (*uR* motifs) in rRNA alignments

To evaluate the statistical significance of both the occurrence number and the nucleotide length of the *uX* motifs identified in the rRNA alignments, we generated 100 'random' codes. The random codes represent a purposive sampling of extreme cases, and were designed to have

similar properties to the *X* circular code except its circularity, as described in Michel et al. 2017. Thus, a random code *R* has 20 trinucleotides; the total number of each nucleotide A, C, G and T in *R* is 15; *R* has no stop codons and no periodic trinucleotides {AAA, CCC, GGG, TTT}. Motifs from each of the 100 random codes were identified in each rRNA alignment and their universality was calculated as for *X* motifs. Thus, we defined a universal *R* motif (denoted *uR* motif) as a region in the alignment with at least 6 consecutive positions and >=90% *R* universality.

To estimate the expected enrichment of *uX* motifs, we calculated the ±0.99 confidence levels for the mean values of the *uR* motifs. We then used a one-sided Student's t-test to evaluate whether the observed number and length of *uX* motifs were significantly higher than expected for random *uR* motifs.

Secondary structures

The secondary structures of LSU and SSU rRNAs for *E. coli* were downloaded from http://apollo.chemistry.gatech.edu/RibosomeGallery/. Mapping of information on to secondary structures was performed with RiboVision (apollo.chemistry.gatech.edu/RiboVision) (Bernier et al. 2014). Positions of the expansion segments for LSU and SSU rRNAs and phases in the accretion model were obtained from (Petrov et al. 2015).

Three-dimensional structures

Coordinates of the high-resolution crystal structure of the *T. thermophilus* ribosome were obtained from the PDB database (https://www.rcsb.org/). The PDB entry 4W2F was chosen because it contains mRNA nucleotides, an antibiotic (amicoumacin A) and three deacylated tRNAs in the A, P and E sites. Numbering of the *T. thermophilus* SSU rRNA is the same as for *E. coli*. For the LSU rRNA, *E. coli* numbering is used.

Visualization and analysis of the three-dimensional structures, as well as image preparation were performed with PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC).

Acknowledgements

This work was supported by Institute funds from the French Centre National de la Recherche Scientifique, and the University of Strasbourg. The authors would like to thank the BISTRO and BICS Bioinformatics Platforms for their assistance.

References

Achenbach J, Nierhaus KH. 2015. The mechanics of ribosomal translocation. *Biochimie* **114:** 80-89.

Agmon I. 2009. The dimeric proto-ribosome: Structural details and possible implications on the origin of life. *Int. J. Mol. Sci.* **10:** 2921-2934.

Agmon I. 2017. Sequence complementarity at the ribosomal Peptidyl Transferase Centre implies self-replicating origin. *FEBS Letters* **591:** 3252–3258.

Agmon I. 2018. Hypothesis: Spontaneous Advent of the Prebiotic Translation System via the Accumulation of L-Shaped RNA Elements. *Int J Mol Sci* **19:** E4021.

Ahmed A, Frey G, Michel CJ. 2010. Essential molecular functions associated with the circular code evolution. *J Theor Biol* **264:** 613–622.

Arquès DG, Michel CJ. 1996. A complementary circular code in the protein coding genes. *J Theor Biol* **182:** 45–58.

Banwell EF, Piette BMAG, Taormina A, Heddle JG. 2018. Reciprocal Nucleopeptides as the Ancestral Darwinian Self-Replicator. *Mol Biol Evol* **35:** 404–416.

Belardinelli R, Sharma H, Caliskan N, Cunha CE, Peske F, Wintermeyer W, Rodnina MV. 2016. Choreography of molecular movements during ribosome progression along mRNA. *Nat Struct Mol Biol* **23:** 342–348.

Bernier CR, Petrov AS, Waterbury CC, Jett J, Li F, Freil LE, Xiong X, Wang L, Migliozzi BL, Hershkovits E, et al. 2014. RiboVision: Visualization and analysis of ribosomes. *Faraday Discuss* **169:** 195–207.

