TRANSCRIPTION
2021, VOL. 12, NO. 1, 28–53
https://doi.org/10.1080/21541264.2021.1927652

Taylor & Francis
Taylor & Francis Group

POINT-OF-VIEW

Evolution of the genetic code

Lei Lei[a] and Zachary Frome Burton [b]

[a]Department of Biology, University of New England, Biddeford, ME, USA; [b]Department of Biochemistry and Molecular Biology, Michigan State University, E. Lansing, MI, USA

ABSTRACT
Diverse models have been advanced for the evolution of the genetic code. Here, models for tRNA, aminoacyl-tRNA synthetase (aaRS) and genetic code evolution were combined with an understanding of EF-Tu suppression of tRNA 3rd anticodon position wobbling. The result is a highly detailed scheme that describes the placements of all amino acids in the standard genetic code. The model describes evolution of 6-, 4-, 3-, 2- and 1-codon sectors. Innovation in column 3 of the code is explained. Wobbling and code degeneracy are explained. Separate distribution of serine sectors between columns 2 and 4 of the code is described. We conclude that very little chaos contributed to evolution of the genetic code and that the pattern of evolution of aaRS enzymes describes a history of the evolution of the code. A model is proposed to describe the biological selection for the earliest evolution of the code and for protocell evolution.

ARTICLE HISTORY
Received 5 March 2021
Revised 3 May 2021
Accepted 5 May 2021

KEWORDS
Archaea; aminoacyl-tRNA synthetases; amyloids; bacteria; elongation factor-Tu; liquid–liquid phase separation; polyglycine; protocells; standard genetic code; tRNA

1. Introduction

A model is presented for evolution of the genetic code based on analyses of tRNA and aminoacyl-tRNA synthetase (aaRS) evolution. The model is highly detailed and provides simple rules for filling code sectors. Strong selection rules are also apparent for the earliest evolution of the code.

A primordial tRNA^{Pri} was comprised of ordered sequences, GCG, CGC and UAGCC repeats and inverted repeats with 7-nt U-turn loops (homologous 17-nt anticodon and T stem-loop-stems) [1–6]. With the exception of a few anticodon loop and T loop bases, the tRNA^{Pri} sequence is completely known. Three tRNA evolution models are considered here [1,2,7–11]. Only the 3-31-nt minihelix model can be correct. The 3-31-nt minihelix model has been referred to as a theorem [2]. There are no theorems in evolutionary biology, but the 3-31-nt minihelix model for tRNA evolution is very close to being one.

AaRS enzymes attach amino acids to the 3 -end of tRNAs [12]. Much has been published on evolution of aaRS (i.e. GlyRS-IIA; IIA indicates the class (I or II) and structural subclass (i.e. A-E)) [1,3,13–18]. We have simplified the understanding of aaRS evolution and brought it in line with the evolution of the genetic code [1,3].

Evolution of tRNA and the genetic code provides new models for evolution of life on Earth and the pre-life to life transition. In agreement with some others, we posit that Archaea are the oldest organisms, and Archaea are the most similar to the last universal common (cellular) ancestor (LUCA) (Figure 1) [1,3,19–21]. Our interest in this issue comes from studies of earliest evolution of transcription [22,23] and translation systems [1,3,24]. We find that tRNAomes (all the tRNAs of an organism) are simpler in archaeal systems relative to bacterial systems [2,4,24,25]. AaRS enzymes are closer to root sequences in archaeal systems, and aaRS evolution falls more in line with the simpler archaeal genetic code [1,3,24,26]. Archaeal TFB is a homolog of bacterial σ factors [23,27], indicating that evolution of bacterial transcription systems may have largely drove divergence of Archaea and Bacteria [28]. In this work, we concentrate on archaeal systems for these reasons.

2. Methods

Sequences of tRNAs were obtained from the tRNA database [29] and the genomic tRNA database [30,31]. Typical tRNA diagrams were generated using tRNAdb tools and modified as necessary. Longer V loops in type II tRNA^{Leu} and tRNA^{Ser} were analyzed using WebLogo 3.7.4 [32].

CONTACT Zachary Frome Burton burton@msu.edu Department of Biochemistry and Molecular Biology, Michigan State University, E. Lansing, MI, USA

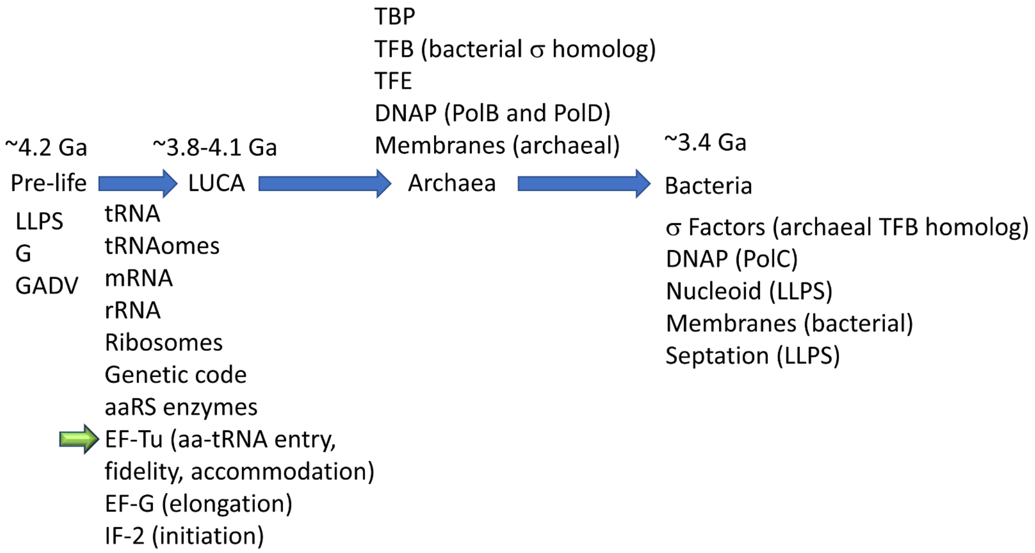


Figure 1. A working model for the pre-life to life transition and for divergence of Archaea and Bacteria. A major driving force for the divergence of Bacteria and Archaea is posited to have been divergence of transcription systems [28]. EF-Tu is highlighted (small green arrow) because EF-Tu evolution allowed expansion of the genetic code from an 8-aa bottleneck to the standard code. Abbreviations: LLPS) liquid–liquid phase separation; G) glycine; GADV) glycine, alanine, aspartic acid, valine; EF) translational elongation factor; IF) translational initiation factor; TBP) TATA-box binding protein; TF) transcription factor; DNAP (Pol)) DNA polymerase; 1 Ga) 1 billion years ago.

Molecular graphics was done using UCSF ChimeraX [33,34].

Evolution of aaRS enzymes was analyzed using the Phyre2 protein fold recognition server (see Figure 7c) [35]. Phyre2 identifies nearest and distant matches to a seed sequence, in the RCSB protein data bank. Phyre2 is a very useful tool for identifying close and distant protein family members that are related by both structure and sequence. Using only sequence-based tools, it is difficult to relate aaRS enzymes, which were driven to differentiate in order to establish and maintain translational accuracy of tRNA charging. Phyre2 was used to build a lineage of class I aaRS enzymes in which all class I aaRS were connected by both close and distant metrics (Phyre2 homology scores). Phyre2 could be used to build a model for a class II aaRS lineage. Some distantly related class II aaRS, however, could not be scored to one another. For instance, class IIA and class IID enzymes could not be scored as homologs without connecting intermediate class II aaRS. Because class I and class II aaRS have different folds, Phyre2 could not be used to identify homology of class I and class II aaRS enzymes. We used standard NCBI (National Center for Biotechnology Information) tools, such as Blast, to relate GlyRS-IIA to IleRS-IA and ValRS-IA (see Figure 7a-b) [1,3].

3. Pre-life evolution of transcription, metabolism and translation

There are some shared concepts comparing the earliest evolution of transcription and translation systems. On Earth, complexity was often generated from repetition of a motif. We imagine a pre-life mechanism to duplicate and multimerize RNAs (i.e. by ribozyme-mediated ligation and replication), often resulting in duplication of a common or related sequence [22,23,28,36]. If the RNA was protein-encoding, dimerization would generate a protein motif duplication. In this way, a β-β-α-β unit was duplicated to create a β-β-α-β – β-β-α-β motif refolded into a 6-β-sheet barrel [37–40]. Double-Ψ-β-barrels were generated in this way. Cellular RNAPs are 2-double-Ψ-β-barrel type enzymes [27,28,41,42]. PolD is a 2-double-Ψ-β-barrel type replicative DNAP from Archaea, and may be the most ancient replicative DNA polymerase [42–45]. RNA template-dependent RNAPs can also be of the 2-double-Ψ-β-barrel type, and these appear to be the oldest form of the enzyme class [46]. In Archaea, TFB includes a duplication of a helix-turn-helix motif ($(HTH)_2$), also referred to as a cyclin-like repeat. In Bacteria, σA is a homolog of TFB that appears to be derived from a $(HTH)_4$ repeat, which probably arose as a TFB $(HTH)_2$ duplication [22,23,36]. TBP was generated by duplication of a motif encoding multiple β-sheets and may be coevolved with DNA [47]. We consider TFB and TBP to be founding general transcription factors. It appears that Bacteria evolved sigma factors from TFB and lost TBP and that these were defining events in the divergence of Archaea and Bacteria. In many ways, Bacteria seem to be more successful prokaryotes than Archaea. For instance, many Archaea appear to be pushed into extremophile environments on Earth. We posit that Bacteria evolved from Archaea, and that Archaea are most similar to LUCA (Figure 1) [28]. For many factors and functions, Bacteria are simplified relative to Archaea, presumably due to genetic loss.

Much of core metabolism, including the glycolysis pathway and the citric acid cycle, evolved around $(\beta\text{-}\alpha)_8$ barrels (i.e. glycolysis; TIM (triose-phosphate isomerase) barrels) and $(\beta\text{-}\alpha)_8$ sheets (i.e. citric acid cycle; Rossmann folds). We posit that Rossmann fold $(\beta\text{-}\alpha)_8$ sheets were refolded from $(\beta\text{-}\alpha)_8$ barrels. We posit that $(\beta\text{-}\alpha)_8$ barrels and sheets were initially generated from two serial duplications of β-α-β-α motifs [22,28].

The evolution of tRNA from ligation of 3–31-nt minihelices, two of which were identical, is described below. Because of our experience with evolution of transcription systems, we searched for repeating motifs in tRNAs and found them easily. In pre-life, mostly, ribozyme-dependent mechanisms must have existed to replicate 31-nt minihelices and tRNAs. We posit that genetic code evolution was mostly driven by replication of tRNAs, mutation of the tRNA anticodon and coevolution of tRNAs with aaRS. AaRS enzymes evolved by a chaotic pathway, described below. Class I aaRS have their active site mounted on a platform of parallel β-sheets [12]. For this reason, class I aaRS are often referred to as “Rossmann folds”, but this is improper, as described below. Remarkably, the lineages of aaRS enzymes in Archaea give the pattern of genetic code evolution. The somewhat more complex pattern of genetic code evolution in Bacteria can be derived from the older archaeal pattern [1,3].

4. Evolution of tRNA

4.1. Concepts

We posit that evolution of tRNA, from an RNA – and ribozyme-dominated world, laid the foundation for evolution of the genetic code. We posit that the genetic code sectored according to the tRNA anticodon. Initially, code columns were selected because the central anticodon base (2nd position) was easiest to read on a primitive ribosome. Initially, both the 1st and 3rd anticodon positions were read as wobble positions with pyrimidine-purine discrimination. Evolution of EF-Tu suppressed wobbling at the 3rd anticodon position allowing expansion of the code. Because of the pathway to evolution of tRNA, the anticodon is read in a register of 3-nt, so 2-nt code registers could not be supported using tRNAs. Wobbling and code degeneracy are described by the evolution of tRNA reading on a primitive ribosome and coevolution of EF-Tu. Sequence analyses of type II tRNAs with longer V loops in ancient Archaea provides reasonable models for serine jumping during evolution of the code. Sequences of tRNAs in Archaea show that tRNA evolved from repeat and inverted repeat sequences. Therefore, before evolution of tRNA and protein encoding, there must have been ribozyme-based mechanisms to generate RNA repeats and inverted repeats. In this way, tRNA is a central key to understand the pre-life to life transition and evolution of the genetic code. TRNA is uniquely suited as a genetic adapter to support evolution of a genetic code. To generate a genetic code with a different adapter than tRNA presents many problems that we would not know how to solve.

4.2. The 3-31-nt minihelix model

The pathway of evolution of tRNA has been controversial, but tRNA evolution is essential to grasp, in order to understand the pre-life to life transition and evolution of the genetic code. Here, three models are considered, but we focus on one model, the 3–31-nt minihelix model [1,2]. So far as we can discern, the 3–31-nt minihelix model is the only viable model, and alternate models are falsified. The alternate models can be described as the Uroboros model [7,48,49] and the 2-minihelix model [8,10,50]. Both of these models are accretion models, in which tRNA evolves in expanding and/or contracting segments. Because proposed expansion and contraction segments in tRNA would derive from highly ordered sequences (i.e. repeats and inverted repeats), no accretion model can reasonably describe early tRNA evolution. In an accretion model, expansions and contractions would need to result in ordered sequences [1–3]. Furthermore, analyses of archaeal tRNA sequences provides an irrefutable record of the 3–31-nt minihelix model.

The 3–31-nt minihelix model is summarized in Figure 2. 3–31-nt minihelices of mostly known sequence were ligated to form a 93-nt tRNA precursor, which was then processed by internal 9-nt deletion(s) into type I and type II tRNAs [2]. 31-nt minihelices were comprised of a 5 -7-nt acceptor stem, a 17-nt core, and a 3 -7-nt acceptor stem. The sequence of the 5 -7-nt acceptor stem was originally GCGGCGG, which is a truncated GCG repeat. The sequence of the 3 -7-nt acceptor stem was originally CCGCCGC, which is a complementary, truncated CGC repeat. For the D loop, the 17-nt core sequence was originally UAGCCUAGCCUAGCCUA, which is a truncated UAGCC repeat. Remarkably, the anticodon and T loop 17-nt core sequences were both originally close to CCGGGUU/AAAAACCCGG (/indicates a U-turn in a 7-nt loop). These 17-nt sequences form a stem-loop-stem with 5-nt 5 -stems (CCGGG), a 7-nt U-turn loop (~UU/AAAAA) and 5-nt 3 -stems (CCCGG). There is only slight sequence ambiguity in the primordial 7-nt U-turn loop, not in the stems.

