Title: Coding triplets in the transfer RNA arm and their role in present and past tRNA recognition

Authors: Ilana Agmon[a,b] E-Mail: chilana@technion.ac.il

Itay Fayerverker[c] E-Mail: fire@cs.technion.ac.il

Tal Mor[c] E-Mail: talmo@cs.technion.ac.il

Authors affiliation:

a. Institute for Advanced Studies in Theoretical Chemistry, Schulich Faculty of Chemistry - Technion - Israel Institute of Technology, Haifa 3200003, Israel

b. Fritz Haber Research Center for Molecular Dynamics, Hebrew University Jerusalem, 9190401, Israel

c. Department of Computer Science, Technion 3200003, Haifa, Israel

Corresponding author: Ilana Agmon

Schulich Faculty of Chemistry, Technion, Haifa 3200003, Israel

Tel: +972 4 8293824

E-Mail: chilana@technion.ac.il

Abstract

Background

The evolutionary pathway of tRNA established the linking of the genetic code and protein synthesis, thus, revealing the characteristics of the prebiotic tRNA is crucial for presenting a feasible scenario for the emergence of life.

Results

Analysis of bacterial tRNA sequences reveals that for nine amino acids, mostly those considered to be the most ancient ones, the acceptor-TΨC arm carries cognate coding triplets far beyond statistical expectation. The contemporary determinants for the recognition of tRNAs and minihelices by the cognate synthetases and by EF-Tu, are part of these conserved triplets. We also find a strong correlation between the extent of the coding triplet conservation and amino acids whose contemporary synthetases belong to class IIa.

Conclusions

Identification of conserved coding triplets within the tRNA acceptor-TΨC stem, along with the analysis of their distinctive layout, is suggestive of a continuous evolutionary path that could have led to the appearance of the early translation system.

Keywords: tRNA, tRNA aminoacylation, Origin of the genetic code, Translation, Origin of life, Synthetase

Background

tRNA is a critical component of the translation system due to its role in linking the genetic code with the synthesized protein. The modern tRNA is an L-shaped molecule. One end of the molecule is the universally conserved single stranded NCCA tail which carries the cognate amino acid (aa), while at the other end, about 75Å away, the anticodon (AC) loop that fully characterizes the attached amino acid is located. The mature tRNA interacts with three key components of the translation system - the cognate aminoacyl-tRNA-synthetase (aaRS) that specifically aminoacylates it; the elongation factor (EF-Tu in bacteria) that accommodates the aa-tRNAs and carries them, in a ternary complex, to the ribosome; and the ribosome where the attached amino acid is incorporated into the growing polypeptide.

Recognition of aa-tRNA on the ribosome is accomplished via the interactions of the mRNA codon accommodated on the small subunit, with the AC loop. EF-Tu, which carries a variety of different tRNAs, recognizes each of them specifically [1]. The minimal fragment of tRNA that can interact efficiently with EF-Tu, consists of a 10 base-pair helix from the acceptor-TΨC stem, nucleotides 63-72, linked to the aa-3'end tail [1], from which nucleotides 63–67 form the contact interface [2].

Aminoacyl-tRNA-synthetase typically acquires the identity of the cognate amino acid from the anticodon stem-loop [3,4], from the discriminator base (N73) and from specific nucleotides in the acceptor stem of the tRNA, named identity determinants [5,6,7,8]. Minimal RNA substrates comprising 4-12 base-pairs, derived from the acceptor-TΨC stem of the tRNA, were shown to be specifically aminoacylated by their cognate synthetases [5,6,7,9,10,11], even though the main provider of identity information, i.e. in the AC loop, was absent. This non-AC aminoacylation process was proposed to be controlled by an “operational code” [6] that relates few nucleotides in the acceptor stem to specific amino acids. The major recognition determinants for this enzymatic process were found to lie in the first five base-pairs (nucleotides 68-72) of the acceptor stem [7,9,11]. They were proposed to be the remnants of an early mechanism of specifically aminoacylating proto-tRNAs, possibly associated with a second code [11,12], distinct from the nucleotide triplets of the standard genetic code [6,11].

The possible existence of non-enzymatic aminoacylation process, which could have preceded the enzymatic charging, is of significance because it could have provided aa-tRNAs for the synthesis of the initial coded peptides, prior to the appearance of the proto-synthetase. Such processes were demonstrated experimentally: Hairpins were aminoacylated with their cognate amino acids by reaction with aminoacyl-adenylates in the presence of a dipeptide, where the specificity was likely to be provided by the codon at the 5'end [13]. In a different line of research, aminoacylation of tRNAs with their cognate amino acids, in the absence of the specific aminoacyl-tRNA synthetase and ATP, was reported to occur at high pressure of 6 kbar [14,15]. The aa-tRNAs obtained were shown to bind to ribosomes and participates in the synthesis of a polypeptide in vitro. Specificity was assumed to be encoded in a nucleotide sequence, probably of the amino acid acceptor stem.

Here we analyze sequence data from bacterial tRNA genes and report the occurrence, to large extent, of cognate coding triplets in the acceptor-T□C arm of the modern bacterial tRNAs belonging to nine amino acids (the "conserving" amino acids), and to a very high extent in six of these nine amino acids (the "robustly conserving" amino acids"): Ala, Asp, Gly, His, Pro and Ser. High correlation is found between the 9 conserving amino acids and the amino acids believed to be among the most ancient ones [Refs16,17]. High correlation is also found between the robustly conserving amino acids and the amino acids that are charged by class IIa synthetases [Ref23]. In contrast, for five amino acids (the "avoiding" amino acids), the string lacks the cognate triplets far beyond statistical expectation.

The highly conserved coding triplets of six amino acids, which accommodate the identity determinants for the recognition of minihelices and tRNAs by their cognate synthetases and by EF-Tu, are proposed to be the precursors of the modern recognition determinants. Their possible involvement in a non-enzymatic aminoacylation of proto-tRNAs, which could have preceded the enzymatic process, is suggestive of an evolutionary scheme that resolves the chicken and egg paradox concerned with the appearance of translation, where a synthetase is required for charging tRNAs, but charged tRNAs are required for the formation of synthetases [18]. The conserved coding triplets, assumed to have a prebiotic origin, constitute a subset of the contemporary codon set, suggesting the significance of this subset of the standard codon table in the emergence of life as we know it.