Bokov K, Steinberg SV. 2009. A hierarchical model for evolution of 23S ribosomal RNA. *Nature* **457:** 977–980.

Bowman JC, Hud NV, Williams LD. 2015. The ribosome challenge to the RNA world. *J Mol Evol* **80:** 143–161.

Carter CW Jr. 2015. What RNA world? Why a peptide/RNA partnership merits renewed experimental attention. *Life* **5:** 294–320.

Carter CW Jr. 2016. An alternative to the RNA world. *Nat Hist* **125:** 28–33.

Chatterjee S, Yadav S. 2019. The Origin of Prebiotic Information System in the Peptide/RNA World: A Simulation Model of the Evolution of Translation and the Genetic Code. *Life (Basel)* **9:** E25.

Crick FH, Griffith JS, Orgel LE. 1957. Codes without commas. *Proc Natl Acad Sci U S A* **43:** 416–421.

Crick FH, Brenner S, Klug A, Pieczenik G. 1976. A speculation on the origin of protein synthesis. *Origin of Life* **7:** 389–397.

Dao Duc K, Batra SS, Bhattacharya N, Cate JHD, Song YS. 2019. Differences in the path to exit the ribosome across the three domains of life. *Nucleic Acids Res* gkz106.

De Farias ST, Rêgo TG, José MV. 2019. Origin of the 16S Ribosomal Molecule from Ancestor tRNAs. *Sci* **1:** 8.

Dila G, Michel CJ, Poch O, Ripp R, Thompson JD. 2019. Evolutionary conservation and functional implications of circular code motifs in eukaryotic genomes. *Biosystems* **175:** 57–74.

Eigen M, Schuster P. 1978. The hypercycle. A principle of natural self-organization. Part C: the realistic hypercycle. *Naturwissenschaften* **65:** 341–369.

El Soufi K, Michel CJ. 2015. Circular code motifs near the ribosome decoding center. *Comput Biol Chem* **59:** 158–176.

El Soufi K, Michel CJ. 2016. Circular code motifs in genomes of eukaryotes. *J Theor Biol* **408:** 198–212.

Fimmel E, Michel CJ, Strüngmann L. 2016. n-Nucleotide circular codes in graph theory. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374:** 20150058.Fimmel E, Strüngmann L. 2018. Mathematical fundamentals for the noise immunity of the genetic code. *Biosystems* **164:** 186–198.

Fournier GP, Neumann JE, Gogarten JP. 2010. Inferring the ancient history of the translation machinery and genetic code via recapitulation of ribosomal subunit assembly orders. *PLoS One* **5:** e9437.

Frey G, Michel CJ. 2003. Circular codes in archaeal genomes. *J Theor Biol* **223:** 413–431.

Frey G, Michel CJ. 2006. Identification of circular codes in bacterial genomes and their use in a factorization method for retrieving the reading frames of genes. *Comput Biol Chem* **30:** 87–101.

Gilbert W. 1986. The RNA world. *Nature* **319:** 618.

Grosjean H, Westhof E. 2016. An integrated, structure- and energy-based view of the genetic code. *Nucleic Acids Res* **44:** 8020-8040. Hartman H, Smith TF. 2014. The evolution of the ribosome and the genetic code. *Life (Basel)* **4:** 227–249.

Hsiao C, Mohan S, Kalahar BK, Williams LD. 2009. Peeling the onion: Ribosomes are ancient molecular fossils. *Mol Biol Evol* **26:** 2415–2425.

Hsiao C, Lenz TK, Peters JK, Fang PY, Schneider DM, Anderson EJ, Preeprem T, Bowman JC, O'Neill EB, Lie L, et al. 2013. Molecular paleontology: a biochemical model of the ancestral ribosome. *Nucl Acids Res* **41:** 3373–3385.