Type II tRNAs have a longer V loop (V for variable) relative to type I tRNAs [4]. The model shown in Figure 2 describes the evolution of both type I and type II tRNAs. To generate a type II tRNA, the 93-nt tRNA precursor was processed by a single 9-nt internal deletion, as shown. The 9-nt deletion occurred within ligated 3 – and 5 -7-nt acceptor stems. The 5-nt segment that remains

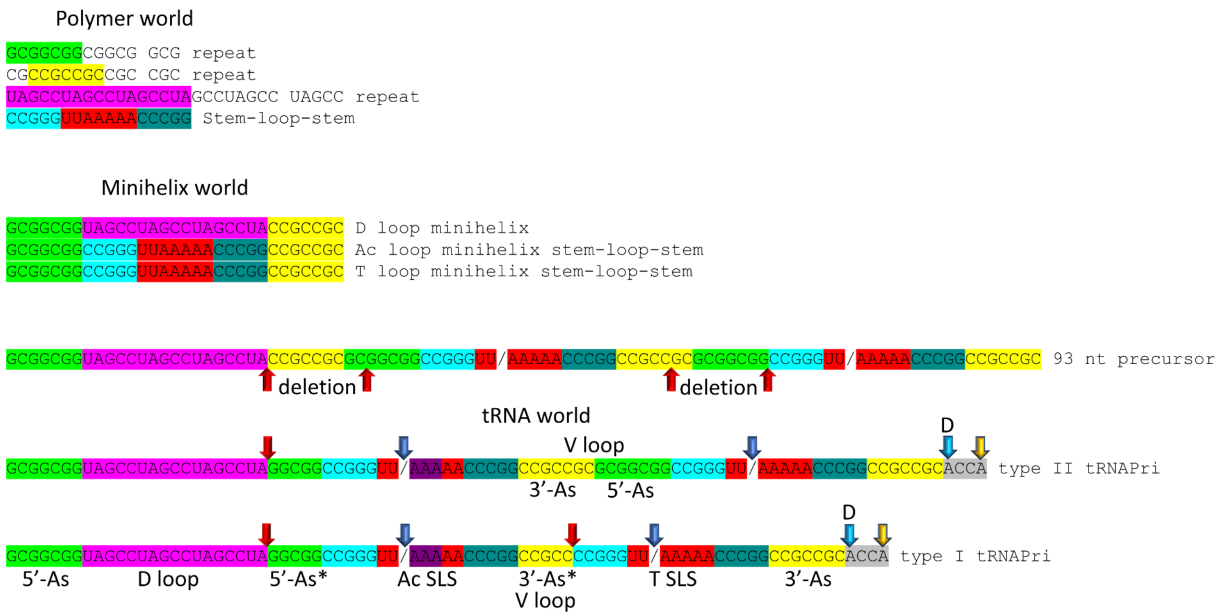

Polymer world
GCGGCGGCGGCG GCG repeat
CGCCGCCGCCGC CGC repeat
UAGCCUAGCCUAGCCUAGCCUAGCC UAGCC repeat
CCGGGUUAAAAACCCGG Stem-loop-stem
Minihelix world
GCGGCGGUAGCCUAGCCUAGCCUACCGCCGC D loop minihelix
GCGGCGGCCGGGUUAAAAACCCGGCCGCCGC Ac loop minihelix stem-loop-stem
GCGGCGGCCGGGUUAAAAACCCGGCCGCCGC T loop minihelix stem-loop-stem
GCGGCGGUAGCCUAGCCUAGCCUACCGCCGCGCGGCGGCCGGGUU/AAAAACCCGGCCGCCGCGCGGCGGCCGGGUU/AAAAACCCGGCCGCCGC 93 nt precursor
deletion
deletion
tRNA world
V loop
D
GCGGCGGUAGCCUAGCCUAGCCUAGGCGGCCGGGUU/AAAAACCCGGCCGCCGCGCGGCGGCCGGGUU/AAAAACCCGGCCGCCGCACCA type II tRNAPri
3’-As
5’-As
D
GCGGCGGUAGCCUAGCCUAGCCUAGGCGGCCGGGUU/AAAAACCCGGCCGCCCCGGGUU/AAAAACCCGGCCGCCGCACCA type I tRNAPri
5’-As
D loop
5’-As*
Ac SLS
3’-As*
V loop
T SLS
3’-As

Figure 2. The 3–31-nt minihelix model for evolution of tRNA. Sequences are primordial but remain represented in archaeal tRNAs. A 93-nt tRNA precursor formed from ligation of 3–31-nt minihelices, as shown. Type I and type II tRNAs were processed by 9-nt internal deletion(s) from the 93-nt precursor. Minihelix world was preceded by polymer world. Colors: green) 5 -acceptor stems (5 -As) and derived 5 -As*; magenta) D loop; cyan) 5 -stem for the anticodon (Ac) and T stem-loop-stems (SLS); red) anticodon and T loops; purple) anticodon; cornflower blue) 3 – stems for the anticodon and T stem-loop-stem; yellow) 3 -As* and 3 -As sequences. Arrow colors: red) internal deletion endpoints; blue) U-turns; cyan) discriminator (D); and gold) amino acid placements.

after deletion (originally GGCGG) became the last 5-nt of the D loop region just before the anticodon stem-loop-stem. The original type II tRNA, therefore, was 84-nt before addition of 3 -ACCA (A is the primordial discriminator base), presumably via ligation. The type II tRNA V loop, therefore, was initially 14-nt (7-nt + 7-nt). 14-nt remains a dominant length of $tRNA^{Leu}$ V loops in Archaea.

To generate type I tRNA from the 93-nt precursor, required two internal 9-nt deletions, as shown [2]. The 5 -9-nt deletion was identical to the processing event in type II tRNAs. The 3 -9-nt internal deletion in type I tRNA was also within ligated 3 – and 5 -7-nt acceptor stem segments. The 9-nt deletion generated the type I tRNA V loop, which was initially 5-nt (originally CCGCC). The primordial type I tRNA was 75-nt before addition of 3 -ACCA.

Folding into the tRNA L-shaped structure brought the D loop into contact with the V loop and the D stem in contact with the V region, causing a small number of systematic changes in tRNA sequences [2]. The 5 -acceptor stem remnant (5 -As*) initially changed from GGCGG to GGGCG to form the D stem and to break base pairing contacts to the V loop (numbered V1-V5). The V loop (3 -As*) changed with time from CCGCC to ~UGGUC. The V1 base is often U to form a wobble pair with G26. The V5 base tends to remain C to form the Levitt reverse Watson-Crick base pair to G15. Statistical tests support the homology of bases 3–7 of the 5 -acceptor stem to the 5 -As* sequence, with a p-value of 0.001 (highest indication of homology). Statistical tests support the homology of bases 1–5 of the 3 -acceptor stem to the 3 -As* sequence (the V loop), with a p-value of 0.001 (highest indication of homology) [5].

4.3. Type I tRNA

Figure 3a shows a type I tRNA colored according to the 3–31-nt minihelix model. Notice that the anticodon stem-loop-stem and the T stem-loop-stem are homologs (cyan-red-cornflower blue segments) with 7-nt U-turn loops, as indicated in the model (Figure 2). As in the model, the 5 -acceptor stem is a homolog of the 5 -As* sequence, and the 3 -As* is a homolog of the 3 -acceptor stem sequence (compare green and yellow segments).

Figure 3b shows a typical tRNA diagram (as DNA sequence) from *Pyrococcus furiosis* [29],

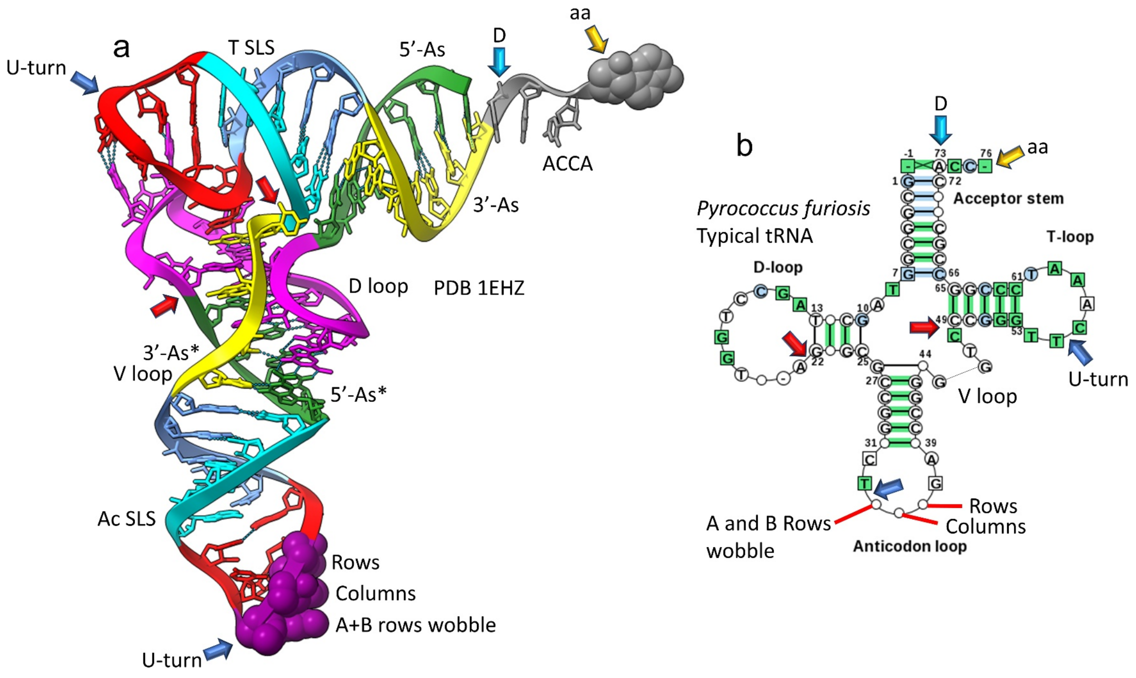


Figure 3. Type I tRNA. A) structure of a type I tRNA colored according to the model (Figure 2). B) A typical type I tRNA diagram from *Pyrococcus furiosis* [29]. Typical tRNA diagrams from the tRNAdb website are exported as DNA sequences rather than RNA. Arrow colors are as in Figure 2

which is an ancient Archaeon. The typical tRNA sequence is almost identical to the tRNA^{Pri} sequence shown in the model (Figure 2), indicating that the model is correct. The D loop sequence is UAGCNUAGCC, indicating conservation of the UAGCCUAGCC repeat sequence of tRNA^{Pri}. The homology of the anticodon stem-loop-stem (CCGGNCU/NNNGANCCGG) and the T stem-loop-stem (CCGGGUU/CAAAUCCCGG) is obvious by inspection. Statistical tests support this homology, with a p-value of 0.001 (highest indication of homology) [5]. The typical 5 -acceptor stem sequence is GCGGCGG, identical to the tRNA^{Pri} sequence. The 3 -acceptor stem sequence is CCGCNNC, consistent with the tRNA^{Pri} sequence (CCGCCGC). Acceptor stem sequences vary among tRNAs because acceptor stems are determinants for amino acid placements by aaRS enzymes at 3 -ACCA. The anticodon is highlighted (purple) because the genetic code evolved around the tRNA anticodon and its reading on the evolving ribosome.

4.4. *TRNA^{Gly} was the first tRNA*

The closest tRNA in Archaea to tRNA^{Pri} is tRNA^{Gly} (Figure 4) [1–3,24,25]. Figure 4a shows a typical tRNA^{Gly} from three *Pyrococcus* species. The typical sequence of tRNA^{Gly} is almost identical to tRNA^{Pri}. The sequence alignment is shown in Figure 4b. The tRNA^{Pri}, typical tRNA^{Gly} (*Pyrococcus*) and typical tRNA (*Pyrococcus furiosis*) are nearly identical sequences. This result indicates that tRNA^{Gly} was the first tRNA and that all archaeal tRNAs radiated from tRNA^{Gly}. This observation is relevant to the evolution of aaRS enzymes and the genetic code, as described below.

The typical D loop sequence of *Pyrococcus* tRNA^{Gly} is UAGUCUAGCCUGGUCUA (D1 to D17) versus UAGCCUAGCCUAGCCUA in tRNA^{Pri}. These sequences are nearly identical. The D12 A G shift from the primordial sequence appears to be universal in Archaea. D12G (19 G in standard tRNA numbering) intercalates into the T loop between 57A and 58A ((54-UU/CAAAU-60); standard numbering) and forms a specific

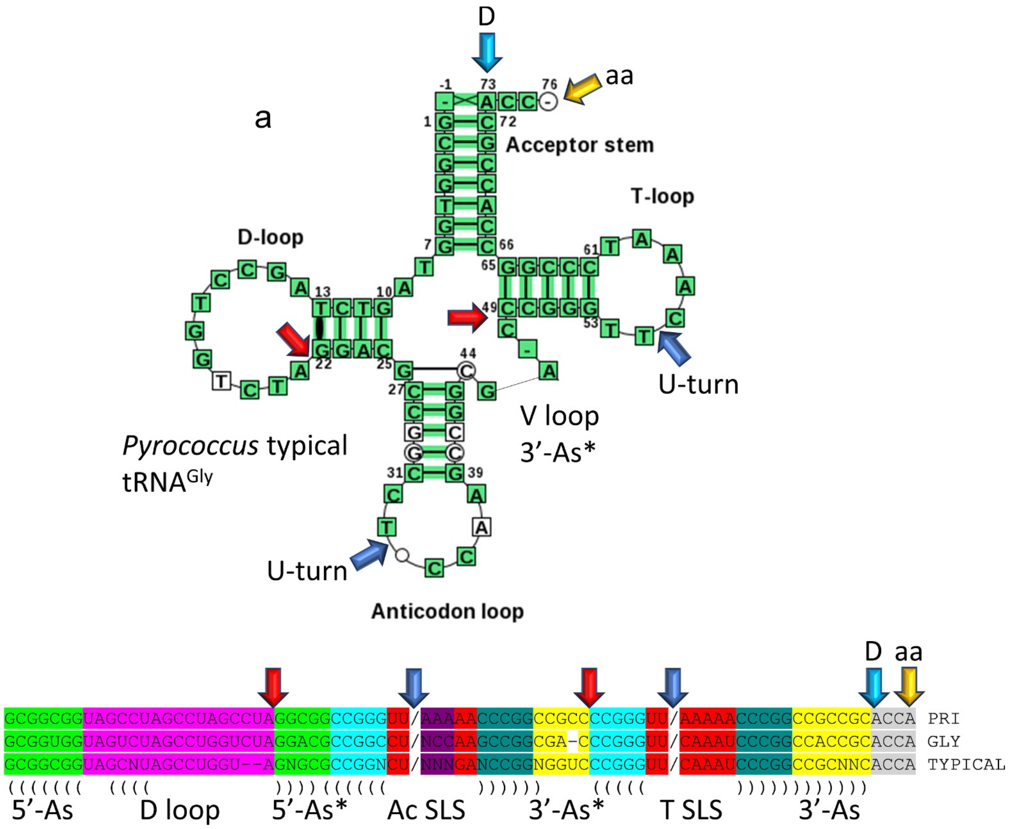

a
D
aa
-1
73
76
1
72
Acceptor stem
T-loop
D-loop
7
66
65
61
13
10
49
53
22
25
44
27
U-turn
V loop
3'-As*
Pyrococcus typical
tRNAGly
31
39
U-turn
Anticodon loop
b
D aa
GCGGCGGUAGCCUAGCCUAGCCUAGGCGGCCGGGUU/AAAAACCCGGCCGCCCCGGGUU/AAAAACCCGGCCGCCGCACCA PRI
GCGGUGGUAGUCUAGCCUGGUCUAGGACGCCGGCCU/NCCAAGCCGGCGA-CCCGGGUU/CAAAUCCCGGCCACCGCACCA GLY
GCGGCGGUAGCNUAGCCUGGU--AGNGCGCCGGNCU/NNNGANCCGGNGGUCCCGGGUU/CAAAUCCCGGCCGCNNCACCA TYPICAL
((((((((((( )))) (((((( )))))) ((((( ))))))))))))
5'-As
D loop
5'-As*
Ac SLS
3'-As*
T SLS
3'-As

Figure 4. TRNA^{Gly} was the first tRNA. A) A typical tRNA^{Gly} from three *Pyrococcus* species (*P. furiosis, P. abyssi and P. horikoshii*; 9 sequences). B) TRNA^{Pri}, tRNA^{Gly} and tRNA^{Typical} (*P. furiosis*) are close homologs. TRNA secondary structure is indicated. Colors and arrows are as in Figure 2

H-bond contact, explaining the A G sequence change. Apparently, 19 G is a preferred intercalating base to the primordial 19A. Note that the lengths of the D loop are identical in tRNA^{Gly} and tRNA^{Pri}. Standard tRNA numbering can be somewhat confusing compared to tRNA^{Pri} because standard numbering is based on a 2-nt deletion in the D loop in eukaryotic tRNAs.

4.5. Type II tRNAs

In Archaea, tRNA^{Leu} and tRNA^{Ser} are type II tRNAs, with longer V loops [4]. As we have shown previously, tRNA^{Leu} is most similar in overall sequence to type II tRNA^{Pri} (Figure 5). Figure 5a shows a tRNA^{Leu} from *Pyrococcus horikoshii* (PDB 1WZ2). The typical V loop has the 14-nt sequence UCCCGUAGGGGUUC (V1-V14). The V loop sequence varies from the primordial 14-nt CCGCCGCGCGGCGG because the primordial sequence can pair along its entire length, which would be awkward for tRNA folding and for functional contacts (i.e., with aaRS enzymes). Instead, the tRNA^{Leu} V loop evolved to form a short stem (i.e., 3-nt) and loop (i.e., 4-nt). Also, V1C typically became V1U to form a wobble pair with 26 G, and V14G became V14C to form a Levitt reverse Watson-Crick base pair with 15 G. Statistical tests support the model that the V loop is derived from a 3 -acceptor stem ligated to a 5 -acceptor stem (Figure 2), with a p-value of 0.001 (the highest indication of homology) [4].