Results

In bacterial tRNAs, the sequence of the string preceding the 3'end (pre-3'end), ranging from nucleotide 63 to 72 is diverse [19] even among different strains of the same bacterium (Tab. 1). Yet, when read from the 5' to the 3', this string holds codon and/or anticodon triplets (termed here C/ACs) of its cognate amino acid in the tRNAs of about half of the amino acids, while the cognate C/ACs are absent from this string for other amino acids (Tab. 1, 2, additional file 1 Tab. S1 and additional file 2 DataSet S1).

Table 1. Pre-3'end sequences

Amino acid (codons/anti-codons)	Pre-3'end sequence (nucleotides 63-72)**	Organism	Code in tRNAdb*
Ala (Conserved codons-GCC,GCU)	GCUAGGCUCC	*Bacillus subtilis*	tdbR00000007
	GCUUAGCUCC	*Escherichia coli*	tdbR00000008
	GCAUAGCUCC	*Escherichia coli*	tdbR00000010
	GUUAGGGUCC	*Mycoplasma capricolum*	tdbR00000005
Pro (Conserved codon-CCG)	UGUCUUCCCG	*Bacillus subtilis*	tdbR00000316
	UCUAUCACCG	*Escherichia coli*	tdbR00000623
	UCUCGUGCCG	*Salmonella typhimurium*	tdbR00000318
	UGUCUUCCCG	*Mycoplasma mycoides*	tdbR00000413
Gly (Conserved Anticodons-UCC, CCC)	CAUCACCCGC	*Bacillus subtilis*	tdbR00000116
	CUUCGCCCGC	*Escherichia coli*	tdbR00000117
	CGUUUCCCGC	*Escherichia coli*	tdbR00000118
	CGCUGCCCGC	*Salmonella typhimurium*	tdbR00000121
	UAUCUUCCGC	*Staphylococcus_epidermidis*	tdbR00000114
	CGUCACCCGC	*Streptomyces_coelicolor_A3*	tdbR00000113
	CUUCACCUGC	*Mycoplasma capricolum*	tdbR00000111
Tyr (Codons-UAU, UAC; AC-AUA,GUA)	GCCCCCCUCC	*Bacillus subtilis*	tdbR00000543
	UUCCCCCACC	*Escherichia coli*	tdbR00000545
	GUCCCCCUCC	*Geobacillus stearotherm.*	tdbR00000547
	CUCCCCCUCC	*Mycoplasma capricolum*	tdbR00000542

** All the unique sequences existing for the elongator-tRNAs of Ala, Pro, Gly and Tyr, as found in the tRNAdb database [20,21].*

***The conserved codons are framed and the anticodons underlined. In $tRNA^{Ala}$ the location of the conserved coding triplets varies, in $tRNA^{Pro}$ (codon) and $tRNA^{Gly}$ (anticodon) it keeps a fixed position and in $tRNA^{Tyr}$ the cognate C/ACs are absent. The Mycoplasma (genus of the Mollicutes class) sequences, at the bottom of each box, often do not comply with the other bacteria.*

Sequences of the pre-3'end strings were acquired from the tRNAdb database[20,21], which holds 128 bacterial tRNA sequences obtained from elongator-tRNAs involved in protein synthesis, and over 6200 bacterial tRNA genes, belonging to the 20 canonical amino acids. To avoid the inclusion of pseudo-tRNA genes, i.e. those not coding for elongator-tRNAs, only tRNA sequences having a full CCA 3'end and a consensus discriminator base [8] (N73) were incorporated into the statistics (Methods), retaining only 60% of the original sequences. The exclusion of one class of bacteria, the Mollicutes, which does not share the statistical behavior of the other bacteria (Tab. 1, additional file 1 Tab. S2), resulted in a sample of 3324 tRNA gene-sequences from over 100 bacterial species. Statistical results that took into account the 350 tRNA-genes of Mollicutes (additional file 1 Tab. S3), didn't differ significantly from the data sample used in the current study.

The occurrence of the cognate codons and anticodons in the pre-3'end strings of the 20 canonical amino acids was determined and compared with their expected occurrences, as obtained from randomized 10-mer oligonucleotides simulation. The construction of the randomized sequences and the estimation of the expected occurrence of a specific set of C/ACs are described in the Methods section. The randomized strings were constructed with three nucleotide distributions: 1. Distribution identical to the average found in all the bacterial pre-3'end strings analyzed ("sample" distribution), i.e.: A=10%, C=46%, G=24%, T=20%; 2. Equal distribution of the four nucleotides; 3. The individual nucleotide distribution found in the pre-3'end string of each amino acid (additional file 1 Tab. S1). Amino acids whose pre-3'end strings significantly conserved or lacked the cognate C/ACs ($P<0.01$), when compared with the expected occurrence values obtained with all three distributions, were classified as "conserving" or "avoiding" (Tab. 2, additional file 1 Tab. S1). The cognate C/ACs conserved in the pre-3'end string are often not correlated with the anticodon triplet in the AC loop (additional file 2 DataSet S1) and they typically constitute only a subset of the standard C/ACs table of each amino acid.

The conserved cognate C/ACs are present in 81-99% of the pre-3'end strings belonging to Ala, Arg, Asp, Gly, His, Leu, Pro, Ser and Val, far beyond statistical expectation. In contrast, the pre-3'end strings of the remaining amino acids exhibit the cognate C/ACs in 0-40% of their sequences, which, for Gln, Glu, Lys, Met and Tyr is less than statistically expected. For the remaining six amino acids the occurrences apparently accord with the statistical expectations

(Tab. 2 and additional file 1 Tab. S1). To further benchmark these statistical results, the occurrence of the triplet CCC, which is expected to be the most prevalent in 14 out of 20 amino acids due to the excess of C in their sequences, was examined. Strings from the conserving amino acids display CCC considerably less than inferred from their nucleotide distribution and surprisingly, Ala, His and Asp, hardly show any CCC triplets in their strings, while the statistically most avoiding amino acid, Gln, displays the most significant access of CCC (additional file 1 Fig. S1).