Ikehara K. 2002. Origins of gene, genetic code, protein and life: comprehensive view of life systems from a GNC-SNS primitive genetic code hypothesis. *Journal of Biosciences* **27:** 165–186.

Jenner LB, Demeshkina N, Yusupova G, Yusupov M. 2010. Structural aspects of messenger RNA reading frame maintenance by the ribosome. *Nat Struct Mol Biol* **17:** 555–560. |

Khade P, Joseph S. 2010. Functional interactions by transfer RNAs in the ribosome. *FEBS Lett* **584:** 420–426.

Koonin EV, Novozhilov AS. 2017. Origin and Evolution of the Universal Genetic Code. *Annu Rev Genet.* **51:** 45–62.

Koonin EV. 2017. Frozen Accident Pushing 50: Stereochemistry, Expansion, and Chance in the Evolution of the Genetic Code. *Life (Basel)* **7:** 22.

Kovacs NA, Petrov AS, Lanier KA, Williams LD. 2017. Frozen in Time: The History of Proteins. *Mol Biol Evol* **34:** 1252–1260.

Kunnev D, Gospodinov A. 2018. Possible Emergence of Sequence Specific RNA Aminoacylation via Peptide Intermediary to Initiate Darwinian Evolution and Code Through Origin of Life. *Life (Basel)* **8:** E44.

Lecompte O, Ripp R, Thierry JC, Moras D, Poch O. 2002. Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res* **30:** 5382-5390.

Lupas AN, Alva V. 2017. Ribosomal proteins as documents of the transition from unstructured (poly)peptides to folded proteins. *J Struct Biol* **198:** 74–81.

Ma W. 2010. The scenario on the origin of translation in the RNA world: In principle of replication parsimony. *Biol Direct* **5:** 65.

Maier UG, Zauner S, Woehle C, Bolte K, Hempel F, Allen JF, Martin WF. 2013. Massively convergent evolution for ribosomal protein gene content in plastid and mitochondrial genomes. *Genome Biol Evol* **5:** 2318–2329.

Melnikov S, Ben-Shem A, Garreau de Loubresse N, Jenner L, Yusupova G, Yusupov M. 2012. One core, two shells: bacterial and eukaryotic ribosomes. *Nat Struct Mol Biol* **19:** 560–567.

Michel CJ. 2008. A 2006 review of circular codes in genes. *Comput Math Appl* **55:** 984–988.

Michel CJ. 2012. Circular code motifs in transfer and 16S ribosomal RNAs: a possible translation code in genes. *Comput Biol Chem* **37**: 24-37.

Michel CJ. 2013. Circular code motifs in transfer RNAs. *Comput Biol Chem* **45:** 17–29.

Michel CJ, Pirillo G. 2013. A permuted set of a trinucleotide circular code coding the 20 amino acids in variant nuclear codes. *J Theor Biol* **319:** 116-121.

Michel CJ. 2015. The maximal C3 self-complementary trinucleotide circular code X in genes of bacteria, eukaryotes, plasmids and viruses. *J Theor Biol* **380:** 156-177.

Michel CJ. 2017. The maximal C3 self-complementary trinucleotide circular code X in genes of bacteria, archaea, eukaryotes, plasmids and viruses. *Life* **20:** 1–16.

Michel CJ, Ngoune VN, Poch O, Ripp R, Thompson JD. 2017. Enrichment of Circular Code Motifs in the Genes of the Yeast Saccharomyces cerevisiae. *Life (Basel)* **7:** E52.

Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*. 29: 2933-2935.

Nirenberg MW, Matthaei JH. 1961. The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proc Natl Acad Sci U S A* **47:** 1588–1602.

Noller HF. 2012. Evolution of protein synthesis from an RNA world. *Cold Spring Harb Perspect Biol* **4:** a003681. Nutman AP, Bennett VC, Friend CR, Van Kranendonk MJ, Chivas AR. 2016. Rapid emergence of life shown by discovery of 3,700-million-year-old microbial structures. *Nature* **537:** 535–538.