TRNA^{Leu} has a 5 -acceptor stem with the typical sequence GCGGGGG versus GCGGCGG in tRNA^{Pri}. As in tRNA^{Gly} (Figure 4), the typical *Pyrococcus* tRNA^{Leu} D loop (Figure 5b) includes no deleted bases, with the 17-nt typical sequence UUGCCGAGCCUGGUCAA versus UAGCCUAGCCUAGCCUA in tRNA^{Pri}. These sequences are very similar.

The tRNA^{Ser} V loop evolved from the tRNA^{Leu} and tRNA^{Pri} sequences, in order to be distinguished from the tRNA^{Leu} V loop. Notably, SerRS-IIA interacts with the tRNA^{Ser} V loop directly as a determinant in order to recognize and accurately charge tRNA^{Ser} [12]. By contrast, the tRNA^{Leu} V loop is an anti-determinant to reduce inaccurate aminoacylation of tRNA^{Leu} by SerRS-IIA. LeuRS-IA recognizes the opposite face of tRNA^{Leu} from the expanded V loop, so LeuRS-IA recognizes other determinants such as the acceptor stem and discriminator base. A comparison of tRNA^{Ser} and tRNA^{Leu} V loop sequences is shown in Figure 6. Figure 6a is a typical tRNA^{Ser} diagram from

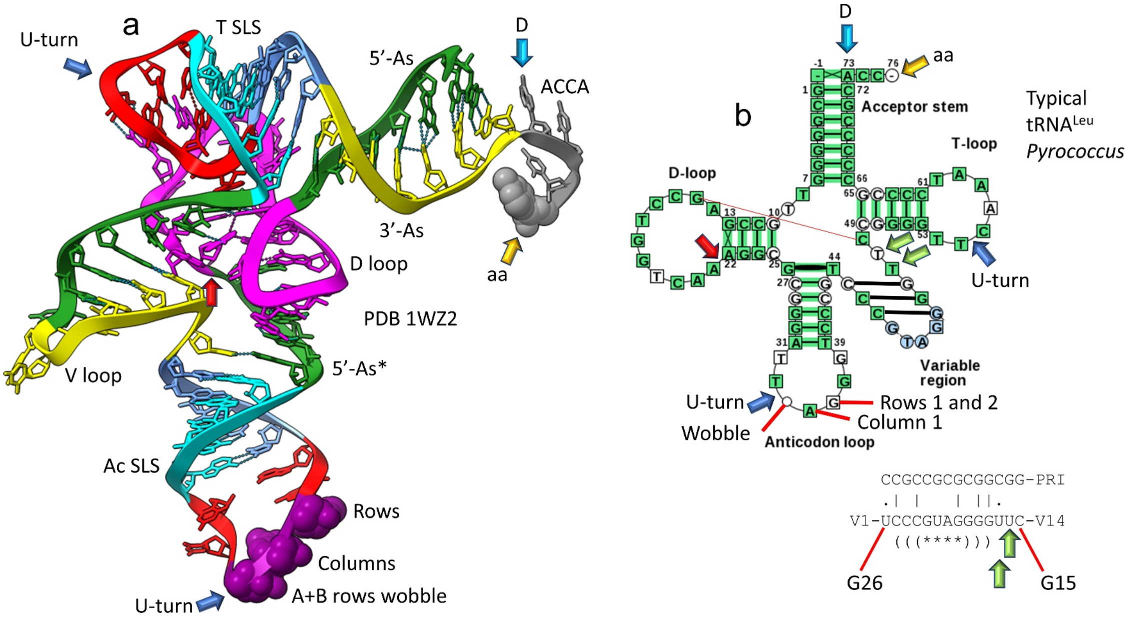


Figure 5. Type II tRNA. a) Structure of tRNA^{Leu} from *Pyrococcus horikoshii*. The larger V loop of type II tRNA is the 7-nt yellow segment linked to the 7-nt green segment. b) Typical tRNA^{Leu} from three *Pyrococcus* species (*P. furiosis, P. abyssi* and *P. horikoshii*; 15 sequences). Light green arrows indicate unpaired bases (V12U and V13U) just before the Levitt base pair (V14C = 15 G) (thin red line). Parentheses indicate paired bases; * indicates loop. Arrows and colors are as in .Figure 2

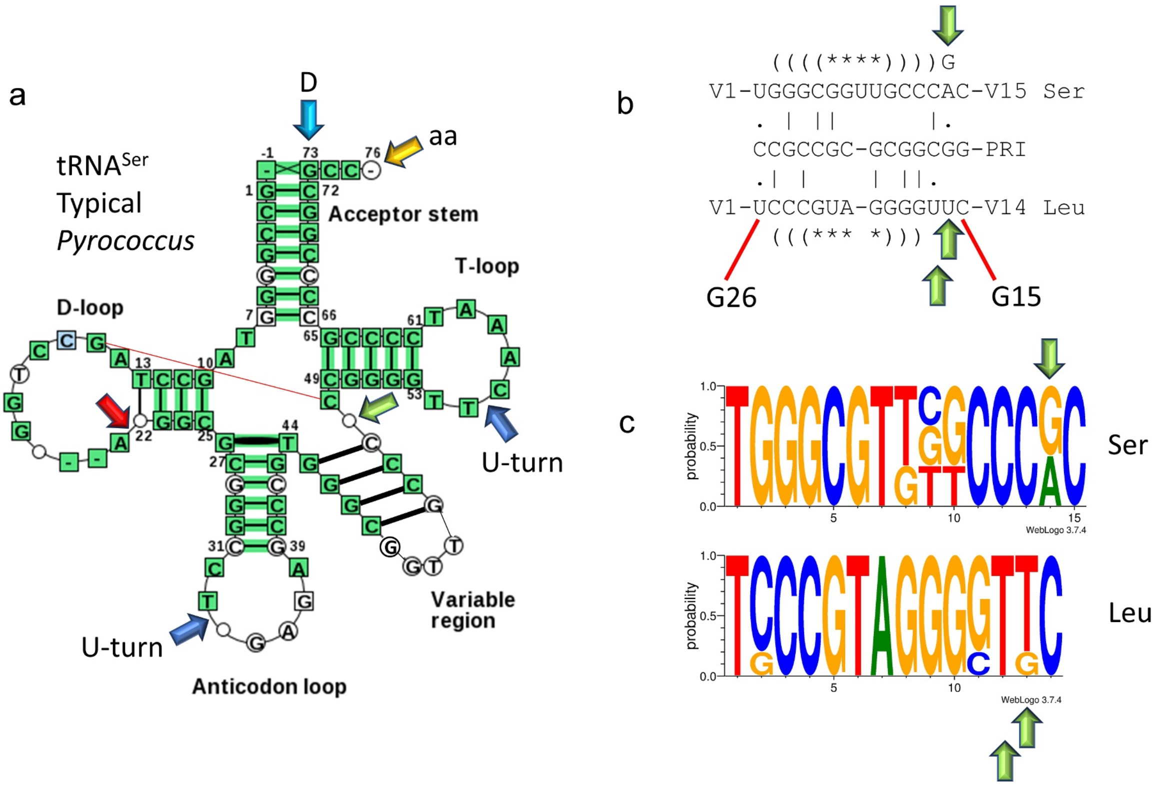


Figure 6. Comparison of tRNA^{Leu} and tRNA^{Ser} V loops. A) A typical tRNA^{Ser} from three *Pyrococcus* species (*P. furiosis, P. abyssi* and *P. horikoshii*; 12 sequences) (as DNA sequence). B) The alignment compares V loops of tRNA^{Pri} to tRNA^{Leu} and tRNA^{Ser} typical sequences. Parentheses indicate stems; * indicates loops. C) Sequence logos comparing tRNA^{Leu} and tRNA^{Ser} V loops (as DNA sequences). Light green arrows indicate unpaired bases after the V loop stem.

3-*Pyrococcus* species. Figure 6b shows an alignment of typical tRNA^{Ser} and tRNA^{Leu} V loop sequences versus tRNA^{Pri}. Figure 6c shows the comparison of tRNA^{Ser} and tRNA^{Leu} V loop sequence logos. We could not find a suitable tRNA^{Ser} structure (i.e., from an ancient Archaeon) to compare to the tRNA^{Leu} structure in Figure 5a, for instance, to compare the expected different trajectories of the V loops.

Accurate recognition of tRNA^{Ser} by SerRS-IIA is important in understanding the evolution of the genetic code. Serine is the only amino acid that sectors within two code columns, as described below. We posit that differences in sequence, stem positions and unpaired bases were important to discriminate the tRNA^{Ser} and tRNA^{Leu} V loops. Specifically, in *Pyrococcus*, tRNA^{Ser} has a single unpaired base just 3 of its short stem, while tRNA^{Leu} has two unpaired bases (light green arrows in Figures 5 and Figures 6). This difference should change the trajectory of the type II V loop stems. Also, the tRNA^{Ser} V loop is G-rich at its 5 stem, while the tRNA^{Leu} V loop is C-rich at its 5 stem. These and possibly other differences in type II V loops are expected to contribute to discrimination. The sequence logos in Figure 6c show that, in their major features, V loops are highly conserved in *Pyrococcus*. Solution of a structure of tRNA^{Ser} from a *Pyrococcus* species would contribute to this discussion.

5. Evolution of aaRS enzymes

5.1. Concepts

Aminoacyl-tRNA synthetases (aaRS) place amino acids at the 3 -end of tRNAs [12]. The idea behind this paper is that insight can be gained into the evolution of the genetic code based on coevolution of aaRS enzymes, tRNAomes and EF-Tu [1,3]. Using the Phyre2 protein fold recognition server [35], we were able to establish simplified pathways of aaRS evolution that appear to describe routes for genetic code evolution often within code

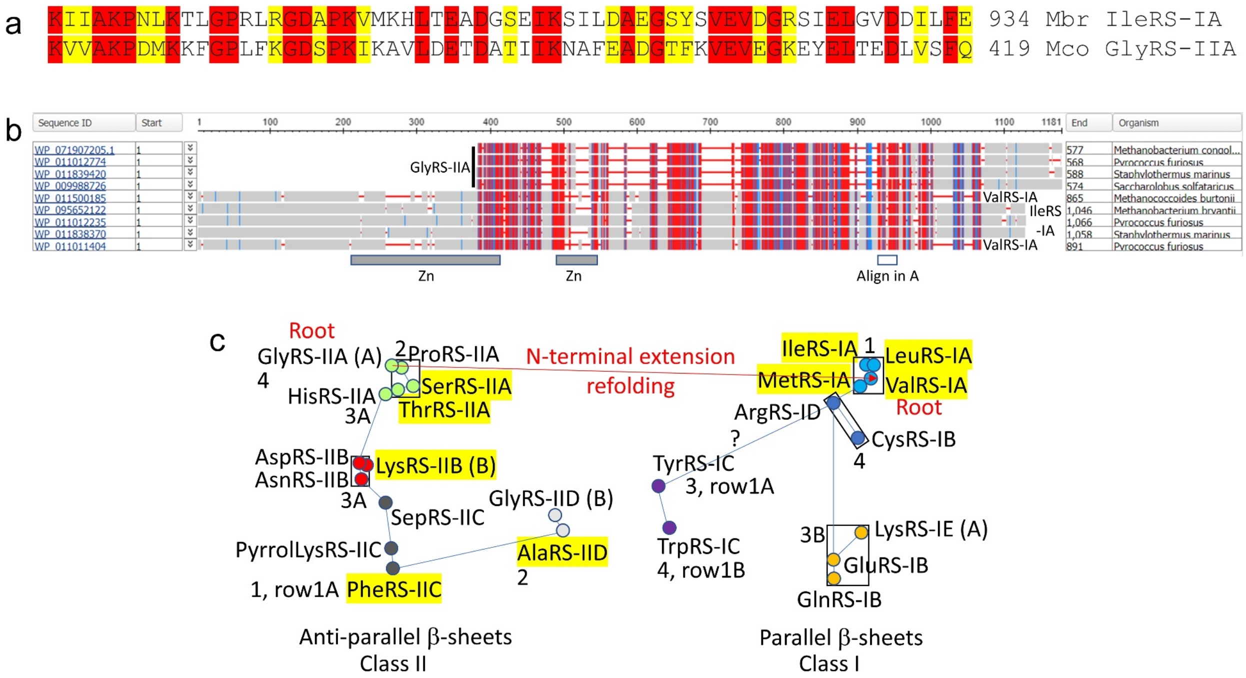


Figure 7. Evolution of aaRS enzymes. A) Local alignment of a class I IleRS-IA and a class II GlyRS-IIA aaRS. Identities are shaded red and similarities are shaded yellow. Mbr) *Methanobacterium bryantii*; Mco) *Methanobacterium congolense*. B) Class II and class I aaRS with incompatible folds are simple sequence homologs. Red boxes indicate sequence homologies in the multiple alignment. Two Zn fingers are indicated. The position of the alignment in panel A is shown. C) Evolution of aaRS enzymes based on Phyre2 homology scores [1,3]. Distances in the map represent structural and genetic relatedness. In Archaea, related aaRS enzymes mostly cluster according to genetic code columns (boxes; numbers indicate genetic code columns; some rows are specified). Yellow highlighting indicates aaRS with editing active sites. Red arrow indicates homology of GlyRS-IIA and ValRS-IA. (A) Archaea-specific; (B) Bacteria-specific. (?) indicates that TyrRS-IC and TrpRS-IC may be derived from a primitive ArgRS-ID.

columns (2nd anticodon position). Also, we noted sequence homology between class I and class II aaRS enzymes with different folds [26], which initially was unexpected. Class I aaRS enzymes have been termed "Rossmann folds", but this is a mischaracterization, as we describe.

5.2. Class II and class I aaRS enzymes

There are two structural classes of aaRS enzymes, termed class II and class I, and multiple structural subclasses (i.e., A-E) [12]. Some aaRS enzymes have a separate active site from the aminoacylating active site that removes non-cognate amino acids from their cognate tRNA. This process is referred to as "proofreading" or "editing". Remarkably, in Archaea, aaRS enzymes that edit non-cognate amino acids are found in the left half of the genetic code, in columns 1 and 2. Column 1 encodes hydrophobic amino acids Val, Met, Ile, Leu and Phe. Column 2 encodes neutral amino acids Ala, Thr, Pro and Ser. Ser is found also in column 4. We posit that Ser jumped from column 2 to column 4 in establishment of the code. In Archaea, ProRS-IIA does not include an editing active site, but ProRS-IIA does edit in Bacteria. Because neutral and hydrophobic amino acids are limited in forming specific hydrogen bonds and ion pairs, editing may be necessary to reduce tRNA charging errors.

Evolution of aaRS enzymes is described in Figure 7. In Figure 7a, a detail of an alignment of IleRS-IA from *Methanobacterium bryantii* to GlyRS-IIA from *Methanobacterium congolense* is shown. The e-value for the alignment is 4 × 10^{-12}, so, qualitatively, about a 1 chance in 2.5 × 10^{11} of the alignment being due to random chance rather than homology. Class I and class II aaRS enzymes, however, have incompatible folds, so these enzymes were not thought to be homologous. Class II aaRS appear to be the older fold, indicating that class I aaRS may be derived from class II aaRS [12]. In class II aaRS, the active site is mounted on a scaffold of antiparallel β-sheets. In class I aaRS, by contrast, the active site is mounted on a scaffold of parallel β-sheets. Commonly, class I aaRS are referred to as "Rossmann folds", although there is no genetic relation of class I aaRS and Rossmann fold enzymes. Rossmann fold enzymes derive from $(\beta\text{-}\alpha)_8$ sheets that appear to derive from refolding $(\beta\text{-}\alpha)_8$ barrels (i.e. TIM barrels; TIM for triosephosphate isomerase) [22].