Table 2: Statistics of the C/ACs occurrence in the pre-3'end of bacterial tRNA genes

Amino acid (no. sequences)	Conserved C/ACs set*		Observed Occurrence **	Expected occurrence *** (±0.8%)	*P*
Conserving amino acids	codons	anticodons			
Ala (209)	**GCU**, GCC		83 %	50 %	E-22
Asp (110)		**GUC**	96 %	17 %	E-106
Gly (243)		**CCC,UCC**	94 %	57 %	E-31
His (83)	**CAC, CAU**		90 %	22 %	E-51
Pro (180)	**CCG**		98 %	37 %	E-64
Ser (290)	**UCA,UCC,UCU**		99 %	46 %	E-72
Arg (305)	CGN, AGG		92 %	68 %	E-20
Leu (366)	CUN		81 %	57 %	E-20
Val (201)	GUN	NAC	88 %	58 %	E-17
Avoiding amino acids	Avoided cognate C/ACs Codons ; Anticodons		Observed Occurrence **	Expected Occurrence *** (±0.8%)	*P*
Gln (120)	CAA,CAG ; CUG,UUG		8 %	35 %	E-9
Glu (113)	GAA,GAG ; CUC,UUC		22 %	45 %	E-6
Lys (122)	AAA,AAG ; CUU,UUU		0 %	19 %	E-7
Met (218)	AUG ; CAU		4 %	9 %	E-3
Tyr (98)	UAU,UAC ; AUA,GUA		2 %	12 %	E-3

** Cognate triplets whose conservation is considered as robust (Methods) are depicted by bold letters, along with their corresponding amino acids.*

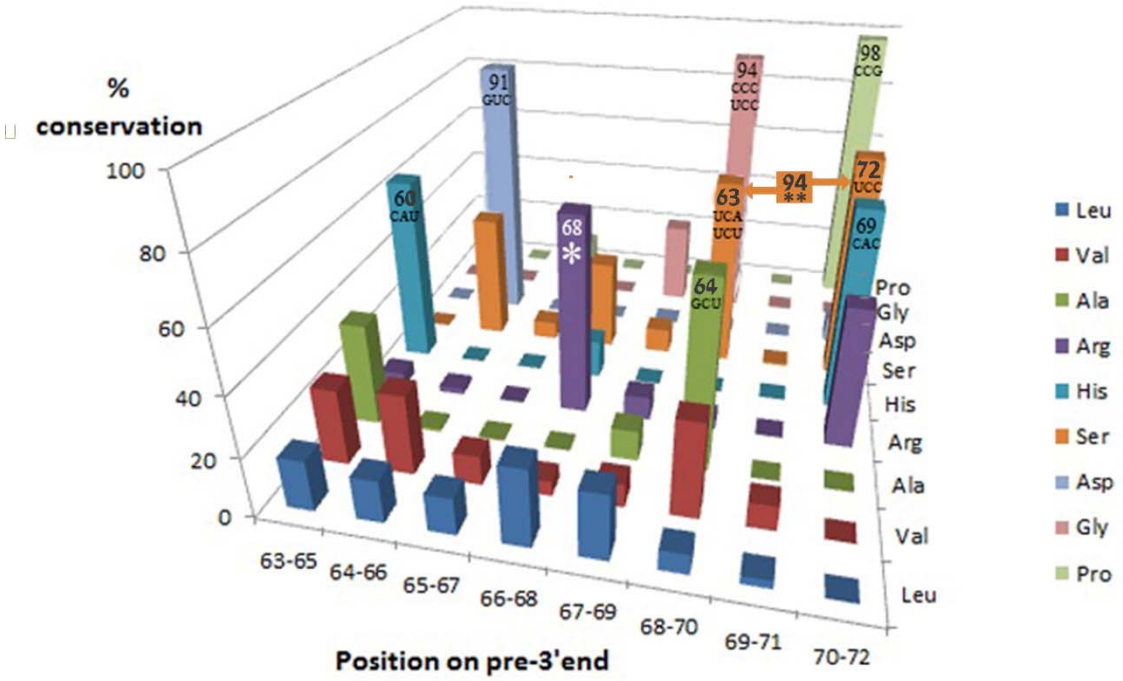
*** % of pre-3'end strings holding at least one C/AC from their conserved set at any location on the string.*

**** Derived from random oligonucleotides simulation with "sample" distribution (A,C,G,U)=(10%,46%,24%,20%)*

N- any nucleotide. P- the probability that the observed occurrence of a particular C/AC set is random. P ≤ 0.01 is considered statistically significant.

The C/ACs in the pre-3'end string of the conserving amino acids may appear as a single site which is fully conserved although the rest of the sequence varies (Pro, Tab. 1, Fig. 1) or as an accumulation of occurrences from as many as seven different sites (Leu, Val, Fig. 1). The conserved C/ACs of Ala, Asp, Gly, His, Pro and Ser are concentrated in one or two sites, each holding one or two coding triplets. In contrast, the overall high occurrence of the conserved C/ACs of Leu, Val and Arg is the sum of small occurrences scattered throughout their pre-3'end strings (Leu, Val) or of five different codons largely concentrated in a single position (Arg).

Figure 1: Occurrence of the conserved codons and/or anticodons in the pre-3'end string of the conserving amino acids.



The column of each "C/AC conservation site" is marked by the identity of its conserved triplets and by the percentages of sequences carrying these triplets in the site.

**- Arg codons CGG, AGG, CGA, CGT.*

***- The overall occurrence of Ser conserved C/ACs in its two conservation sites (positions 68-70 and 70-72).*

Components of a conserved set are considered as robustly conserved when they appear, alone or with only one another coding triplet, in a specific location in the majority of the pre-3'end strings belonging to a certain amino acid. This location is termed a "C/AC conservation site". Overall the occurrence histogram (Fig. 1) contains eight such sites: one for Pro (positions 70-72, 98% occurrence of the codon CCG), Gly (positions 68-70, 94% occurrence of the anticodons CCC or UCC), Asp (positions 64-66, 91% occurrence of the anticodon GUC), Ala (positions 68-70, 64% occurrence of the codon GCU. In the remaining sequences only C69 and U70 are conserved), and two for Ser (in two partially overlapping positions; 68-70, 63% occurrence of the codons UCU or UCA and positions 70-72, 72% occurrence of UCC, which combine to 94% occurrence of a conserved codon at positions 68-72) and His (positions 63-65, 60% occurrence of the codon CAU. In the remaining sequences only C63 and U65 are conserved; Positions 70-72, 69% occurrence of the codon CAC. In the remaining sequences only C70 and C72 are conserved).