Opron K, Burton ZF. 2018. Ribosome Structure, Function, and Early Evolution. *Int J Mol Sci* **20:** E40.

Petrov AS, Gulen B, Norris AM, Kovacs NA, Bernier CR, Lanier KA, Fox GE, Harvey SC, Wartell RM, Hud NV, Williams LD. 2015. History of the ribosome and the origin of translation. *Proc Natl Acad Sci U S A* **112:** 15396–15401.

Plankensteiner K, Reiner H, Rode BM. 2005. Prebiotic Chemistry: The Amino Acid and Peptide World. *Curr Org Chem* **9:** 1107–1114.

Polacek N, Mankin A. 2005. The ribosomal peptidyl transferase center: Structure, function, evolution, inhibition. *Crit Rev Biochem Mol Biol* **40:** 285–311.

Polyansky AA, Hlevnjak M, Zagrovic B. 2013. Proteome-wide analysis reveals clues of complementary interactions between mRNAs and their cognate proteins as the physicochemical foundation of the genetic code. *RNA Biol* **10:** 1248–1254.

Root-Bernstein R and Root-Bernstein M. 2019. The Ribosome as a Missing Link in Prebiotic Evolution III: Over-Representation of tRNA- and rRNA-Like Sequences and Plieofunctionality of Ribosome-Related Molecules Argues for the Evolution of Primitive Genomes from Ribosomal RNA Modules. *Int J Mol Sci.* **20:** E140.

Shepherd JCW. 1981. Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proc Natl Acad Sci U S A* **78:** 1596–1600.

Smith TF, Lee JC, Gutell RR, Hartman H. 2008. The origin and evolution of the ribosome. *Biol Direct* **3:** 16.

Szathmáry E. 1999. The origin of the genetic code: Amino acids as cofactors in an RNA world. *Trends Genet* **15:** 223–229.

van der Gulik PT, Speijer D. 2015. How amino acids and peptides shaped the RNA world. *Life (Basel)* **5:** 230–246.

Vitas M, Dobovišek A. 2018. In the Beginning was a Mutualism - On the Origin of Translation. *Orig Life Evol Biosph* **48:** 223–243.

Woese CR, Dugre SA, Kando M, Saxinger WC. 1966. On the fundamental nature and evolution of the genetic code. *Cold Spring Harbor Symp Quant Biol* **31:** 720–736.

Yarus M, Widmann JJ, Knight R. 2009 RNA-amino acid binding: A stereochemical era for the genetic code. *J Mol Evol* **69:** 406–429.

Yarus M. 2017. The Genetic Code and RNA-Amino Acid Affinities. *Life (Basel)* **7:** 13.

Table 1. Location of the 13 *uX* motifs in the SSU rRNA alignment (prokaryotic 16S and eukaryotic 18S), according to structural domains and helices (*E. coli* numbering). *uX* motifs are labeled according to the accretion model of Petrov et al. 2015. The commas represent the decomposition of the *uX* motifs into trinucleotides of the circular code *X*. The underlined nucleotides in the *uX* motifs are present in more than 90% of the sequences in the alignment.

uX motif	Start	End	Sequence (*E. coli*)	Domain	Helix
a	1396	1404	*AC,ACC,GCC,C*	3'm	h44
b	1492	1501	*G,GGT,GAA,GTC,GTA,AC*	3'm	h44
c	1503	1514	*AG,GTA,ACC,GTA,GG*	3'm	h45
d	918	926	*A,ATT,GAC,GG*	3'M	h28
e	789	797	*TA,GAT,ACC,CTG,GTA,GTC,CA*	C	h24
f	1368	1377	*AC,GGT,GAA,TAC,GTT,C*	3'M	h43
g	520	525	*GC,CAG,CAG,C*	5'	h18
h	527	536	*GC,GGT,AAT,AC*	5'	h18
i	1186	1197	*G,GAT,GAC,GTC,AA*	3'M	h34
j	1333	1338	*AT,GAA,GTC,GG*	3'M	h42
k	249	257	*TA,GTA,GGT,GG*	5'	h11
l	1064	1073	*GT,CAG,CTC,GT*	3'M	h34, h35
m	1099	1107	*GC,AAC,GAG,C*	3'M	h35