5.3. Falsification of the Carter-Rodin-Ohno hypothesis

Figure 7b is a schematic of a multiple sequence alignment comparing GlyRS-IIA, IleRS-IA and ValRS-IA in ancient Archaea [1,3,24,26]. Figure 7b also shows how the aligned genes encoding class II and class I aaRS compare. Archaeal GlyRS-IIA is a simple sequence homolog of IleRS-IA and ValRS-IA (i.e., Figure 7a) showing that class IA aaRS were derived from GlyRS-IIA by N-terminal extension (i.e., upstream transcription and in-frame translation start) and refolding. Probably, this sequence comparison can only be done successfully with aaRS enzymes from ancient Archaea. One reason this homology is relevant is that a model (referred to as the Carter-Rodin-Ohno model) has been proposed that class I and class II aaRS were derived from "molten globule" smaller "Urzymes" encoded on complementary DNA strands [13,51–54]. These primitive Urzymes were posited to have expanded to full-length aaRS enzymes. Molten globule Urzymes must expand to a more complex version to take on their eventual specificity and refined functions. The Carter-Rodin-Ohno model is certainly incorrect. GlyRS-IIA is the root of all aaRS evolution, including all class II enzymes and all class I enzymes.

ValRS-IA and IleRS-IA enzymes include an N-terminal extension relative to GlyRS-IIA (Figure 7b). The N-terminal extension includes part of the class I aaRS active site scaffold and, in ancient Archaea, also, a Zn-finger lacking in GlyRS-IIA [26]. Also, GlyRS-IIA, IleRS-IA and ValRS-IA can share a Zn finger, as indicated. It is not possible that class II and class I aaRS are simultaneously simple homologs, as we show here, and also that class II and class I aaRS were generated as molten globule Urzymes from an ancestral bi-directional gene. Rather, GlyRS-IIA was a large and complex protein, not a molten globule, at the time of its refolding to (probably) a primitive ValRS-IA. Furthermore, ValRS-IA was a large and complex protein and not a molten globule Urzyme, from its first formation. We speculate that GlyRS-IIA and ValRS-IA assumed their initial and incompatible folds based, in part, on Zn and tRNA binding. Class II and class I aaRS bind opposite faces of their cognate tRNAs [12].

5.4. Lineages of aaRS enzymes

Figure 7c summarizes the following: 1) lineage information of aaRS enzymes; 2) aaRS enzymes with editing active sites; and 3) relationships of the aaRS lineages to the pattern of the genetic code [1,3,24,26]. We used the Phyre2 protein fold recognition server [35] in order to determine close and distant structural and sequence homologs among class II and class I aaRS [1,3,26]. What Phyre2 does is to score nearest and more distant homologs using both structure and sequence. As seed sequences, we used aaRS enzymes mostly from *Pyrococcus furiosis*, an ancient Archaeon. The Phyre2 server searches all sequences in the Protein Data Bank to find matches. Based on the homology scores, the lineage in Figure 7c was drawn. Distances in the map represent evolutionary distances, so clustered nodes indicate closely related enzymes. The map represents both close and distant relationships in the placements of the nodes [1,3]. AaRS enzymes with editing active sites are highlighted in yellow. Remarkably, the map closely matches the evolution and structure of the genetic code, indicating that the analysis of aaRS enzymes is reliable. Other approaches have not been as informative for the structure and evolution of the code [16–18,55–57].

The genetic code evolved primarily in columns, which represent the 2nd anticodon position. The anticodon central position is most important for translational accuracy. Closely related aaRS enzymes, therefore, tend to group within columns. This observation is explained in detail below. Strikingly, the pattern of aaRS evolution in Figure 7c describes a history of genetic code evolution.

6. EF-Tu and coding degeneracy

We posit that the translation functions of EF-Tu describe the evolution of coding degeneracy [24]. EF-Tu is a GTPase RNA chaperone that binds aminoacylated tRNA (aa-tRNA) and docks the aa-tRNA-EF-Tu complex on the ribosome. EF-Tu (translational elongation factor Tu) is a homolog of GTPases IF-2 (translational initiation factor 2) and EF-G (translational elongation factor G) [58]. These homologous GTPases occupy the same site on the ribosome during different phases of protein synthesis: initiation (IF-2), tRNA loading, clamping, accommodation (EF-Tu) and elongation (EF-G). EF-Tu is the major factor regulating translational fidelity on the ribosome [59–63]. With the incoming aa-tRNA-EF-Tu in the hybrid A/T ribosome docking site, EF-Tu hydrolyzes GTP and sets the aa-tRNA-mRNA "latch" or clamp. The latch tightens the tRNA anticodon-mRNA codon attachment. Specifically, the latch checks for Watson-Crick geometry at the 2nd and 3rd anticodon positions. The latch allows wobbling at the 1st anticodon position, the wobble position. At a wobble position, without modification of the wobble tRNA base, only pyrimidine-purine discrimination is achieved. The aa-tRNA-mRNA connection is monitored (latched) by 50S subunit 23S rRNA A1913 and by 30S subunit S12 and 16S rRNA G530, A1492 and A1493 (*Thermus thermophilus* ribosome numbering). After setting the latch, EF-Tu dissociates, and the released aminoacylated 3 -end of the aa-tRNA makes a long rotation into the peptidyl-transferase center (the A/A site), where peptide bond formation occurs. If a latched aa-tRNA-mRNA connection cannot be formed because of a base mismatch or inappropriate wobble pair, the inaccurately loaded aa-tRNA dissociates.

Before evolution of EF-Tu, therefore, the tRNA 3rd anticodon position could not have been read with 4-base accuracy. So, the 3rd anticodon position must have been a wobble position, limiting the complexity of the genetic code to 2 × 4 = 8-aa complexity. We posit that, before EF-Tu evolved, only one wobble position (the 1st or 3rd anticodon position) could be efficiently read at a time. Also, in Archaea, A is not read in the anticodon wobble position, so A (i.e., row 1 of the genetic code is 3rd position A), in a wobble position, formed an inefficiently utilized tRNA that functioned as a primitive translation stop signal. After evolution of EF-Tu, the 3rd anticodon position was locked down by the latch, and the maximum complexity of the genetic code became 2×4×4 = 32-assignments. Because of translational fidelity mechanisms, the standard genetic code froze at a complexity of 21-assignments: 20-aa + stops. EF-Tu allowed the genetic code, therefore, to escape an 8-aa bottleneck and expand to the standard code, as described below. Significantly, we posit that coding degeneracy evolved as a natural consequence of how tRNA was read on the primitive ribosome. EF-Tu evolved to expand the genetic code beyond 8-aa. Of course, it is possible that protein EF-Tu evolved to replace a ribozyme with some shared properties in locking down the 3rd anticodon position. At this time, it is difficult to know how complex the genetic code needed to become to encode functional enzymes. Here, we indicate that a code of 8-aa may have been sufficient to encode a primitive EF-Tu enzyme. Note that the 8-aa code we describe includes bending (G), bulky hydrophobics (A, V, L), hydrogen bonding (S), positive (R) and negative (D, E) amino acids. In Archaea, aa-tRNAs can be modified. Amination of D and E and addition of C for metal chelation could enrich an evolving code. C entered the code by a circuitous path described below.

7. Evolution of the genetic code

7.1. Overview

The genetic code evolved around the tRNA anticodon. In the wobble position of tRNA, only purine-pyrimidine resolution was typically achieved. Because of this limitation in reading tRNA on a ribosome, the genetic code evolved to have a maximum potential complexity of 32-assignments rather than 64-assignments, as in

row	1st									3rd	
		1		2		3		4			Col
		U		C		A		G			2nd
1A	U	PHE-IIC	A/GAA	SER-IIA	A/GGA	TYR-IC	A/GUA	CYS-IB	A/GCA	U/C	
1B		LEU-IA	U/CAA	SER-IIA	U/CGA	STOP	U/CUA	TRP-IC	U/CCA	A/G	
2A	C	LEU-IA	A/GAG	PRO-IIA	A/GGG	HIS-IIA	A/GUG	ARG-ID	A/GCG	U/C	
2B		LEU-IA	U/CAG	PRO-IIA	U/CGG	GLN-IB	U/CUG	ARG-ID	U/CCG	A/G	
3A	A	ILE-IA	A/GAU	THR-IIA	A/GGU	ASN-IIB	A/GUU	SER-IIA	A/GCU	U/C	
3B		MET-IA	U/CAU	THR-IIA	U/CGU	LYS-IE	U/CUU	ARG-ID	U/CCU	A/G	
4A	G	VAL-IA	A/GAC	ALA-IID	A/GGC	ASP-IIB	A/GUC	GLY-IIA	A/GCC	U/C	
4B		VAL-IA	U/CAC	ALA-IID	U/CGC	GLU-IB	U/CUC	GLY-IIA	U/CCC	A/G	
		aa-aaRS	Ac	aa-aaRS	Ac	aa-aaRS	Ac	aa-aaRS	Ac		

Figure 8. The standard genetic code in Archaea as a codon-anticodon table with a complexity of 32-assignments. Amino acids and aaRS (aa-aaRS) are colored according to closely related aaRS enzymes to emphasize evolution within code columns (Col) (Figure 7c). Anticodon (Ac) bases in red are rarely or never used in Archaea. Boxes highlighted in gray have aaRS enzymes with separate editing active sites. Codon 5 3 positions are labeled 1st, 2nd and 3rd.

DNA and mRNA. In Figure 8, we show an annotated standard genetic code for Archaea with a maximum complexity of 32-assignments. The code is shown as a codon-anticodon table because the tRNA anticodon limits the complexity of the code. The amino acids are colored according to closely related aaRS enzymes (Figure 7c) to emphasize that most evolution occurred in genetic code columns, which represent the central position of the anticodon. AaRS enzymes that have editing active sites are highlighted in gray. Note that, in Archaea, aaRS that edit are limited to hydrophobic and neutral amino acids found in the leftmost two columns of the code. SerRS-IIA, which edits, is found in both column 2 and column 4. We posit that serine first invaded column 2 and then jumped to column 4 (see below).

We posit an approximate order of addition for amino acids entering the genetic code (Figure 9). Glycine appears to be the first amino acid to enter the code [1,3,64,65]. Two reasons to consider glycine as the first encoded amino acid are: 1) tRNA^{Gly} appears to be the first tRNA (Figure 4); and Figures 2) GlyRS-IIA appears to be the first aaRS enzyme (Figure 7c). Glycine appears to occupy the most favored position in the code (anticodons GCC, UCC and CCC; 2nd and 3rd anticodon position C). Glycine, alanine, aspartic acid and valine (GADV) appear to be the first four amino acids to enter the code [66–70]. GADV are the four simplest amino acids chemically. These

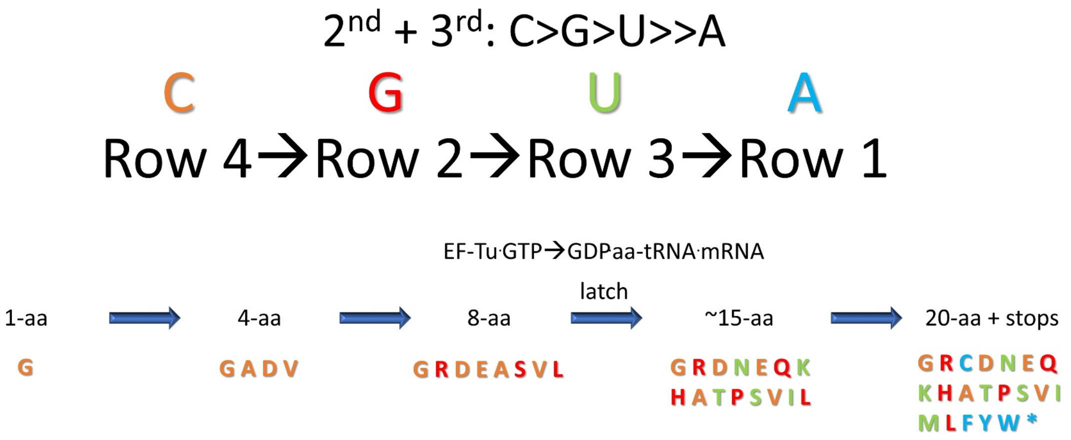
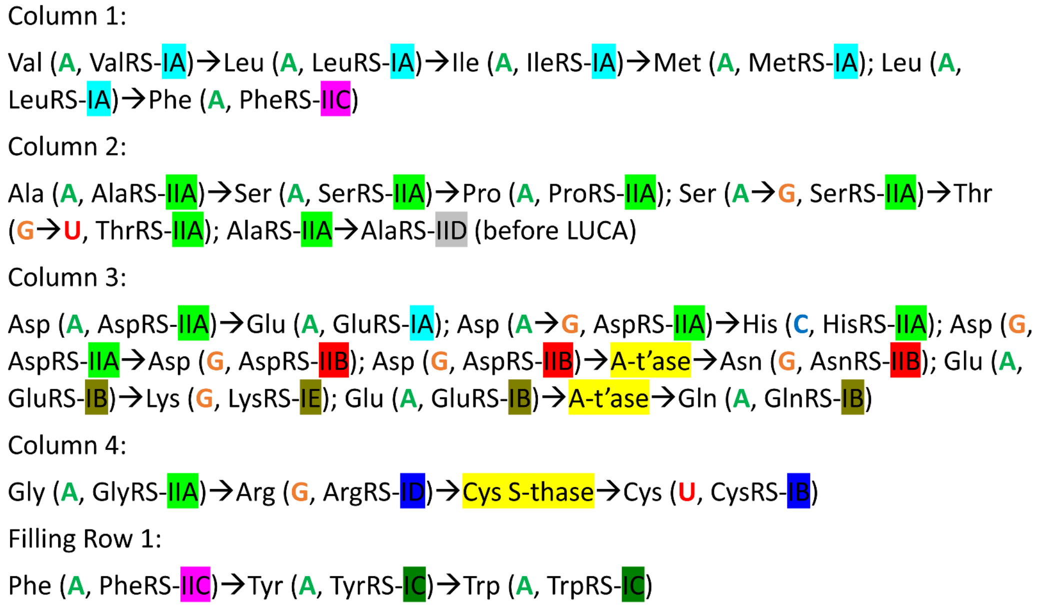


Figure 9. Proposed order of addition of amino acids into the genetic code. Amino acids appear to fill the code mostly by rows. 6-codon sectors for Leu, Ser and Arg were scored for their most favored anticodon. Ser jumps from column 2 to favored column 4, changing the Ser most-favored row assignment. Row 1 F, Y, C and W and row 3 M are posited to be the last additions to the code.

amino acids occupy row 4 of the code that appears to be the most favored row (3rd anticodon position C). We posit that Arg, Glu, Ser and Leu enter the code next. Arg, Ser and Leu end up occupying 6-codon sectors in the code. This is described in more detail below.

As described above, to progress beyond an 8-aa code required evolution of EF-Tu to suppress wobbling at the 3rd anticodon position [24]. After filling rows 4 and 2, row 3 and finally row 1 could be filled. Row 1 was disfavored because, initially, the 3rd anticodon position was a wobble position, and A is not allowed in a wobble anticodon position in Archaea. Because row 1 was disfavored, stop codons located to row 1. Stop codons are read by protein release factors that bind to mRNA codons [71], so no tRNA is associated with stop codons, except in suppressor tRNA strains. We judge the order of amino acid additions that we propose to be consistent with the following rules for tRNA anticodon position preferences: 1) for the 2nd and 3rd anticodon positions, C > G > U A; preferences are more extreme for the 3rd anticodon position because the 2nd anticodon position is central and easier to read; and 2) for the 1st anticodon (wobble) position, G>(U ~ C).