To benchmark the resulting list of conserving aa's and robustly conserving aa's, and the resulting high correlations with aa's being charged by class IIa synthetase on the other hand, we add two steps to our analysis: The first step is a second method, named here "the alternative method" (see Methods), for finding conserving and robustly conserving aa's. The second step is to prove high correlations between our list of robustly conserving aa's and aa's charged by Class IIa synthetase; we leave precise analysis of correlation to anciency (namely, ancientness or antiquity) of aa's to future work.

The alternative method differs from the main one in three ways: First, it counts appearances of triplets per site, and order these in a list, listing at the top of the list - the top appearing ones, etc. See additional file 2 DataSet S4. Second, it does not involve a removal of strings; namely, if a specific triplet appears several times (each in a different location) in one string, each appearance contributes to the counting. Last, as this alternative method is mainly for benchmarking by comparing to the main method, we make it as simple as possible: 1.- We use the uniform distribution of 25% per nucleotide; note that this is similar to one of the statistical analysis of

the main method - shown in Tab. s1, and also it is similar to the analysis in [Ref22]), namely we do not use here the "sample distribution". 2.- We only analyze conservation and not avoidance. 3.- We do not repeat proving that the strings are extremely far from random.

Using the alternative method we obtain exactly the same six robustly conserving aa's, as when using the main method. However, we obtain only Val and Arg as "conserved yet not robustly conserved", namely the alternative method – with the choice of 50% and 20% as thresholds (see Methods) removes Leu from the list of conserving aa's. We note that relatively small changes of the thresholds used in the alternative method could lead to two interesting results: Modifying the result for Arg so that this aa becomes robustly conserving if threshold for robust conservation is changed from 50% to 45% (or even to 49%). And adding back Leu to the list of conserving aa's, however, along also with adding Trp (and respectively also adding Thr) if threshold for conservation is modified from 20% to 14% (and respectively to 12%). To the other direction, note that even if we were to increase the threshold for robust conservation to 60%, the list of robustly conserving aa's will not be modified.

Similar conclusion for Arg being at the border between conserving and robustly conserving, and Leu being at the border of being non-conserving, may also be observed using small changes of the thresholds of the main method (as can be seen from looking at the observed occurrence in Tab. 2). We summarize the results in Tab. 3.

TABLE 3: The list obtained using the alternative assignment method

	Main method	Alternative method				
	Conserved set	Conserved set*		Percentage of appearance**	The relevant site	"Relevant" non-coding triplets***
Ala	GCU, GCC	GCU [GCU,AGC,GGC]		64%	68-70	2
Asp	GUC	GUC [GAC]		91%	64-66	0
Gly	CCC,UCC	CCC [UCC,ACC,GCC]		78%	68-70	2
His	CAC,CAU	CAC [CAU]		69%	70-72	0
Pro	CCG	CCG [CCC,CCC]		98%	70-72	0
Ser	UCA,UCC,UCU	UCC [UCU, UCU,UCU]		74%	70-72	0
Arg	CGN,AGG	CGG [GCG,CGC,UCG,CCG]		49%	66-68	1
Val	GUN	GTC		23%	63-65	6
Leu	CUN	CUU (may be deleted from list)		14%	66-68	11

This table compares the Results of the main method and the alternative method. In bold we note the robustly conserved coding triplets, and the robustly conserving aa's.

**For the Conserved C/AC set, out of parenthesis appears the most conserved C/AC.*

***Percentage of appearance stands for the percentage of 10-mer strings containing the most appearing coding triplet (per its site), while that relevant site is presented in the next column.*

****The last column presents the number of non-coding triplets that are more common than the most common coding triplet.*

For example, let us observe more closely a few lines in the Tab. 3: For Pro and Asp, the most popular triplets are coding ones, appearing in 98% and 91% of the strings respectively, and the relevant robustly conserved C/AC are CCG for Pro and GUC for Asp. As a different example, for Val, the most appearing coding triplet is GTC appearing (in location 63-65) in 23% of the strings, and 6 non-coding triplets appear more than 23% times. We see that Leu ought to be removed from the list according to the alternative method, due to appearing in less than 20% of the

strings; we may also notice that many non-coding triplets appear more than the coding triplets, for Leu.

As a result of the differences in defining what aa's ought to be in the list, the alternative method does not lead to the same conserving aa's (and obviously, not to the same conserved coding triplets) as the main method. We may also observe that a change of threshold for robustly conserved C/AC would potentially make Arg a robustly conserving aa.

To connect the resulting list of robustly conserving aa's to aa's being charged by class IIa synthetase is trivial: Except for Asp, the other top five in our table are all charged by class IIa synthetase. For a more formal description of the correlation we apply a statistical test, and we consider its resulting outcome as significant if the P-value is below 0.005. Formally, we know that 6 out of 20 aa's are aminoacylated by type IIa synthetase, and we name their list as list-A. We found k robustly conserving aa's; we found that k=6 if the threshold is 50% (or increased to 60%), and k=7 if threshold is decreased (say to 45%). We analyze here only the P-value in the case of k=6 (namely not including Arg in the robustly-conserving list), using a statistical test counting the total number of robustly conserving aa's in list-A:
Our result is that 5 out of the 6 aa's in list-A are robustly conserving. The (one-sided) probability of uniformly distributed choice of 6 (random) aa's out of the 20 to include 5 or 6 of the 6 robustly conserving aa's is exactly (1+6*14)/(20 choose 6) = 85/38760=1/456~0.00219. Namely, this number is the resulting P-value, proving, far beyond statistical significance of 0.005, that our null hypothesis of no correlations between the two lists is wrong.

Let us add a brief explanation: There are (20 choose 6) possibilities to choose 6 aa's out of 20. Only one of these choices will contain exactly all the 6 type IIa sythetase aminoacylated aa's, another 6*14=84 of these choices will contain 5 out of the 6 type IIa sythetase aminoacylated aa's (as there are 6 options for "which type IIa sythetase aminoacylated aa's to omit" and 14 options for its replacement in the 6 aa's we choose).