Table 2. Location of the 19 *uX* motifs in the LSU rRNA alignment (prokaryotic 23S and eukaryotic 25/28S), according to structural domains and helices (*E. coli* numbering). *uX* motifs are labeled according to the accretion model of Petrov et al. 2015. The commas represent the decomposition of the *uX* motifs into trinucleotides of the circular code *X*. The underlined nucleotides in the *uX* motifs are present in more than 90% of the sequences in the alignment.

uX motif	Start	End	Sequence (*E. coli*)	Domain	Helix
A	2479	2484	*AT,ATC,GAC,GGC,GGT,GTT,T*	V	H89
B	2497	2511	*AC,CTC,GAT,GTC,GGC,T*	V	H89, H90
C	2516	2525	*AC,ATC,CTG,GG*	V	H91
D	2574	2586	*GC,GAG,CTG,GGT,TT*	V	H90, H93
E	2587	2596	*AG,AAC,GTC,GT*	V	H90, H93
F	2550	2561	*G,CTG,TTC,GCC,ATT,TA*	V	H92
G	2010	2015	*GT,GAA,ATT,GAA,CTC,GC*	0	H26a
H	513	519	*T,GAA,ACC,GT*	I	H2
I	724	732	*AA,CTG,GAG,GAC,C*	II	H34
J	699	708	*G,CAG,GTT,GAA,GGT,T*	II	H34
K	1975	1983	*GT,AAT,GAT,GGC,CAG,GC*	IV	H65, H67
L	804	812	*AG,CTG,GTT,CTC,C*	II	H32
M	1896	1905	*G,GTA,AAC,GGC,GGC,C*	IV	H68
N	1848	1853	*G,GAA,GGT,TA*	IV	H68
O	2654	2662	*AG,TAC,GAG,A*	V1	H95
P	1124	1131	*G,GAA,GAT,GTA,AC*	II	H41, H42
Q	1057	1062	*GC,CAG,GAT,GTT,GGC,TT*	II	H43
R	47	55	*A,GGC,GAT,GAA,GG*	I	H5
S	1388	1398	*AA,CAG,GTT,AAT,ATT,C*	III	H53

Table 3. Coverage of rRNA structural domains by *uX* motifs, in the LSU and SSU. Domain length corresponds to nucleotide length, and *uX* motif length is the total length of *X* motifs located in the domain in nucleotides. % coverage is the percentage of nucleotides in each domain covered by the universal *X* motifs.

rRNA domain	Domain start	Domain end	Domain length	*uX* motif length	% coverage
SSU 5'	1	559	559	25	4.5
SSU central	560	920	361	12	3.3
SSU 3'M	921	1398	478	56	11.7
SSU 3'm	1399	1542	144	28	19.4
Total SSU	1	1542	1542	121	7.8
LSU 0	disjoint	disjoint	159	6	3.8
LSU I	1	561	561	16	2.9
LSU II	587	1250	664	42	6.3
LSU III	1271	1647	377	11	2.9
LSU IV	1679	1989	311	25	8.0
LSU V	2058	2610	553	66	11.9
LSU VI	2626	2895	270	9	3.3
Total LSU	1	2895	2895	175	6

Table 4. Contacts (<5 Å) of the 13 *uX* motifs in the SSU rRNA alignment, with other *uX* motifs, mRNA, tRNA or ribosomal proteins. *uX* motifs are labeled according to the accretion model of Petrov et al. 2015. A, P, E in the tRNA column indicate contacts with the A-site, P-site and E-site tRNAs.