The genetic code evolved mostly within columns (a working model is summarized in Figure 10). Please refer back to Figures 7–10 as reference figures for the details of amino acid placements in the code. Genetic code columns represent the 2nd position in the tRNA anticodon, which is the most important position for translational accuracy. In the final steps, the genetic code filled row 1, which was initially disfavored. Row 1 was difficult to fill, because of wobbling of the 3rd anticodon position. Wobbling at the 3rd anticodon position was suppressed by evolution of EF-Tu [1,3,24,59,60]. We posit systematic rules for population of the code with amino acids (see above). These rules reflect preferences for the sequence of the tRNA anticodon. We posit that the genetic code evolved around the tRNA anticodon and the evolution of its reading on the ribosome. This mode of thinking describes the following features of the genetic code: 1) evolution in columns (Figures 7C and Figures 8); 2) evolution in rows (Figure 9); 3) the order of additions of amino acids (Figure 9); 4) late occupancy of row 1 (Figures 7C and Figures



Column 1:

Val (A, ValRS-IA)→Leu (A, LeuRS-IA)→Ile (A, IleRS-IA)→Met (A, MetRS-IA); Leu (A, LeuRS-IA)→Phe (A, PheRS-IIC)

Column 2:

Ala (A, AlaRS-IIA)→Ser (A, SerRS-IIA)→Pro (A, ProRS-IIA); Ser (A→G, SerRS-IIA)→Thr (G→U, ThrRS-IIA); AlaRS-IIA→AlaRS-IID (before LUCA)

Column 3:

Asp (A, AspRS-IIA)→Glu (A, GluRS-IA); Asp (A→G, AspRS-IIA)→His (C, HisRS-IIA); Asp (G, AspRS-IIA→Asp (G, AspRS-IIB); Asp (G, AspRS-IIB)→A-t'ase→Asn (G, AsnRS-IIB); Glu (A, GluRS-IB)→Lys (G, LysRS-IE); Glu (A, GluRS-IB)→A-t'ase→Gln (A, GlnRS-IB)

Column 4:

Gly (A, GlyRS-IIA)→Arg (G, ArgRS-ID)→Cys S-thase→Cys (U, CysRS-IB)

Filling Row 1:

Phe (A, PheRS-IIC)→Tyr (A, TyrRS-IC)→Trp (A, TrpRS-IC)

Figure 10. Summary of the proposed order of events in evolution of the genetic code, mostly by columns. In parentheses: (discriminator base (3 -XCCA; X = the discriminator) from *Pyrococcus*, aaRS). Colors are used as a guide for closely related aaRS enzymes and tRNA discriminator sequences. Classic identifications of aaRS subclass (i.e., ArgRS-ID) are not necessarily reliable (see Figure 7c). Yellow highlighting indicates tRNA-mediated enzymatic reactions: A-t'ase (Asp-tRNA^{Asn} and Glu-tRNA^{Gln} amidotransferase); Cys S-thase) Sep-tRNA^{Cys} Cys synthase (Sep for o-phosphoserine). See the text for details. Arrows indicate the order of amino acid entries into the code, mostly within columns, and not necessarily the lineage of aaRS evolution. TRNAs and aaRS can be reassigned in evolution.

9); 5) the complexity of the code; 6) evolution of 6-, 4-, 3-, 2- and 1-codon sectors; 7) evolution of stop codons; 8) coevolution of translation factors such as EF-Tu; 9) coevolution of aaRS enzymes with the code (Figures 7C and Figures 8); 10) complexity and structure in the 3rd genetic code column; 11) selections of code structures and amino acid placements; 12) serine jumping during code evolution; 13) evolution of translational fidelity; and 14) freezing of the code. The model, therefore, is highly predictive and descriptive for the final structure and sectoring of the code. Figures 11–17 describe a proposed order of addition of amino acids into the code. We know of no comparable model for genetic code evolution.

7.2. *Polyglycine world*

We posit that glycine was the first encoded amino acid [64,65], and that the genetic code first evolved to synthesize polyglycine (Figure 11) [1,3,24,25]. Initially, this was an assumption, but it turned out to be such a useful assumption, it should not be rejected easily. Also, if one were to choose another amino acid (i.e., Ala or Pro) as the first encoded amino acid, we do not believe reasonable rules can be as easily established for filling the code. Selecting Gly, as the initial encoded amino acid, however, a reasonable mechanism and simple rules for populating the code became apparent. We posit that the entire primitive code, including all anticodons and all codons, evolved to encode glycine. Row 1 tRNAs (1st anticodon position wobble A and 3rd anticodon position (initially) wobble A) were utilized inefficiently, so these tended to function as stops. We posit that wobbling at the 3rd anticodon position was suppressed by evolution of EF-Tu. In Archaea, A is not allowed in a wobble position [26]. Basically, we posit that a ribozyme mechanism existed to replicate templated tRNA (and other RNA) sequences. The anticodon is the easiest feature of tRNA to mutate without consequence for folding, so proliferation of tRNAs and mutations rapidly resulted in all possible anticodon sequences. We posit that a GlyRS ribozyme (GlyRS-RBZ) charged diverse tRNA^{Gly} with glycine because the code was not yet sufficiently evolved to generate protein catalysts. Hemolithin, recovered from meteorite samples, is a polyglycine peptide from outer space, indicating that a polyglycine world existed, even beyond an Earth environment [72].

One advantage of this mode of thinking is that it gives insight into the sectoring of the genetic code. If the entire primitive code encoded glycine, then invasion by other amino acids caused the glycine sector to contract. Invasion of the code by newly encoded amino acids, therefore, resulted in shrinking of previously occupied sectors. We then realized that clear rules could be stated for how incoming amino acids displaced previously added amino acids. Currently, glycine occupies the most favored anticodon positions in the code, which are GCC, UCC and CCC. If amino acids

			1		2		3		4	
			U		C		A		G	
1A	U	**STOP**	**A**/GA**A**	**STOP**	**A**/GG**A**	**STOP**	**A**/GU**A**	**STOP**	**A**/GC**A**	U/C
1B		**STOP**	U/CA**A**	**STOP**	U/CG**A**	**STOP**	U/CU**A**	**STOP**	U/CC**A**	A/G
2A	C	GLY-RBZ	**A**/GAG	GLY-RBZ	**A**/GGG	GLY-RBZ	**A**/GUG	GLY-RBZ	**A**/GCG	U/C
2B		GLY-RBZ	U/CAG	GLY-RBZ	U/CGG	GLY-RBZ	U/CUG	GLY-RBZ	U/CCG	A/G
3A	A	GLY-RBZ	**A**/GAU	GLY-RBZ	**A**/GGU	GLY-RBZ	**A**/GUU	GLY-RBZ	**A**/GCU	U/C
3B		GLY-RBZ	U/CAU	GLY-RBZ	U/CGU	GLY-RBZ	U/CUU	GLY-RBZ	U/CCU	A/G
4A	G	GLY-RBZ	**A**/GAC	GLY-RBZ	**A**/GGC	GLY-RBZ	**A**/GUC	GLY-RBZ	**A**/GCC	U/C
4B		GLY-RBZ	U/CAC	GLY-RBZ	U/CGC	GLY-RBZ	U/CUC	GLY-RBZ	U/CCC	A/G
		aa-aaRS	Ac	aa-aaRS	Ac	aa-aaRS	Ac	aa-aaRS	Ac	

Figure 11. Polyglycine world. A is inefficiently read in a wobble position (at this stage, both 1st and 3rd anticodon positions were wobble positions). Aa-aaRS) amino acid-aminoacyl-tRNA synthetase; RBZ) ribozyme; Ac) anticodon. Letters around the periphery indicate codon (mRNA) sequence. Colored shading for amino acids is maintained in Figures 11–Figures 17, so that placements of amino acids can be tracked. Red letters indicate anticodons that are not used or are used inefficiently.

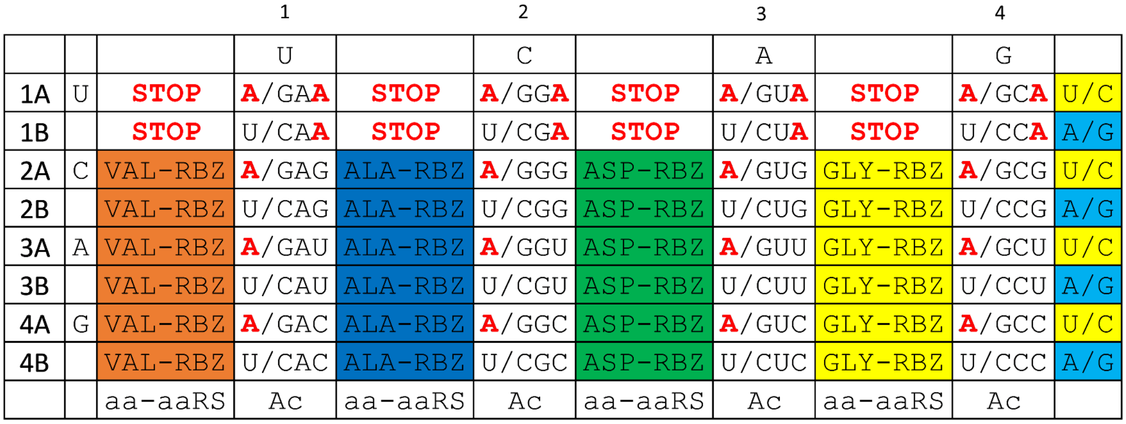
that entered the code at an early time protected the most advantageous sectors, then C was favored in the 2nd and 3rd anticodon positions. The rules for the 2nd and 3rd anticodon positions began to emerge as C > G > U A [1,3]. Preferences were strongest for the 3rd anticodon position, because the 2nd anticodon position was easier to read.

7.3. GADV world

We posit that polyglycine world gave way to GADV world (GADV for glycine, alanine, aspartic acid and valine) (Figure 12) [66–68,73]. Positing GADV world explains why the genetic code sectored in columns (Figures 7C and Figures 8). Glycine, alanine, aspartic acid and valine are the simplest amino acids chemically, so it is reasonable that these might be the first four amino acids to enter the code [66,67,69,70,73–75]. Also, GADV are amino acids that end up on the 4th row of the code, which corresponds to 3rd anticodon position C, in keeping with the C > G > U A rule for the 2nd and 3rd anticodon positions. At the GADV stage, we posit that tRNAs were charged by ribozymes.

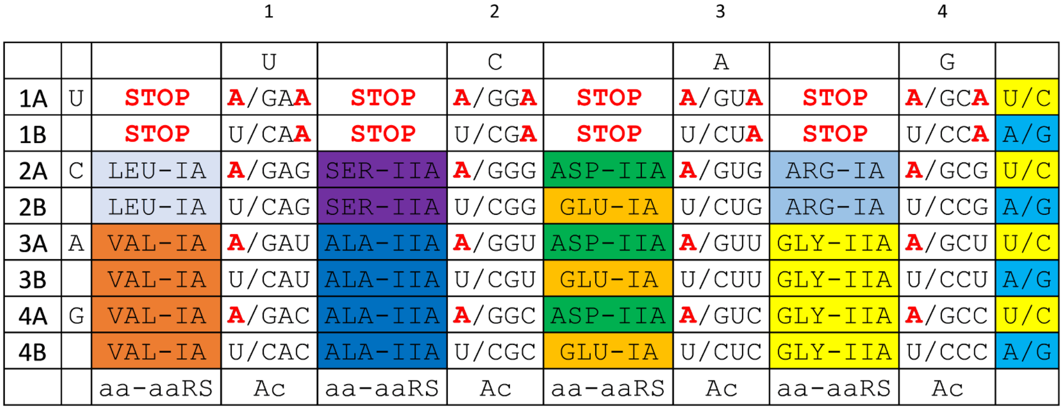
7.4. The 8-aa bottleneck

Before evolution of EF-Tu, the genetic code froze at 8-aa (Figures 13–15). EF-Tu suppresses wobbling at the 3rd anticodon position. Before EF-Tu, therefore, 3rd anticodon position A could not easily be read on the primitive ribosome. We note here that 1st anticodon position wobble A is not utilized in Archaea [1,3,26]. In Bacteria and Eukaryotes, wobble A is encoded only when it is modified to inosine, which broadens the set of recognized codons [76–79]. Interestingly, columns 1, 2 and 4 in the code sectored by one mechanism, and column 3, which became the most innovated



			1		2		3		4	
			U		C		A		G	
1A	U	STOP	A/GAA	STOP	A/GGA	STOP	A/GUA	STOP	A/GCA	U/C
1B		STOP	U/CAA	STOP	U/CGA	STOP	U/CUA	STOP	U/CCA	A/G
2A	C	VAL-RBZ	A/GAG	ALA-RBZ	A/GGG	ASP-RBZ	A/GUG	GLY-RBZ	A/GCG	U/C
2B		VAL-RBZ	U/CAG	ALA-RBZ	U/CGG	ASP-RBZ	U/CUG	GLY-RBZ	U/CCG	A/G
3A	A	VAL-RBZ	A/GAU	ALA-RBZ	A/GGU	ASP-RBZ	A/GUU	GLY-RBZ	A/GCU	U/C
3B		VAL-RBZ	U/CAU	ALA-RBZ	U/CGU	ASP-RBZ	U/CUU	GLY-RBZ	U/CCU	A/G
4A	G	VAL-RBZ	A/GAC	ALA-RBZ	A/GGC	ASP-RBZ	A/GUC	GLY-RBZ	A/GCC	U/C
4B		VAL-RBZ	U/CAC	ALA-RBZ	U/CGC	ASP-RBZ	U/CUC	GLY-RBZ	U/CCC	A/G
		aa-aaRS	Ac	aa-aaRS	Ac	aa-aaRS	Ac	aa-aaRS	Ac	

Figure 12. GADV-world. Explaining evolution in code columns (Figures 7C and Figures 8).



			1		2		3		4	
			U		C		A		G	
1A	U	STOP	A/GAA	STOP	A/GGA	STOP	A/GUA	STOP	A/GCA	U/C
1B		STOP	U/CAA	STOP	U/CGA	STOP	U/CUA	STOP	U/CCA	A/G
2A	C	LEU-IA	A/GAG	SER-IIA	A/GGG	ASP-IIA	A/GUG	ARG-IA	A/GCG	U/C
2B		LEU-IA	U/CAG	SER-IIA	U/CGG	GLU-IA	U/CUG	ARG-IA	U/CCG	A/G
3A	A	VAL-IA	A/GAU	ALA-IIA	A/GGU	ASP-IIA	A/GUU	GLY-IIA	A/GCU	U/C
3B		VAL-IA	U/CAU	ALA-IIA	U/CGU	GLU-IA	U/CUU	GLY-IIA	U/CCU	A/G
4A	G	VAL-IA	A/GAC	ALA-IIA	A/GGC	ASP-IIA	A/GUC	GLY-IIA	A/GCC	U/C
4B		VAL-IA	U/CAC	ALA-IIA	U/CGC	GLU-IA	U/CUC	GLY-IIA	U/CCC	A/G
		aa-aaRS	Ac	aa-aaRS	Ac	aa-aaRS	Ac	aa-aaRS	Ac	

Figure 13. The 8-aa bottleneck. Columns 1, 2 and 4 sectored differently than column 3 because of the selection of the wobble position (3rd anticodon position for columns 1, 2 and 4; 1st anticodon position for column 3).

column, encoding the most amino acids, sectored by a slightly different mechanism. We posit that the genetic code froze at 8-aa because both anticodon 1st and 3rd positions were wobble positions, and wobble positions are read with pyrimidine-purine (2-assignment) resolution. The 2nd anticodon position was much easier to read because it is the middle position. We further posit that only a single wobble position could be read at a time. Because, at a wobble position, only purine versus pyrimidine discrimination was initially achieved, the maximum complexity of the code was 2 × 4 = 8-aa. In columns 1, 2 and 4, the 2nd and 3rd anticodon positions were primarily read. Interestingly, leucine, serine and arginine are the amino acids that maintained 6-codon sectors (3 boxes in the genetic code tables shown). By contrast, in column 3, the 1st and 2nd anticodon positions were primarily read, giving the striped pattern of the related amino acids Asp and Glu. These differences in wobble selection gave rise to 6-codon sectors, for leucine, serine and arginine, and high innovation in column 3 (encoding many amino acids). Because of the geometry of the 7-nt U-turn anticodon loop in tRNA, the reading register for the primitive ribosome was always 3-nt.

In Figure 14, we posit that leucine, serine and arginine invaded row 3 of the code, displacing valine, alanine and glycine into favored row 4 (3rd anticodon position C). Positing this invasion of row 3 helps to describe the evolution of 6-codon sectors, the placement of threonine in the code and the jumping of serine from column 2 to column 4. Because valine, alanine and glycine retained the favored 4th row (3rd anticodon position C), these invasions were tolerated.