To connect the resulting robustly conserving list to the antiquity of aa's is more involved, and is left for future work: In brief, except for His, the other top five in our robustly conserving list are also among the six oldest aa's according to convention [Ref 16, Table V], and the top eight in our

list of nine conserving aa's, are among the oldest in [16]; in particular, Gly, Ala, Asp, and Val are the oldest four according to [16,17].

The genomic tRNA sequences of archaea and eukarya were not subjected to the aforementioned analysis due to the difficulty to distinguish pseudo-tRNA gene-sequences, which are prevalent in the database (Sprinzl, M., private communication), from the authentic ones. The 3'end tail – NCCA – used for screening the bacterial data, which excluded 40% of the bacterial sequences, appears only in few archaeal tRNA-genes and is absent in eukaryotes. The more reliable elongator-tRNA sequences are too scarce to ensure any statistical validity. Part of the 63 unique archaeal and 117 unique eukaryotic sequences, belonging to the tRNAs of the 20 amino acids, do display access of cognate C/ACs in their pre-3'end strings (additional file 2 DataSets S2, S3), mostly for the same amino acids classified as conserving in bacteria. However, the position and identity of these cognate triplets differ from bacteria, leaving open the question of whether the cognate C/ACs are conserved in the other realms of life besides bacteria.

Discussion

1. Coding triplets in the acceptor-TΨC stem

In total, over 90% of the bacterial pre-3'end strings from the conserving amino acids contain a conserved codon and/or anticodon triplet (Tab. 2), although the data probably includes unidentified pseudo-tRNA genes randomly carrying the consensus discriminator base, and possibly sequences from additional species besides the Mollicutes, whose tRNAs do not comply with the bacterial pattern of C/ACs conservation. Moreover, tRNAs from the three most populated bacterial genera in the data, i.e. *Bacillus*, *E. coli*, and *Salmonella* (comprising 20% of the data; 659 sequences altogether), having varying phylogeny, exhibit complete conservation

(99% occurrence) and avoidance (0% occurrence) of the cognate C/ACs in the pre-3'end strings of the conserving and avoiding amino acids, respectively (additional file 1 Tab. S2). Coding triplets, which are complementary to the conserved C/ACs presented here, were previously identified at positions 3-5 of prokaryotic tRNAs from four of the conserving amino acids, i.e. ala, Asp, Gly and Val [22]. Collectively, these observations point to a highly non-random occurrence of conserved C/ACs, distributed along the pre-3'end strings of the conserving amino acid.

Nucleotide C is the most abundant (53%) of the nucleotides constituting the 10 robustly-conserved C/ACs, while within the avoided C/ACs, nucleotides A and U are frequent (33% each) (Tab. 2). The conservation and avoidance are therefore facilitated by the elevated occurrence of nucleotide C (46%) and reduced occurrence of A (10%) in the bacterial pre-3'end sequences. Yet, the C/ACs occurrence in the pre-3'end strings of the nine conserving and the five avoiding amino acids is of consequence with regard to both the expectation values derived from the randomized oligonucleotides constructed with the C-rich and A-poor "sample" distribution (Tab. 2) and with the individual nucleotide distributions (additional file 1 Tab. S1). Thus, the conservation and avoidance of the cognate codons and/or anticodons in the pre-3'end string is a genuine phenomenon and not an artifact of the nucleotide distribution. Correspondingly, the triplet CCC, which is expected to be highly frequent due to the significant access of C nucleotides in the pre-3'end of 14 amino acids, displays CCC abundantly for avoiding amino acids, while CCC is nearly absent in the sequences of the robustly conserving amino acids (additional file 1 Fig. S1), being probably correlated with the marked occurrence of the conserved C/ACs.

For the six robustly conserving amino acids, i.e. for Ala, Asp, Gly, His, Pro and Ser, the conserved C/ACs are mainly located in the C/AC conservation sites (Fig. 1), by definition. All of their tRNAs are charged by class IIa synthetases [23], except from $tRNA^{Asp}$ which is charged by class IIb synthetase. Similarly, all of them except $tRNA^{Asp}$, exhibit conservation sites in the range 68-72, which forms a contact interface with the synthetase. In the data of Gly, Pro and Ser the conservation sites show essentially full occupancy (94-98% occurrence), but for Ala and His, although a large majority of their pre-3'end sequences contain a coding triplet (83% and 90%, respectively), only about 2/3 of the sequences have a full codon located at their conservation sites (Fig. 1). In the remaining third of the data, only two out of the three coding letters are fully conserved at the conservation sites, possibly reflecting the involvement of distinct nucleotides

in the contemporary aminoacylation process, either directly as identity determinants, or indirectly by providing a certain sequence contexts essential for the tRNA recognition [6].

At the other half of the pre-3'end string, nucleotides 63–67 from the TΨC-acceptor stem establish the contact interface with EF-Tu [2]. The C/AC conservation sites in that range, i.e. the Asp codon GUC at positions 64-66, which is conserved in 91% of the sequences and the His codon CAU at positions 63-65, which is conserved in 60% of the sequences (Fig. 1), are suggested to contribute to the binding affinity for EF-Tu. In particular, EF-Tu was shown to use the $tRNA^{Asp}$ base-pairs G51:C63, C50:**G64** and G49:**U65** for tightly binding it [24], where **G64** and **U65** are included in the conserved Asp codon GUC. For His, only C63 and U65 are completely conserved, which suggests, analogously to synthetase recognition scheme, that the present role in the recognition of tRNA by EF-Tu is likely to determine the degree of nucleotide conservation.

The C/ACs conserved in tRNAs of the non-robustly conserving amino acids Arg, Leu and Val, which are charged by class I synthetases, **seem** to lack a clear evolutionary role, as far as our results shows: The conserved C/ACs of Leu and Val are dispersed throughout the pre-3'end string (Fig. 1) rather than holding specific sites, as would be expected from functional C/ACs. 68% of the sequences of the pre-3'end from $tRNA^{Arg}$ contain a conserved codon at positions 66-68, but this site is unlikely to serve as an identity determinant, because the elevated presence of conserved C/ACs is the sum of occurrences of four different codons, and because its location partially overlaps the ranges of tRNA interaction with EF-Tu and with the synthetase. Correspondingly, the level of C/ACs conservation in the 3'end strings of these three amino acids is somewhat lower compared to the six robustly-conserving amino acids (Tab. 2), and the number of conserved C/ACs per amino acid is larger (four, five and eight for Leu, Arg and Val, respectively, compared to one-three for the robustly conserving amino acids). These statistical observations may point to a gradual process by which, in the absence of contemporary function, the conserved C/ACs were shifted throughout the acceptor-TΨC stem, while random mutations supplemented the list of conserved C/ACs and lowered the overall level of conservation.