uX motif	Contacts				Functional site
	uX motif	mRNA	tRNA	Protein	
a	*b*	+	P	S5	P site; Ratchet pawl
b	*a*	+	A,P	S12	A site; P site
c	-	+		-	Ratchet pawl
d	-	+	P	S5	P site; Head swivel hinge
e	-	+	P,E	S11	P site; E site
f	-			S7,S9,S10,S14	
g	*h*			S12	
h	*g*	+	A	S3,S12	A site
i	*l*	+		S3,S5,S9,S10,S14	
j	-			-	PE loop
k	-			S17	
l	*i,m*			S2,S3,S5	Head swivel hinge
m	*l*			S2,S3	

Table 5. Contacts of the 19 *uX* motifs in the LSU rRNA alignment, with other *uX* motifs, tRNA or ribosomal proteins. Contacts are defined as <5 Å unless specified otherwise. *uX* motifs are labeled according to the accretion model of Petrov et al. 2015. A, P, E in the tRNA column indicate contacts with the A-site, P-site and E-site tRNAs.
* indicates bacteria specific ribosomal proteins.

uX motif	Contacts			Functional site
	uX motif	tRNA	Protein	
A	-	A	L16	
B	*D*	P	L3,L32*	PTC (<10 Å), exit tunnel
C	-	A	-	PTC (<30 Å)
D	*B*	A,P	L3,L32*	PTC (<10 Å), exit tunnel
E	-		L2	PTC (<30 Å), exit tunnel
F	*F*	A	L14	PTC (<10 Å), exit tunnel
G	-		L22,L32*	Exit tunnel
H	-		L20*,L22,L32*	Exit tunnel
I	*J*		L2	
J	*I*		L2	
K	-		L2	
L	-		L4,L15,L20*	Exit tunnel
M	-		L2	
N	-	E	-	
O	-		L6	SRL
P	-		L3,L13,L36*	L11 stalk
Q	-		-	L11 stalk - GAC
R	-		L34*	
S	-		L23	Exit tunnel

Figure 1. Properties of the *X* circular code. A. The definition of circularity implies that any word of the *X* code written on a circle has a unique decomposition. B. The *X* circular code is maximal (with 20 trinucleotides) and codes for 12 amino acids. C. The *X* code is composed of 10 trinucleotides and their complementary trinucleotides. D. The permutations of the *X* code associated with the shifted frames 1 and 2, named X_1 and X_2 respectively, are circular codes (C^3) and in addition are complementary to each other: a word in the shifted frame 1 of the strand 5'-3' is complementary to the word in the shifted frame 2 of the strand 3'-5', and *vice versa*. Note that X_1 and X_2 are shown in only one strand for simplicity, although they exist in both strands. E. According to the definition of a comma-free code, all words in the reading frame (frame 0) are valid (shown in blue), while all out-of-frame words are invalid (grey). For the *X* circular code, valid words may be present in frames 1 or 2, up to a length of at most 13 nucleotides.

Figure 2. Hypothesis of circular codes as a missing link in the early evolution of the translation system. The prebiotic contained RNA oligomers and amino acids that interacted non-specifically. They then coevolved to form an ancestral RNA-based 'translation' system, with more specific mapping between trinucleotides and amino acids. The RNA template evolved to form the RNA building blocks of the modern ribosome.

Figure 3. A. Location of the 13 *uX* motifs in the SSU rRNA alignments (prokaryotic 16S and eukaryotic 18S). The abscissa gives the nucleotide position referenced according to the *E. coli* 16S rRNA and the ordinate indicates the level of sequence conservation observed in the *uX* motifs. B. Location of the 19 *uX* motifs in the LSU rRNA alignments (prokaryotic 23S and eukaryotic 25/28S). The abscissa gives the nucleotide position referenced according to the *E. coli* 23S rRNA and the ordinate indicates the level of sequence conservation observed in the *uX*

motifs. Colored boxes indicate rRNA domains (positions in Table 3): for the SSU, light blue for domain 5', olive for the central domain, pink for 3'M and green for 3'm domains and for the LSU, magenta for domain I, blue for domain II, violet for domain III, white for domain 0, yellow for domain IV, pink for domain V, green for domain VI.