			1		2		3		4	
			U		C		A		G	
1A	U	**STOP**	**A**/GA**A**	**STOP**	**A**/GG**A**	**STOP**	**A**/GU**A**	**STOP**	**A**/GC**A**	U/C
1B		**STOP**	U/CA**A**	**STOP**	U/CG**A**	**STOP**	U/CU**A**	**STOP**	U/CC**A**	A/G
2A	C	LEU-IA	**A**/GAG	SER-IIA	**A**/GGG	ASP-IIA	**A**/GUG	ARG-IA	**A**/GCG	U/C
2B		LEU-IA	U/CAG	SER-IIA	U/CGG	GLU-IA	U/CUG	ARG-IA	U/CCG	A/G
3A	A	LEU-IA	**A**/GAU	SER-IIA	**A**/GGU	ASP-IIA	**A**/GUU	ARG-IA	**A**/GCU	U/C
3B		LEU-IA	U/CAU	SER-IIA	U/CGU	GLU-IA	U/CUU	ARG-IA	U/CCU	A/G
4A	G	VAL-IA	**A**/GAC	ALA-IIA	**A**/GGC	ASP-IIA	**A**/GUC	GLY-IIA	**A**/GCC	U/C
4B		VAL-IA	U/CAC	ALA-IIA	U/CGC	GLU-IA	U/CUC	GLY-IIA	U/CCC	A/G
		aa-aaRS	Ac	aa-aaRS	Ac	aa-aaRS	Ac	aa-aaRS	Ac	

Figure 14. The 8-aa bottleneck: Leu, Ser and Arg invaded row 3, helping to describe evolution of 6-codon sectors.

			1		2		3		4	
			U		C		A		G	
1A	U	LEU-IA	**A**/GA**A**	SER-IIA	**A**/GG**A**	**STOP**	**A**/GU**A**	ARG-IA	**A**/GC**A**	U/C
1B		LEU-IA	U/CA**A**	SER-IIA	U/CG**A**	**STOP**	U/CU**A**	ARG-IA	U/CC**A**	A/G
2A	C	LEU-IA	**A**/GAG	SER-IIA	**A**/GGG	ASP-IIA	**A**/GUG	ARG-IA	**A**/GCG	U/C
2B		LEU-IA	U/CAG	SER-IIA	U/CGG	GLU-IA	U/CUG	ARG-IA	U/CCG	A/G
3A	A	LEU-IA	**A**/GAU	SER-IIA	**A**/GGU	ASP-IIA	**A**/GUU	SER-IIA	**A**/GCU	U/C
3B		LEU-IA	U/CAU	SER-IIA	U/CGU	GLU-IA	U/CUU	ARG-IA	U/CCU	A/G
4A	G	VAL-IA	**A**/GAC	ALA-IIA	**A**/GGC	ASP-IIA	**A**/GUC	GLY-IIA	**A**/GCC	U/C
4B		VAL-IA	U/CAC	ALA-IIA	U/CGC	GLU-IA	U/CUC	GLY-IIA	U/CCC	A/G
		aa-aaRS	Ac	aa-aaRS	Ac	aa-aaRS	Ac	aa-aaRS	Ac	

Figure 15. The 8-aa bottleneck. Ser jumped from column 2 to column 4, invading the Arg sector. We posit evolution of a primitive EF-Tu or a ribozyme with EF-Tu-like activity suppressing wobbling at the 3rd anticodon position.

In Figure 15, we posit that serine jumped from column 2 into column 4. This event is described in more detail below. Leucine, serine and arginine began to invade disfavored row 1. This is considered in the model because leucine and serine end up in row 1 of the code. Also, a primitive ArgRS-ID may have evolved to TyrRS-IC, TrpRS-IC and CysRS-IB (Figure 7c) that end up in row 1. Evolution of a primitive EF-Tu at the 8-aa stage, or else evolution of a ribozyme to partly lock down the 3rd anticodon wobble position, might have this effect. We do not propose a specific order of events. We note that, at this stage, row 1 probably included tRNAs that were charged but not efficiently utilized, until EF-Tu or a corresponding ribozyme evolved. Ser jumping from column 2 to column 4 required only a single base change in the tRNA anticodon at the 2nd position (GGU GCU). The jump was favorable for placement of serine in the code because column 4 (2nd anticodon position C) was favored over column 2 (2nd anticodon position G).

7.5. Evolution of EF-Tu suppressed wobbling at the 3rd anticodon position and broke the 8-aa bottleneck

We posit that evolution of EF-Tu converted the 3rd anticodon position from a wobble position with pyrimidine-purine (2-assignment) resolution to a position with 4-base (A, G, C, U) resolution. This advance expanded the genetic code from a maximum complexity of 2 × 4 = 8-aa to a maximum complexity of 2×4×4 = 32-assignments. It should be noted that, in mRNA, the maximum complexity of the code is 4×4×4 = 64-assignments. The complexity of the code, which froze at 20-aa + stops = 21-assignments, was limited to a large extent by reading of tRNA on the ribosome.

			1		2		3		4	
			U		C		A		G	
1A	U	PHE-IIC	**A**/GAA	SER-IIA	**A**/GGA	TYR-IC	**A**/GUA	TYR-IC	**A**/GCA	U/C
1B		LEU-IA	U/CAA	SER-IIA	U/CGA	**STOP**	**U**/**C**UA	TYR-IC	U/CCA	A/G
2A	C	LEU-IA	**A**/GAG	PRO-IIA	**A**/GGG	HIS-IIA	**A**/GUG	ARG-ID	**A**/GCG	U/C
2B		LEU-IA	U/CAG	PRO-IIA	U/CGG	GLN-IB	U/CUG	ARG-ID	U/CCG	A/G
3A	A	ILE-IA	**A**/GAU	THR-IIA	**A**/GGU	ASN-IIB	**A**/GUU	SER-IIA	**A**/GCU	U/C
3B		ILE-IA	U/CAU	THR-IIA	U/CGU	LYS-IE	U/CUU	ARG-ID	U/CCU	A/G
4A	G	VAL-IA	**A**/GAC	ALA-IID	**A**/GGC	ASP-IIB	**A**/GUC	GLY-IIA	**A**/GCC	U/C
4B		VAL-IA	U/CAC	ALA-IID	U/CGC	GLU-IB	U/CUC	GLY-IIA	U/CCC	A/G
		aa-aaRS	Ac	aa-aaRS	Ac	aa-aaRS	Ac	aa-aaRS	Ac	

Figure 16. An intermediate ~17-aa stage and evolution of high innovation in column 3. We speculate that TyrRS-IC may be derived from a primitive ArgRS-ID and that tyrosine may have jumped from column 4 to column 3.

			1		2		3		4	
			U		C		A		G	
1A	U	PHE-IIC	**A**/GAA	SER-IIA	**A**/GGA	TYR-IC	**A**/GUA	CYS-IB	**A**/GCA	U/C
1B		LEU-IA	U/CAA	SER-IIA	U/CGA	**STOP**	**U**/**C**UA	TRP-IC	**U**/CCA	A/G
2A	C	LEU-IA	**A**/GAG	PRO-IIA	**A**/GGG	HIS-IIA	**A**/GUG	ARG-ID	**A**/GCG	U/C
2B		LEU-IA	U/CAG	PRO-IIA	U/CGG	GLN-IB	U/CUG	ARG-ID	U/CCG	A/G
3A	A	ILE-IA	**A**/GAU	THR-IIA	**A**/GGU	ASN-IIB	**A**/GUU	SER-IIA	**A**/GCU	U/C
3B		MET-IA	**U**/CAU	THR-IIA	U/CGU	LYS-IE	U/CUU	ARG-ID	U/CCU	A/G
4A	G	VAL-IA	**A**/GAC	ALA-IID	**A**/GGC	ASP-IIB	**A**/GUC	GLY-IIA	**A**/GCC	U/C
4B		VAL-IA	U/CAC	ALA-IID	U/CGC	GLU-IB	U/CUC	GLY-IIA	U/CCC	A/G
		aa-aaRS	Ac	aa-aaRS	Ac	aa-aaRS	Ac	aa-aaRS	Ac	

Figure 17. The standard genetic code. Amino acids in charcoal and gray were the last to enter the code. Disfavored row 1 was filled. Stop codons evolved on disfavored row 1. Met invaded the Ile sector. The code now has stops and starts.

An intermediate state of the code with ~17-aa is posited in Figure 16. In column 1, Ile was added as a 4-codon sector. Because Leu entered the code before Ile, and because Ile displaced Leu, Ile occupied row 3 (3rd anticodon position U), which was disfavored compared to row 2 (3rd anticodon position G) (C > G > U A). In column 2, Thr replaced Ser in row 3, and Pro replaced Ser in row 2. Because Thr and Ser are related amino acids, it is easy to see how SerRS-IIA could have duplicated and a copy could have diverged to evolve ThrRS-IIA (Figure 7c). ProRS-IIA may be derived from GlyRS-IIA. We note that, in Archaea, ProRS-IIA appears most closely related to GlyRS-IIA, and that neither enzyme evolved an editing active site (Figure 7c). Ser now occupied row 1. Column 3 filled to become the most innovated column encoding the most amino acids. The two founding amino acids in column 3, Asp and Glu, remained on favored row 4 (3rd anticodon position C). We posit that Asp, which entered the code first, occupied row 4A and Glu occupied row 4B because tRNA wobble G was favored over wobble U/C. We posit that AspRS was originally AspRS-IIA and evolved to AspRS-IIB after evolution of HisRS-IIA. AsnRS-IIB was later derived from AspRS-IIB. This is a continuing evolution in some archaeal species. GluRS-IB gave rise to LysRS-IE and GlnRS-IB (Figure 7c). GluRS-IB GlnRS-IB is ongoing in some archaeal species.

We begin to model the filling of disfavored row 1. We posit that phenylalanine and tyrosine may have entered the code at about this stage. PheRS-IIC may be derived from a primitive AspRS-IIB in several steps (Figure 7c). TRNA$^{\text{Phe}}$, in *Pyrococcus*, appears to be derived from tRNA$^{\text{Lys}}$ [25]. These identifications do not provide a simple model for placement of phenylalanine in the code. TyrRS-IC may be derived from a primitive ArgRS-ID, filling column 4, row 1, and jumping to column 3 by a single base change in the anticodon. We posit that pressure is building to evolve modern stop codons at this stage of evolution.

7.6. Filling disfavored row 1, evolution of stop codons and Met invasion of an Ile sector

To evolve the standard genetic code (Figure 17), then required filling in disfavored row 1 (3rd anticodon position A). We posit that Phe invaded column 1, row 1A, displacing Leu. Tyr invaded column 3, row 1A, perhaps as described above. Stop codons located to columns 3 and 4, row 1B. Cys invaded column 4, row 1A. We posit that CysRS-IB was derived from duplication and repurposing of a primitive ArgRS-ID. The classic naming of these aaRS enzymes is deceptive. CysRS-IB and ArgRS-ID are closely related enzymes, despite their IB and ID structural subclass designations (Figure 7c). Trp invaded column 4, row 1B. There are few 1-codon sectors in the genetic code, but Trp shares a sector with a stop codon, which is read in mRNA by a protein release factor [71], so there is no competing tRNA occupying the Trp sector (column 4, row 1B).

Evolution of row 1 appears somewhat chaotic. Some chaos might be expected in filling the last available positions in the code. It appears from Figure 7c that TyrRS-IC and TrpRS-IC might have been derived from a primitive ArgRS-ID. Tyrosine and tryptophan may have been two of the last amino acids added to the code [80]. CysRS-IB appears to be derived from a primitive ArgRS-ID. PheRS-IIC appears to have evolved in steps from a primitive AspRS-IIB. In *Pyrococcus*, tRNA$^{\text{Phe}}$ (discriminator A) appears to be derived from tRNA$^{\text{Lys}}$ (discriminator G). TRNA$^{\text{Tyr}}$ (discriminator A) appears to be derived from tRNA$^{\text{Asn}}$ (discriminator G). TRNA$^{\text{Trp}}$ (discriminator A) appears to be derived from tRNA$^{\text{Pro}}$ (discriminator A). TRNA$^{\text{Cys}}$ (discriminator U) appears to be derived from tRNA$^{\text{Thr}}$ (discriminator U) [25]. We guess that all of these tRNAs were assigned and then reassigned, and, therefore, their apparent derivations do not indicate the precise steps in adding these amino acids to the code. Reassignments of aaRS enzymes and tRNAs in evolution enhances translational accuracy by suppressing mischarging of tRNAs.

Cys appears to have entered the genetic code via a circuitous path [81–83]. Notably, some Archaea generate Cys from Sep-tRNA$^{\text{Cys}}$ charged by SepRS-IIC (Sep for o-phosphoserine). Modification of amino acids bound to tRNAs is a repeated theme in ancient Archaea that may reflect chemistry from a pre-life world [84,85]. This is also how Asn and Gln entered the code. Asp-tRNA$^{\text{Asn}}$ and Glu-tRNA$^{\text{Gln}}$ were aminated by Asp-tRNA$^{\text{Asn}}$ and Glu-tRNA$^{\text{Gln}}$ amidotransferase

[86–93]. TRNA-linked and RNA-linked chemistry must have been common in the pre-life world before evolution to cellular life [84,85].

In column 1, row 3B, Met invaded a 4-codon Ile sector. MetRS-IA and IleRS-IA are closely related enzymes (Figure 7c). Furthermore, tRNA^{Met} and tRNA^{Ile} are closely related tRNAs, in ancient Archaea such as *Pyrococcus* [25]. In keeping with our contention that 1-codon sectors were difficult to form and maintain, tRNA^{Ile}(CAU) and tRNA^{Met}(CAU) are both utilized in Archaea. The UAU anticodon, however, is rarely used. In tRNA^{Ile}(CAU), wobble C is converted to agmatidine to recognize Ile codon AUA but not Met codon AUG [94–100]. In tRNA^{Met}(CAU), wobble C is lightly modified and recognizes Met codon AUG but not Ile codon AUA. Wobble anticodon modification, therefore, resolves the ambiguity of tRNA^{Ile}(CAU) and tRNA^{Met}(CAU) but anticodon UAU was generally lost in the process. Met provides translation starts, in addition to bringing another amino acid into the code.

7.7. Code punctuation

We posit that genetic code starts and stops were late additions to the code. Translation initiation in Archaea has recently been reviewed [101]. Archaea utilize Met-tRNA^{iMet} (iMet for initiator methionine) and a set of translation initiation factors. Bacteria utilize fMet-tRNA^{iMet} (fMet for N-formyl-methionine) and a simplified initiation mechanism. In *Pyrococcus*, tRNA^{iMet}(CAU), tRNA^{Met}(CAU) and tRNA^{Ile}(CAU) are discriminated mostly based on acceptor stems. For tRNA^{iMet}, the 5 -acceptor stem sequence is AGCGGG(G), with the 3 -G uncharacteristically unpaired (opposite 3 -acceptor stem (G) CCCGCU). For tRNA^{Met}, the 5 -acceptor stem sequence is GCCGGGG, with all bases paired. For tRNA^{Ile} (CAU), the 5 -acceptor stem sequence is GGGCCCG, with all bases paired. It appears that selection of methionine as the initiating amino acid in Archaea was a complex coevolution of Met-tRNA^{iMet} and initiation factors. We guess that the need for a translation initiation start signal was a powerful driving force to evolve this system. For instance, to initiate translation at internal mRNA sites for gene expression regulation may have required evolution of the Met-tRNA^{iMet} translation initiation system. We guess that Bacteria simplified the archaeal system in evolution, adopting fMet-tRNA^{iMet} and shedding initiation factors during divergence from Archaea.

Evolution of translation stops appears to have been a complex process with multiple stages [71]. We posit that initially the problem was generating longer peptides to form more complex proteins. The system appears to have started with inefficient tRNAs, i.e., 1st or 3rd anticodon A, as primitive stop signals. The system suppressed 3rd anticodon position wobbling by evolving EF-Tu. Finally, protein translation release factors evolved to recognize stop codons in mRNA. Suppression of wobbling at the 3rd anticodon position by EF-Tu expanded the genetic code and may have driven evolution of protein release factors and stop codons.