Interestingly, eight out of the nine conserving amino acids, i.e. - all except His, are considered to be among the ten earliest amino acids to emerge in the prebiotic world, according to consensus chronology of amino acids built on 60 criteria [16] and to the Miller-Urey experiment [17]. In

contrast, amino acids whose pre-3'end string is devoid from the cognate triplets or exhibit statistically-expected occurrence, are considered to be rather late arrivals.

2. Implications of the common properties of Ala, His, Gly, Ser and Pro

Five robustly-conserving amino acids, i.e. Ala, Gly, His, Pro and Ser, possess common properties:

1. Their synthetases are classified as belonging to class IIa on the basis of their active site architecture [23].

2. Their synthetases were shown to be the most efficient in aminoacylating minihelices [25], except from Pro where relevant data is currently unavailable.

3. All these amino acids, except for His, are believed to be among the most ancient ones [16,17].

4. Their tRNAs exhibit C/AC conservation sites in the range 68-72, which forms a contact interface with the synthetase.

5. Nearly all the identity determinants for the synthetase recognition, present in the 3' side of minihelices derived from the acceptor stem of these five tRNAs, are contained in the conserved triplets (Tab. 4).

These apparently unrelated properties may all be correlated. Properties 1 and 2, i.e. having class IIa synthetases which perform non-AC aminoacylation efficiently, can be linked by suggesting that the efficiency stems from the distinct recognition mode carried out by bacterial class IIa synthetases, which interact with the tRNA acceptor stem in a base-specific fashion via their characteristic longer motif 2 loop [26]. Biochemical studies [27] revealed interactions of this loop with G72 of bacterial $tRNA^{Pro}$, while the X-ray structure of a complex of $tRNA^{Ser}$ with SerRS, showed the loop interactions with U68 and C69 [28], establishing the loop ability to scan the entire 68-72 region. Linking the non-AC aminoacylation efficiency with this distinctive recognition mode is consistent with the lack of C/AC conservation in the pre-3'end strings of the sixth amino acid charged by a class IIa synthetase, Thr (additional file 1 Tab. S1). In bacterial ThrRS, motif 2 loop contacts only the CCA end while the stem identity determinants are recognized by the N terminal domain of the enzyme [29], a functional exception that is likely to account for the lack of C/AC conservation. Interestingly, Thr is listed among the ten amino acids considered to be the most ancient ones [16,17].

Table 4. Synthetase recognition determinants vs. conserved C/ACs

Amino acid		**Ala**	**Gly**	**Ser**		**His**	**Pro**
C/AC conservation sites in the 68-72 range	Positions	68-70	68-70	68-70	70-72	70-72	70-72
	Conserved C/ACs*	GC**U**	CC**C**/UC**C**	**UCU/UCA**	**UCC**	**CA**C	**C**C**G***
Recognition determinants for minimal substrate aminoacylation*		G3**·U70** [5]	G3:**C70,** C2:G71† [30]	(C/**U)68**, (**C**/U)**69**, (**A/U**)**70** [9]	(A/**U)70 C71, C72** [9]	**C70, A71** [30]	No data

**In bold - nucleotides from the conserved coding-triplets serving as identity determinants.*

*** These identity determinants refer to the recognition of full tRNAs [27].*

† The identity determinant for aminoacylating minihelices of Gly, C2:G71, are not part of the conserved coding triplet.

The contemporary non-AC aminoacylation, which is efficiently carried out by class IIa synthetases, was suggested to be a relic from the primordial aminoacylation process [6,12,31] and consistently, Ala, Gly, Ser, and Pro are considered to be among the earliest amino acids [16,17]. This correlation may be explained by proposing, in accord with previous suggestions, that the earliest synthetase that catalyzed the primordial enzymatic charging was the ancestor of class II synthetases [10], or particularly of class IIa synthetases [32].

2.1 Possible implications on the origin of the genetic code

Generating a primordial synthetase, which could have utilized the identity determinants for aminoacylating tRNAs via the ancient non-AC aminoacylation process, would have required a primitive translation system relying on aminoacylated proto-tRNAs, thus presenting a chicken and egg paradox [18]. However, attributing the conserved coding-triplets to a non-enzymatic aminoacylation process that preceded the appearance of a synthetase seems to resolve the dilemma, as the non-enzymatically charged aa-tRNAs could have been engaged in the prebiotic synthesis of the proto-synthetase. A non-enzymatic charging process could have possibly relied on direct interaction between an amino acid and its cognate coding triplet, i.e. on the stereochemical interaction [33] which suggests specific affinity between amino acids and their

cognate coding triplets. Alternatively, the acceptor stems can perhaps acquire subtly-distinct conformations, which are particularly suited for accommodating the cognate amino acid [14,15,34]. Such a specific conformation was possibly induced by the presence of a coding triplet in the sequence. According to both mechanisms, an amino acid would reside on the acceptor stem and its proximity to the 3'end would facilitate the tRNA charging.

The existence of stereochemical affinity between amino acids and their coding triplets is considered non-conclusive at best [6,34], yet, the possible relevance of this idea seems to be supported by an experimental demonstration of non-enzymatic aminoacylation. The specific charging of C4N hairpins, which was catalyzed by di-peptides that were unlikely to provide the specificity, is attributed to the interaction of the amino acid with the cognate codon at the 5'end [13], pointing to the possible existence of sequence-specific affinity. The specificity reported in the self-aminoacylation of tRNAs, which occurred under high pressure in the absence of a synthetase or ATP, is proposed to be encoded in a nucleotide sequence, probably of the amino acid stem [14,15]. This process was demonstrated for $tRNA^{Phe}$ from *E. coli* and yeast. In accord with the two mechanisms proposed for the specific accommodation of amino acids on the acceptor stem, the two unique sequences of elongator-tRNAs and genomic tRNAs from yeast, found in the tRNAdb (additional file 2 DataSet S3), include the Phe codon UUC at positions 68-70. However, the only sequence of $tRNA^{Phe}$ from *E. coli* (additional file 2 DataSet S1), does not display a cognate C/AC of Phe in its pre-3'end, possibly pointing to an alternative self-aminoacylation mechanism. The misfit between the processes implied by the yeast and *E. coli* aminoacylation, calls for further experimental studies of this non-enzymatic mechanism.