Figure 4. Secondary structure schema of the LSU and SSU rRNA (*E. coli*), showing the location of the *uX* motifs (red boxes). The schema is colored according to the six phases of the accretion model (Petrov et al. 2015) of ribosome evolution (phase 1: blue; phase 2: cyan; phase 3: green; phase 4: sepia; phase 5: brown; phase 6: purple). *uX* motifs are labeled with capital letters for LSU motifs and small letters for SSU motifs, according to their order of accretion in the different phases. PTC = peptidyl transferase center, CPK = central pseudoknot.

Figure 5. Distribution of the number and total nucleotide lengths of the *uR* random motifs in the SSU (16/18S) and LSU (23/28S) rRNA multiple alignments. The corresponding values for the *uX* motifs are indicated by a vertical red line. A. 2% of the random codes have the same number of universal motifs compared to *uX* motifs (number=32). B. 3% of the random codes have the same or larger total length of universal motifs compared to *uX* motifs (length=296).

Figure 6. *uX* motifs in the rRNA of *T. thermophilus*. A. LSU rRNA (green ribbon) with mRNA (orange sticks) and surface representations of tRNAs in the A-site (cyan), P-site (light blue) and E-site (deep teal). Nucleotides of the *uX* motifs are shown as magenta spheres. The PTC is identified by a black circle and the exit tunnel by a black arrow. B. SSU rRNA (pink ribbon) with tRNA colored as in A. Nucleotides of the *uX* motifs are shown as red spheres. C. Nucleotides in *uX* motifs close to the PTC (<10 Å in white sticks, <30 Å in magenta sticks, <50 Å in olive sticks). The distances were measured from atom N4 of CYT 2573 (white sphere).

All *uX* motifs are shown as magenta ribbons. D. All rRNA nucleotides (green ribbon) within 20 Å of the exit tunnel (black arrow) as defined by Dao Duc et al. 2019: nucleotides in *uX* motifs are colored according to rRNA domains, magenta for domain I, blue for domain II, violet for domain III, orange for domain 0, yellow for domain IV, pink for domain V (Table 3). tRNA are colored as in A. E. SSU rRNA nucleotides in contact with mRNA (<5 Å): nucleotides in *uX* motifs are colored according to rRNA domains, light blue for domain 5', olive for the central domain, pink for 3'M and green for 3'm domains (Table 3), other nucleotides and amicoumacin A (UAM) are white. Magnesium ions and their coordinated water molecules are represented by white spheres.

Figure 7. Proto-LSU and proto-SSU, with nucleotides and numbering from the contemporary *E. coli* 23S and 16S rRNA. *uX* motifs are highlighted in red and labeled according to the accretion model of Petrov et al. 2015, with 5'-3' direction indicated by red arrows. The dimeric proto-LSU (Agmon 2017) can be divided into A- and P-monomers corresponding to the modern A-tRNA and P-tRNA sites. Sequence complementarity of nucleotides building the conserved PTC walls in bacterial ribosomes is indicated by grey arrows in the PTC loop (connecting *X* trinucleotides shown in bold). The minimal proto-SSU model proposed by Agmon 2018 is shown in brown and the additional core segments identified by Petrov et al. 2015 are shown in yellow. PTC = peptidyl transferase center, CPK = central pseudoknot.

Figure 8. Proposed model of genetic code evolution associating codes, translation systems and peptide products at different stages from the primordial translation building blocks to the ancestor of the modern ribosome present in the Last Universal Common Ancestor (LUCA).