7.8. Serine jumping from column 2 to column 4

We posit that jumping across columns was rare in establishment of the code. The advantage for Ser to jump from column 2 to column 4 was that serine obtained a favored anticodon. GCU was favored over GGU, because 2nd anticodon position C was favored over G. Serine could invade the arginine sector because ArgRS-ID reads type I tRNA^{Arg} and cannot read type II tRNA^{Ser}. Also, SerRS-IIA must bind the type II tRNA^{Ser} V loop to add Ser. There is no advantage to serine invading the leucine sector, and such an invasion would be problematic, partly because both tRNA^{Leu} and tRNA^{Ser} are type II tRNAs. Invasion of row 3 would not be advantageous for serine and would also eliminate an amino acid from the code.

7.9. Summary

In summary, a highly detailed working model is possible for evolution of the genetic code. The model is mostly based on tRNA and aaRS sequence analyses. The genetic code evolved around the tRNA anticodon. The model tracks evolution of aaRS enzymes, indicating that both the genetic code model and the model for aaRS evolution are substantially correct. When we started this work, we did not think such a detailed and predictive model was possible or reasonable. Now, we consider this a very strong model for further analysis of the code.

8. Discussion

8.1. tRNA evolution

Remarkably, tRNA evolution was determined to the last nucleotide (Figure 2) [2]. The solution to tRNA evolution was possible because $tRNA^{Pri}$ sequences were highly ordered and these repeats and inverted repeats can still be detected in ancient Archaea. No accretion model can describe tRNA evolution, because of conservation of highly regular $tRNA^{Pri}$ sequences. For an accretion model to have credence, tRNAs would need to expand, inserting ordered and preordained sequences, which seems unlikely if not impossible. Only the 3–31-nt minihelix model describes tRNA evolution. Accretion models, such as the Uroboros [7] and 2-minihelix [8,10] models are, therefore, falsified. Also, the genetic code evolved around the tRNA anticodon [1–3,24]. The genetic code complexity is determined by the way the tRNA was read on a ribosome, explaining why the genetic code went from being frozen at 8-aa complexity (Figures 13–15), before attaining the standard code (Figure 17) that has 21-assignments (20-aa + stops).

8.2. aaRS evolution

Analysis of aaRS evolution tracks evolution of the genetic code (Figure 7c) [1,3,24]. This analysis is easiest to do in archaeal systems because Archaea are the oldest organisms, and Bacteria are more derived (Figure 1). Because GlyRS-IIA, ValRS-IA and IleRS-IA are simple sequence homologs (Figure 7a and Figure 7b), the Carter-Ohno-Rodin hypothesis for aaRS evolution [13,51–53] is falsified. Class I and class II aaRS were not generated from opposite DNA strands of a primordial, bi-directional gene encoding molten globule Urzymes.

8.3. genetic code evolution

A highly detailed model has been generated that describes standard genetic code evolution in Archaea (Figures 10–17). Every aspect of code evolution is described by this model. Simple rules were developed to describe sectoring of the code. The model can be modified to generate the more derived genetic codes of Bacteria and Eukarya.

8.4. Freezing the code

We posit that new amino acids were introduced through tRNA charging errors and through aa-tRNA linked chemistry, and that translational fidelity mechanisms froze the code [1,3]. Based on archaeal systems, Asn, Gln and Cys appear to have initially entered the code through enzymatic mechanisms in which aa-tRNAs were modified. Subsequently, the tRNA-linked reactions were replaced by evolution of AsnRS-IIB, GlnRS-IB and CysRS-IB. Other amino acids may also have entered the code via tRNA-linked reactions. For instance, Arg may have replaced ornithine early in code evolution. Ornithine can be converted enzymatically to Arg in two steps [102]. Similarly, Leu may have been synthesized from tRNA-linked Val in 5 enzymatic steps. Because of initial wobbling in the 1st and 3rd anticodon positions, EF-Tu evolution was necessary to expand the code beyond 8-aa. Subsequently, EF-Tu contributed to freezing of the code by enforcing translational accuracy. Some aaRS have proofreading (editing) to remove inappropriately added amino acids from their cognate tRNAs [12]. Remarkably, the aaRS that edit are limited in Archaea to amino acids located in the left half of the code (columns 1 and 2; Figures 7C and Figures 8). Hydrophobic and neutral amino acids locate to columns 1 and 2 of the code, so editing helped with translational accuracy for amino acids with limited chemical character, such that these amino acids could not be as easily specified in the aaRS active site. Editing helped to freeze the code by protecting 6- and 4-codon sectors in the left half of the code. To add additional amino acids required splitting larger sectors of the code. 6-codon sectors encoding Leu, Ser and Arg resulted from the history of evolution, as described. Splitting a 2-codon sector into two

1-codon sectors was problematic because of tRNA wobble ambiguity (generally, in Archaea, tRNA wobble U ~ C and only G is allowed, not A). High innovation in column 3 of the code resulted from the history of column 3 sectoring and the initial selection of the wobble base (1st anticodon position). In the case of Ile and Met, co-occupancy of the CAU anticodon through wobble modifications resulted in suppression of the Ile UAU anticodon, as described.

8.5. A model for evolution of protocells

Figure 18 shows a model for evolution of the first cells. We posit that a number of ribozymes must have been present in order for tRNA to evolve prior to evolution of complex proteins [1,3]. TRNA^{Pri} was comprised completely of ordered sequences, notably, repeats and inverted repeats, so ribozymes must have existed to generate repeats, inverted repeats, 31-nt minihelices and tRNA [2] (Figure 2). Furthermore, polyglycine and GADV polymers appear to have been the first products of the evolving genetic code. Therefore, a selection for polyglycine and GADV polymers would describe the selective pressure for evolution of the code before complex proteins became possible. We posit that polyglycine and GADV polymers formed essential structures in protocells. Structures included cell walls, internal structures, amyloid plaques and LLPS (liquid–liquid phase separation) compartments. We note that $(Gly)_5$ is a component of bacterial cell walls [103,104] and may be a relic of a pre-life world.

Amyloid plaques form from assemblies of long, mis-associated β-sheets. In eukaryotic cells, amyloid plaques are a symptom of neurological diseases [105–107]. We posit that amyloid plaques may be generated from misregulation of LLPS membraneless compartments, which are a normal feature of eukaryotic cells. In the ancient world, amyloid plaques would have regulated hydration in protocells to enhance diverse chemistries such as polymerization reactions. LLPS compartments stimulate diverse chemistries in eukaryotic cells and may function similarly in prokaryotic cells [108–110]. We posit that polyglycine and GADV LLPS compartments were essential features of early protocells. Amyloids and LLPS are posited to have been selected because they supported novel and essential protocell chemistries, partly by regulating hydration and dehydration. We posit that amyloids and LLPS provided the selective driving force for the early evolution of the genetic code before complex proteins could be encoded.

8.6. Other models

Here, we briefly contrast our genetic code evolution model with some other models that have been advanced. Koonin and Novozhilov review

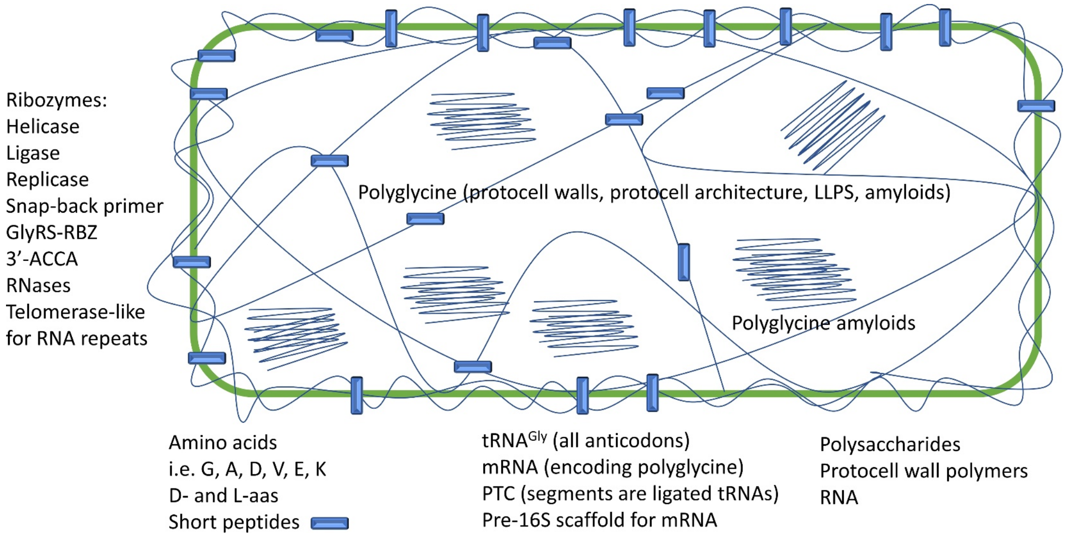


Figure 18. A detailed model for evolution of protocells, polyglycine- and GADV-world. PTC) peptidyl transferase center.

a number of genetic code models [14]. Much emphasis has been given to three models: 1) stereochemical; 2) error-minimization; and 3) coevolution. We do not consider these models to be highly predictive [1]. A new approach has been proposed based on codon energetics [111]. Computational approaches may be of interest [112–114], although the concept of late evolution of degeneracy seems unlikely compared to our model that degeneracy is a natural result of evolution of anticodon reading on the ribosome and evolution of EF-Tu. Kunnev and Gospodinov have proposed models that include RNA-aa linked reactions in pre-life, as we also support [84,85]. Rogers has put forth a model in some ways similar to ours [115]. Another somewhat similar model to ours has been advanced by Chatterjee and Yadav [116]. A different view of serine sectoring than that we propose was recently advanced [117]. Simply stated, our models are more detailed than others and make many more specific predictions. We provide a clear selection strategy and a set of rules for the placements of all amino acids in the standard code. We enrich the discussion of Kunnev and Gospodinov on tRNA – and RNA-linked reactions in the ancient world before the first true cells. Our approach is centered on the tRNA anticodon, and others should adopt the anticodon-centered view. For instance, evolving directly to codons is a mistake, because the genetic code was limited in complexity by the tRNA anticodon [1,3]. Emphasis on the tRNA anticodon and reading of the anticodon on the primitive and modern ribosomes (i.e., +/− EF-Tu) also describes degeneracy of the code. Furthermore, we consider filling the code piecemeal to be a mistake. We take a more orderly approach to code-filling that is based on clear rules for anticodon sequence preference. We provide strong selections for the initial steps of code evolution before complex proteins can be encoded. Our model for tRNA evolution reaches far back into the pre-life world with many predictions for pre-life ribozymes and, surprisingly, ordered pre-life RNA chemistry. Remarkably, existing tRNA sequence provides a record of chemistry in the pre-life world.

Abbreviations

Author Contributions

The authors wrote the paper, made the figures and did the research.

Disclosure of potential conflicts of interest

No potential conflict of interest was reported by the authors.

Funding

This research received no external funding.

ORCID

Zachary Frome Burton http://orcid.org/0000-0003-1065-5222

References

[1] Lei L, Burton ZF. Evolution of life on earth: tRNA, aminoacyl-tRNA synthetases and the genetic code. Life (Basel). 2020;10(3). DOI:10.3390/life10030021

[2] Burton ZF. The 3-minihelix tRNA evolution theorem. J Mol Evol. 2020;88(3):234–242.

[3] Kim Y, Opron K, Burton ZF. A tRNA- and anticodon-centric view of the evolution of aminoacyl-tRNA synthetases, tRNAomes, and the genetic code. Life (Basel). 2019;9(2). DOI:10.3390/life9020037

[4] Kim Y, Kowiatek B, Opron K, et al. Type-II tRNAs and evolution of translation systems and the genetic code. Int J Mol Sci. 2018;19(10):3275.

[5] Pak D, Root-Bernstein R, Burton ZF. tRNA structure and evolution and standardization to the three nucleotide genetic code. Transcription. 2017;8(4):205–219.

[6] Root-Bernstein R, Kim Y, Sanjay A, et al. tRNA evolution from the proto-tRNA minihelix world. Transcription. 2016;7(5):153–163.

[7] Demongeot J, Seligmann H, Rings Strengthen RNA. Hairpin accretion hypotheses for tRNA evolution: a reply to commentaries by Z.F. burton and M. di giulio. J Mol Evol. 2020;88(3):243–252.

[8] Di Giulio M. A comparison between two models for understanding the origin of the tRNA molecule. J Theor Biol. 2019;480:99–103.

[9] Demongeot J, Seligmann H. Codon assignment evolvability in theoretical minimal RNA rings. Gene. 2021;769:145208.

[10] Di Giulio M. An RNA ring was not the progenitor of the tRNA molecule. J Mol Evol. 2020;88(3):228–233.
[11] Demongeot J, Seligmann H. Theoretical minimal RNA rings mimick molecular evolution before tRNA-mediated translation: codon-amino acid affinities increase from early to late RNA rings. C R Biol. 2020;343(1):111–122.
[12] Perona JJ, Gruic-Sovulj I. Synthetic and editing mechanisms of aminoacyl-tRNA synthetases. Top Curr Chem. 2014;344:1–41.
[13] Carter CW Jr., Wills PR. Class I and II aminoacyl-tRNA synthetase tRNA groove discrimination created the first synthetase-tRNA cognate pairs and was therefore essential to the origin of genetic coding. IUBMB Life. 2019;71(8):1088–1098.
[14] Koonin EV, Novozhilov AS. Origin and evolution of the universal genetic code. Annu Rev Genet. 2017;51 (1):45–62.
[15] Perona JJ, Hadd A. Structural diversity and protein engineering of the aminoacyl-tRNA synthetases. Biochemistry. 2012;51(44):8705–8729.
[16] O'Donoghue P, Luthey-Schulten Z. On the evolution of structure in aminoacyl-tRNA synthetases. Microbiol Mol Biol Rev. 2003;67(4):550–573.
[17] Aravind L, Anantharaman V, Koonin EV. Monophyly of class I aminoacyl tRNA synthetase, USPA, ETFP, photolyase, and PP-ATPase nucleotide-binding domains: implications for protein evolution in the RNA. Proteins. 2002;48(1):1–14.
[18] Wolf YI, Aravind L, Grishin NV, et al. Evolution of aminoacyl-tRNA synthetases–analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. Genome Res. 1999;9(8):689–710.
[19] Marin J, Battistuzzi FU, Brown AC, et al. The timetree of prokaryotes: new insights into their evolution and speciation. Mol Biol Evol. 2017;34(2):437–446.
[20] Battistuzzi FU, Feijao A, Hedges SB. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. BMC Evol Biol. 2004;4(1):44.
[21] Long X, Xue H, Wong JT. Descent of bacteria and eukarya from an archaeal root of life. Evol Bioinform Online. 2020;16:1176934320908267.
[22] Burton ZF, Opron K, Wei G, et al. A model for genesis of transcription systems. Transcription. 2016;7(1):1–13.
[23] Burton SP, Burton ZF. The sigma enigma: bacterial sigma factors, archaeal TFB and eukaryotic TFIIB are homologs. Transcription. 2014;5(4):e967599.
[24] Opron K, Burton ZF. Ribosome structure, function, and early evolution. Int J Mol Sci. 2018;20(1):40.
[25] Pak D, Du N, Kim Y, et al. Rooted tRNAomes and evolution of the genetic code. Transcription. 2018;9 (3):137–151.
[26] Pak D, Burton ZF, Burton ZF. Aminoacyl-tRNA synthetase evolution and sectoring of the genetic code. Transcription. 2018;9(3):205–224.
[27] Iyer LM, Aravind L. Insights from the architecture of the bacterial transcription apparatus. J Struct Biol. 2012;179(3):299–319.
[28] Lei L, Burton ZF. Early evolution of transcription systems and divergence of archaea and bacteria. Front Mol Biosci. 2021;8. DOI:10.3389/fmolb.2021.651134
[29] Juhling F, Morl M, Hartmann RK, et al. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. Nucleic Acids Res. 2009;37(Database):D159–162.
[30] Chan PP, Lowe TM. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Res. 2016;44(D1):D184–189.
[31] Chan PP, Lowe TM. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res. 2009;37(Database):D93–97.
[32] Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 1990;18(20):6097–6100.
[33] Pettersen EF, Goddard TD, Huang CC, et al. UCSF ChimeraX : structure visualization for researchers, educators, and developers. Protein Sci. 2021;30(1):70–82.
[34] Goddard TD, Huang CC, Meng EC, et al. UCSF chimeraX: Meeting modern challenges in visualization and analysis. Protein Sci. 2018;27(1):14–25.
[35] Kelley LA, Mezulis S, Yates CM, et al. The phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015;10(6):845–858.
[36] Burton ZF. The old and new testaments of gene regulation. Evolution of multi-subunit RNA polymerases and co-evolution of eukaryote complexity with the RNAP II CTD. Transcription. 2014;5(3):e28674.
[37] Alva V, Dunin-Horkawicz S, Habeck M, et al. The GD box: a widespread noncontiguous supersecondary structural element. Protein Sci. 2009;18(9):1961–1966.
[38] Alva V, Koretke KK, Coles M, et al. Cradle-loop barrels and the concept of metafolds in protein classification by natural descent. Curr Opin Struct Biol. 2008;18 (3):358–365.
[39] Coles M, Hulko M, Djuranovic S, et al. Common evolutionary origin of swapped-hairpin and double-psi beta barrels. Structure. 2006;14 (10):1489–1498.
[40] Coles M, Djuranovic S, Soding J, et al. AbrB-like transcription factors assume a swapped hairpin fold that is evolutionarily related to double-psi beta barrels. Structure. 2005;13(6):919–928.
[41] Iyer LM, Koonin EV, Aravind L. Evolution of bacterial RNA polymerase: implications for large-scale bacterial phylogeny, domain accretion, and horizontal gene transfer. Gene. 2004;335:73–88.
[42] Koonin EV, Krupovic M, Ishino S, et al. The replication machinery of LUCA: common origin of DNA replication and transcription. BMC Biol. 2020;18(1):61.
[43] Madru C, Henneke G, Raia P, et al. Structural basis for the increased processivity of D-family DNA polymerases in complex with PCNA. Nat Commun. 2020;11(1):1591.