The dependence of the non-enzymatic aminoacylation, as well as of the non-AC enzymatic aminoacylation, on a particular combination of nucleotides located in the acceptor-stem, offers two possible scenarios concerned with the origin of the genetic code. Triplets of the standard genetic code could have originated, at least in part, from the acceptor-stem determinants for enzymatic aminoacylation by a proto-synthetase [6,11]; Or alternatively, coding triplets suggested here to be engaged in an earlier process of non-enzymatic aminoacylation, formed the basis for the current genetic code. In the latter case, the ensuing appearance of enzymatic aminoacylation would have retained, as identity determinants, only those nucleotides required for the recognition by the proto-synthetase. The second scenario seems more reasonable

because, evolutionarily speaking, non-enzymatic processes are expected to precede the enzymatic ones. Moreover, the second scenario is supported by the partial conservation of the coding triplets observed in the acceptor stems of Ala and His, where two third of the sequences carry a full cognate codon, but one third conserves only two of the three coding letters. The possibility that random mutations could have supplemented the correct third letter and completed the codon triplet in the majority of Ala and His sequences appears less feasible than the alternative process, by which one letter from the initial triplet was abandoned due to the relaxation of functional requirements in the transition from non-enzymatic to enzymatic aminoacylation. A scenario, in which the identity determinants for non-AC aminoacylation were singled out from their former conserved C/ACs, links the conserved triplets and the identity determinants contained in them, i.e. properties 4 and 5, supporting the notion that the origin of the conserved coding triplets and of the non-AC determinants, is likely to be prebiotic [22]. However, the "operational code" residing in the tRNA acceptor stem, which was associated with a second code distinct from the nucleotide triplets of the standard genetic code [6,11,12], consists primarily of elements from the current coding-triplets set, suggesting the singularity of a subset of the standard genetic code in the evolution of life as we know it.

3. Discussing the results of the alternative method

The alternative method gave rather similar results when it comes to the list of conserving and robustly conserving aa's. On the other hand, the list of conserved and robustly conserved coding triplets is far less similar. We leave further research on the correct meaning of conservation of triplets for future research. The important point made by using two methods and not just one is that we succeeded in benchmarking the resulting lists (up to small changes, clarified once one allows some flexibility in the choice of threshold).

4. Proposed evolutionary path

Incorporating the insight gained from stem coding triplets into the vast experimental and theoretical data accumulated in the field of Origin of Life, may suggest a continuous evolutionary scenario: Long oligonucleotides, reaching over 100 mer, which were demonstrated, both experimentally and theoretically, to polymerize in solution under a thermal gradient [35], would have autonomously folded in the primordial environment into stable RNA molecules,

mostly into hairpins consisted of a single or two helices [36]. Two matching hairpins were proposed to combine to form an L-shaped proto-tRNA carrying a coding triplet both in its AC loop and in the acceptor stem [11,37,38]. Self-aminoacylation, occurring either under high-pressure or via another yet unknown process, is suggested to have allowed certain hairpins, as well as L-shaped RNAs to be autonomously charged with their cognate amino acid.

Self-aminoacylation occurring under high pressure is a tempting possibility, because it matches up with the widespread belief that life started at hydrothermal vents [39], where elevated pressure is prevalent and where liquid under thermal gradient, that can induce the formation of long RNA chains [35], is found. It follows that attempts to reproduce prebiotic processes with contemporary chemistry may benefit from working under high pressure. In particular, it may prove useful to conduct the experimental search for the stereo-chemical effect under high pressure conditions. RNA helices derived from the acceptor stems of Ala, Gly, His, Pro or Ser, which carry full codons and were suggested to accommodate amino acids prebiotically, may offer suitable binding sites for testing the specific affinity of amino acids for their coding triplets.

These self-assembled products of the prebiotic chemistry, i.e. charged hairpins and charged L-shaped RNAs could have been floating around in the prebiotic environment. Charged hairpins were assigned a possible role in prebiotic replication [40], while charged L-shaped RNAs could have started acting as proto-tRNAs with the appearance of a coding proto-ribosome. This ribozyme, that was capable of translating a code into a peptide, could have possibly evolved via the addition of an RNA element that accommodated the translated RNA, i.e. the initial version of the ribosomal small subunit, into an earlier proto-ribosome that catalyzed the association of random amino acids [41,42]. Translation of random RNA chains would produce random polypeptides, which could, in certain cases, possess weak catalytic capabilities, maybe yielding in such manner a simple proto-synthetase. When the earliest specific synthetase emerged, the conserved coding triplets at positions 68-72, which, until that stage are suggested to control the non-enzymatic aminoacylation, could have constituted the template for establishing the mutual recognition scheme between the proto-tRNAs and the primordial synthetases. The vestige of the primordial enzyme was linked with the catalytic core of the modern synthetase, while the AC binding domain would have ensued [6,10,31]. Charging of full tRNAs by this composite synthetase relies, in certain cases, on acceptor-stem determinants distinct from those of the

non-AC aminoacylation [10], expanding the conserved part of the acceptor stem, as is the case with the conserved base-pairs C2:G71, G1:C72 of $tRNA^{Gly}$ [43]. The interaction of tRNA with EF-Tu would join when protein factors were integrated into the function of an elaborated ribosome, giving rise to higher efficiency and accuracy. The three key components of the protein-synthesis machinery, i.e. the tRNA, synthetase and EF-Tu, along with the ribosome, would have coevolved into the intricate contemporary translation system, still keeping traces of their prebiotic roots in the form of the conserved coding-triplets. The evolutionary path proposed for the coding triplets conserved in the tRNA acceptor stem, commencing from specificity control in a prebiotic non-enzymatic aminoacylation to the involvement in the specificity of the contemporary synthetase, outlines a continuous route from the prebiotic era to modern biology.