[44] The Extended SL. "Two-Barrel" polymerases superfamily: Structure, function and evolution. J Mol Biol. 2019;431(20):4167–4183.

[45] Raia P, Carroni M, Henry E, et al. Structure of the DP1-DP2 PolD complex bound with DNA and its implications for the evolutionary history of DNA and RNA polymerases. PLoS Biol. 2019;17(1):e3000122.

[46] Salgado PS, Koivunen MR, Makeyev EV, et al. The structure of an RNAi polymerase links RNA silencing and transcription. PLoS Biol. 2006;4(12):e434.

[47] Brindefalk B, Dessailly BH, Yeats C, et al. Evolutionary history of the TBP-domain superfamily. Nucleic Acids Res. 2013;41(5):2832–2845.

[48] Demongeot J, Seligmann H. The primordial tRNA acceptor stem code from theoretical minimal RNA ring clusters. BMC Genet. 2020;21(1):7.

[49] Demongeot J, Seligmann H. The uroboros theory of life's origin: 22-nucleotide theoretical minimal RNA rings reflect evolution of genetic code and tRNA-rRNA translation machineries. Acta Biotheor. 2019;67(4):273–297.

[50] Di Giulio M. A comparison among the models proposed to explain the origin of the tRNA molecule: a synthesis. J Mol Evol. 2009;69(1):1–9.

[51] Carter CW Jr. Coding of class I and II aminoacyl-tRNA synthetases. Adv Exp Med Biol. 2017;966:103–148.

[52] Carter CW Jr., Li L, Weinreb V, et al. The Rodin-Ohno hypothesis that two enzyme superfamilies descended from one ancestral gene: an unlikely scenario for the origins of translation that will not be dismissed. Biol Direct. 2014;9(1):11.

[53] Carter CW Jr. Urzymology: experimental access to a key transition in the appearance of enzymes. J Biol Chem. 2014;289(44):30213–30220.

[54] Rodin AS, Rodin SN, Carter CW Jr. On primordial sense-antisense coding. J Mol Evol. 2009;69(5):555–567.

[55] Hartman H, Smith TF. Origin of the genetic code is found at the transition between a thioester world of peptides and the phosphoester world of polynucleotides. Life (Basel). 2019;9(3). DOI:10.3390/life9030069

[56] Smith TF, Hartman H. The evolution of class II aminoacyl-tRNA synthetases and the first code. FEBS Lett. 2015;589(23):3499–3507.

[57] Wetzel R. Evolution of the aminoacyl-tRNA synthetases and the origin of the genetic code. J Mol Evol. 1995;40(5):545–550.

[58] Maracci C, Rodnina MV. Review: translational GTPases. Biopolymers. 2016;105(8):463–475.

[59] Loveland AB, Demo G, Korostelev AA. Cryo-EM of elongating ribosome with EF-Tu*GTP elucidates tRNA proofreading. Nature. 2020;584(7822):640–645.

[60] Loveland AB, Demo G, Grigorieff N, et al. Ensemble cryo-EM elucidates the mechanism of translation fidelity. Nature. 2017;546(7656):113–117.

[61] Rozov A, Demeshkina N, Westhof E, et al. New structural insights into translational miscoding. Trends Biochem Sci. 2016;41(9):798–814.

[62] Rozov A, Westhof E, Yusupov M, et al. The ribosome prohibits the G*U wobble geometry at the first position of the codon-anticodon helix. Nucleic Acids Res. 2016;44(13):6434–6441.

[63] Rozov A, Demeshkina N, Westhof E, et al. Structural insights into the translational infidelity mechanism. Nat Commun. 2015;6(1):7251.

[64] Bernhardt HS, Patrick WM. Genetic code evolution started with the incorporation of glycine, followed by other small hydrophilic amino acids. J Mol Evol. 2014;78(6):307–309.

[65] Bernhardt HS, Tate WP. Evidence from glycine transfer RNA of a frozen accident at the dawn of the genetic code. Biol Direct. 2008;3(1):53.

[66] Ikehara K. Evolutionary steps in the emergence of life deduced from the bottom-up approach and GADV hypothesis (top-down approach). Life (Basel). 2016;6(1). DOI:10.3390/life6010006

[67] Ikehara K. [GADV]-protein world hypothesis on the origin of life. Orig Life Evol Biosph. 2014;44(4):299–302.

[68] Ikehara K. Pseudo-replication of [GADV]-proteins and origin of life. Int J Mol Sci. 2009;10(4):1525–1537.

[69] Oba T, Fukushima J, Maruyama M, et al. Catalytic activities of [GADV]-peptides. Formation and establishment of [GADV]-protein world for the emergence of life. Orig Life Evol Biosph. 2005;35(5):447–460.

[70] Ikehara K. Possible steps to the emergence of life: the [GADV]-protein world hypothesis. Chem Rec. 2005;5(2):107–118.

[71] Burroughs AM, Aravind L. The origin and evolution of release factors: implications for translation termination, ribosome rescue, and quality control pathways. Int J Mol Sci. 2019;20(8):1981.

[72] McGeoch MW, Dikler S, McGeoch JEM. Hemolithin: a meteoritic protein containing iron and lithium. arXiv. 2020. arXiv:2002.11688v1. arXiv:2002.11688v1 [astro-ph.EP], [astro-ph.EP].

[73] Ikehara K, Niihara Y. Origin and evolutionary process of the genetic code. Curr Med Chem. 2007;14(30):3221–3231.

[74] Ikehara K. The origin of tRNA deduced from pseudomonas aeruginosa 5 anticodon-stem sequence: Anticodon-stem loop hypothesis. Orig Life Evol Biosph. 2019;49(1–2):61–75.

[75] Ikehara K, Omori Y, Arai R, et al. A novel theory on the origin of the genetic code: a GNC-SNS hypothesis. J Mol Evol. 2002;54(4):530–538.

[76] Rafels-Ybern A, Torres AG, Camacho N, et al. The expansion of inosine at the wobble position of tRNAs, and its role in the evolution of proteomes. Mol Biol Evol. 2019;36(4):650–662.

[77] Rafels-Ybern A, Torres AG, Grau-Bove X, et al. Codon adaptation to tRNAs with Inosine modification at

position 34 is widespread among Eukaryotes and present in two Bacterial phyla. RNA Biol. 2018;15 (4–5):500–507.

[78] Demongeot J, Norris V. Emergence of a "Cyclosome" in a primitive network capable of building "Infinite" proteins. Life (Basel). 2019;9(2). DOI:10.3390/life9020051

[79] Saint-Leger A, Bello C, Dans PD, et al. Saturation of recognition elements blocks evolution of new tRNA identities. Sci Adv. 2016;2(4):e1501860.

[80] Fournier GP, Alm EJ. Ancestral reconstruction of a Pre-LUCA aminoacyl-tRNA synthetase ancestor supports the late addition of Trp to the genetic code. J Mol Evol. 2015;80(3–4):171–185.

[81] Hauenstein SI, Perona JJ. Redundant synthesis of cysteinyl-tRNACys in methanosarcina mazei. J Biol Chem. 2008;283(32):22007–22017.

[82] Turanov AA, Xu XM, Carlson BA, et al. Biosynthesis of selenocysteine, the 21st amino acid in the genetic code, and a novel pathway for cysteine biosynthesis. Adv Nutr. 2011;2(2):122–128.

[83] Mukai T, Crnkovic A, Umehara T, et al. RNA-dependent cysteine biosynthesis in bacteria and archaea. mBio. 2017;8(3). DOI:10.1128/mBio.00561-17

[84] Gospodinov A, Kunnev D. Universal codons with enrichment from GC to AU nucleotide composition reveal a chronological assignment from early to late along with LUCA formation. Life (Basel). 2020;10(6). DOI:10.3390/life10060081

[85] Kunnev D, Gospodinov A. Possible emergence of sequence specific RNA aminoacylation via peptide intermediary to initiate darwinian evolution and code through origin of life. Life (Basel). 2018;8(4). DOI:10.3390/life8040044

[86] Rampias T, Sheppard K, Soll D. The archaeal transamidosome for RNA-dependent glutamine biosynthesis. Nucleic Acids Res. 2010;38(17):5774–5783.

[87] Wu J, Bu W, Sheppard K, et al. Insights into tRNA-dependent amidotransferase evolution and catalysis from the structure of the Aquifex aeolicus enzyme. J Mol Biol. 2009;391(4):703–716.

[88] Feng L, Sheppard K, Tumbula-Hansen D, et al. Gln-tRNAGln formation from Glu-tRNAGln requires cooperation of an asparaginase and a Glu-tRNAGln kinase. J Biol Chem. 2005;280(9):8150–8155.

[89] Feng L, Sheppard K, Namgoong S, et al. Aminoacyl-tRNA synthesis by pre-translational amino acid modification. RNA Biol. 2004;1(1):16–20.

[90] Tumbula-Hansen D, Feng L, Toogood H, et al. Evolutionary divergence of the archaeal aspartyl-tRNA synthetases into discriminating and nondiscriminating forms. J Biol Chem. 2002;277(40):37184–37190.

[91] Min B, Pelaschier JT, Graham DE, et al. RNA-dependent amino acid biosynthesis: an essential route to asparagine formation. Proc Natl Acad Sci U S A. 2002;99(5):2678–2683.

[92] Raczniak G, Becker HD, Min B, et al. A single amidotransferase forms asparaginyl-tRNA and glutaminyl-tRNA in Chlamydia trachomatis. J Biol Chem. 2001;276(49):45862–45867.

[93] Salazar JC, Zuniga R, Raczniak G, et al. A dual-specific Glu-tRNA Gln and Asp-tRNA Asn amidotransferase is involved in decoding glutamine and asparagine codons in Acidithiobacillus ferrooxidans. FEBS Lett. 2001;500 (3):129–131.

[94] Numata T. Mechanisms of the tRNA wobble cytidine modification essential for AUA codon decoding in prokaryotes. Biosci Biotechnol Biochem. 2015;79 (3):347–353.

[95] Satpati P, Bauer P, Aqvist J. Energetic tuning by tRNA modifications ensures correct decoding of isoleucine and methionine on the ribosome. Chemistry. 2014;20 (33):10271–10275.

[96] Voorhees RM, Mandal D, Neubauer C, et al. The structural basis for specific decoding of AUA by isoleucine tRNA on the ribosome. Nat Struct Mol Biol. 2013;20(5):641–643.

[97] Osawa T, Kimura S, Terasaka N, et al. Structural basis of tRNA agmatinylation essential for AUA codon decoding. Nat Struct Mol Biol. 2011;18 (11):1275–1280.

[98] Phillips G, De Crecy-lagard V. Biosynthesis and function of tRNA modifications in Archaea. Curr Opin Microbiol. 2011;14(3):335–341.

[99] Ikeuchi Y, Kimura S, Numata T, et al. Agmatine-conjugated cytidine in a tRNA anticodon is essential for AUA decoding in archaea. Nat Chem Biol. 2010;6 (4):277–282.

[100] Mandal D, Kohrer C, Su D, et al. Agmatidine, a modified cytidine in the anticodon of archaeal tRNA(Ile), base pairs with adenosine but not with guanosine. Proc Natl Acad Sci U S A. 2010;107(7):2872–2877.

[101] Schmitt E, Coureux PD, Kazan R, et al. Recent advances in archaeal translation initiation. Front Microbiol. 2020;11:584152.

[102] Longo LM, Despotovic D, Weil-Ktorza O, et al. Primordial emergence of a nucleic acid-binding protein via phase separation and statistical ornithine-to-arginine conversion. Proc Natl Acad Sci U S A. 2020;117(27):15731–15739.

[103] Vollmer W. Structural variation in the glycan strands of bacterial peptidoglycan. FEMS Microbiol Rev. 2008;32(2):287–306.

[104] Vollmer W, Blanot D, De Pedro MA. Peptidoglycan structure and architecture. FEMS Microbiol Rev. 2008;32(2):149–167.

[105] Singh V, Xu L, Boyko S, et al. Zinc promotes liquid-liquid phase separation of tau protein. J Biol Chem. 2020;295(18):5850–5856.

[106] Elbaum-Garfinkle S. Matter over mind: liquid phase separation and neurodegeneration. J Biol Chem. 2019;294(18):7160–7168.

[107] Puzzo D, Argyrousi EK, Staniszewski A, et al. Tau is not necessary for amyloid-beta-induced synaptic and memory impairments. J Clin Invest. 2020;130(9):4831–4844.

[108] Portz B, Shorter J. Switching condensates: The CTD code goes liquid. Trends Biochem Sci. 2020;45(1):1–3.

[109] Boehning M, Dugast-Darzacq C, Rankovic M, et al. RNA polymerase II clustering through carboxy-terminal domain phase separation. Nat Struct Mol Biol. 2018;25(9):833–840.

[110] Guo C, Che Z, Yue J, et al. ENL initiates multivalent phase separation of the super elongation complex (SEC) in controlling rapid transcriptional activation. Sci Adv. 2020;6(14):eaay4858.

[111] Klump HH, Volker J, Breslauer KJ. Energy mapping of the genetic code and genomic domains: implications for code evolution and molecular Darwinism - CORRIGENDUM. Q Rev Biophys. 2020;53:e14.

[112] Yarus M. Evolution of the standard genetic code. J Mol Evol. 2021. DOI:10.1007/s00239-020-09983-9

[113] Yarus M. Optimal evolution of the standard genetic code. J Mol Evol. 2021. DOI:10.1007/s00239-020-09984-8

[114] Yarus M. Crick wobble and superwobble in standard genetic code evolution. J Mol Evol. 2021. DOI:10.1007/s00239-020-09985-7

[115] Rogers SO. Evolution of the genetic code based on conservative changes of codons, amino acids, and aminoacyl tRNA synthetases. J Theor Biol. 2019;466:1–10.

[116] Chatterjee S, Yadav S. 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). 2019;9(1). DOI:10.3390/life9010025

[117] Inouye M, Takino R, Ishida Y, et al. Evolution of the genetic code; Evidence from serine codon use disparity in Escherichia coli. Proc Natl Acad Sci U S A. 2020;117 (46):28572–28575.