Conclusions

Our results detect highly-conserved cognate coding triplets in the acceptor-TΨC stem of the bacterial tRNAs of six amino acids, mostly those considered to be the earliest. These triplets allow to suggest a temporally order for principal events in the prebiotic emergence of translation. They are suggested to have been involved in an initial non-enzymatic aminoacylation process, a process whose existence can resolve the chicken and egg paradox concerned with the appearance of translation. The triplets are proposed to be the precursors of modern recognition determinants located the acceptor-TΨC arm, and the fact that they constitute a subset of the contemporary codon set, suggests the high significance of (part of) the current set of codons in the emergence of life. The triplets also point at the ancestor of the class IIa synthetases as the earliest aminoacylating enzyme. Overall, the conserved coding triplets are suggestive of a coevolution scheme of the three major components of the protein-synthesis machinery, i.e. the tRNA, synthetase and EF-Tu, laying the ground for the emergence of coded translation.

Methods

Dataset construction:

Sequences of the 3'end side of the acceptor-TΨC stem from the bacterial tRNAs of each of the 20 canonical amino acids, were obtained from the tRNAdb database [20,21],

http://trna.bioinf.uni-leipzig.de. The tRNAdb database holds over 6200 bacterial tRNA genes, belonging to the 20 canonical amino acids. To avoid the inclusion of pseudo-tRNA genes, i.e. those not coding for elongator-tRNAs, only tRNA sequences having a full CCA 3'end and a consensus discriminator base (N73) [8] were incorporated into the statistics, with one exception; The data for Glu contained similar amount of sequences with the canonical G73 and with the non-canonical A73, both groups displaying the same pattern of C/AC occurrence. In order to obtain as much data as possible, tRNAs with non-canonical A73 were incorporated.

The observed occurrence of the cognate codon and anticodon triplets was determined in 10 mer strings composed of nucleotides 63-72 (termed pre-3'end), by scanning the sequences. Nucleotides 61 and 62 were not included in the study, being conserved and semi-conserved, respectively. The scanning indicated that one class of bacteria, i.e. species from six genera of the Mollicutes class, which constituted about 9% of the initial data set, i.e. 350 gene-sequences, does not share the statistical behavior of the other bacteria (additional file 1 Tab. S2) and it was excluded. No other species in the data contained enough entries to be reliably recognized as diverging from the conservation pattern. Under these restrictions, only 60% of the original sequences were retained, comprising an initial sample of 3324 tRNA gene-sequences from over 100 bacterial species.

Conserved sets assignment:

In our main "assignment method", the minimal set of cognate triplets accounting for 95% of the sequences carrying codons and/or anticodons at any position along the 10-mer pre-3'end string, was assigned as the conserved set for each of the conserving amino acids. The robustly conserved C/ACs are those located at specific sites in the pre-3'end string, occurring, alone or with a single additional and less frequent coding triplet (which will also be considered as robustly conserved), in more than 50% of the sequences. Robustly conserving amino acids (Pro, Asp, Ala, His, Glu, Ser) are those whose pre-3'end string contains at least one robustly conserved C/AC. The choice of the conserved sets for the non-robustly conserving amino acids (Arg, Leu and Val), is not unique.

Calculation of expected occurrences:

The expected occurrence of the coding triplets in the pre-3'end string was determined by analyzing their occurrence in simulated 10-mer randomized oligonucleotides. A pool of 10^5

randomized strings was generated using MATLAB and Statistics Toolbox (Release 2012b, The MathWorks, Inc., Natick, Massachusetts), by producing a string of 10^6 nucleotides with a predetermined nucleotide-distribution, randomizing it and cutting it into 10^5 10-mer oligonucleotides. For the conserving amino acids, the percentage of the 10-mer randomized strings carrying at least one cognate triplet from the conserved triplet-set is defined as the expected occurrence. For the remaining amino acids, the percentage of the 10-mer randomized strings carrying at least one cognate codon or anticodon triplet was set to be the expected occurrence. The error value for the expected percentage, ±0.8%, includes a statistical error of 0.6% and a rounding error of 0.5%. The probability of random occurrence, *P*, was calculated for each amino acid, from the results of a χ^2 test, using a statistical cut off of $P \leq 0.01$.

Alternative Conserved sets assignment:

The above assignment method is not the only possible method that makes sense. Therefore, it is important to check if a somewhat different method will lead to similar results: To benchmark the resulting list of robustly conserved aa's and their robustly conserved coding triplets, we add here an alternative assignment method.

We note that in our main method, the assignment is based on the removal of strings that are already counted due to the appearance of one highly appearing C/AC, and on completing the desired goal of having C/AC in 95% by counting other coding triplets, but only in the remaining strings; such a method can miss some relatively highly occurring coding triplets, while adding to the list of conserved C/AC, far less occurring coding triplets. This method also removes symmetry between coding and non-coding triplets, as only the most appearing coding triplet (and not the most occurring non-coding triplets) results in a removal of strings.

We therefore choose an alternative method that is complementary to the main one in those respects: It counts and takes into account all highly appearing coding triplets, without a removal of strings, and furthermore, it considers only the frequency of appearance per site and not per string. Any coding triplet appearing in a specific location in 50% or more of the strings is considered robustly conserved. Accordingly, the aa's having a robustly conserved C/AC are considered to form the set of robustly conserving aa's (as well as being part of the conserving aa's set). In addition, any C/AC appearing in 20-49% of the strings in a specific location are considered to be conserved (and not robustly conserved); the resulting aa's having no robustly conserved coding triplet yet having conserved coding triplet(s) are considered to complete the

set of conserving aa's. We further analyze in the Results Section slight variations of the conclusions about conservation and robust conservation, by checking what kind of slight modification of these two thresholds - from 50% and from 20% to slightly lower thresholds – will change the list of robustly conserving and conserving aa's.

Finally, we apply a simple statistical test to find a relevant P-value for checking how significant is the correlation between robustly conserving aa's versus aa's having a Class IIa synthetase: Given that there are exactly 6 out of the 20 aa's that are charged by Class IIa synthetase, we compare our result (of k robustly conserved aa's) being correlated with the above 6, if the null hypothesis is that there is no correlation, namely if the choice of our list (given the other) had been taken from a uniform distribution of k out of 20 aa's. The statistical test is counting the total number of aa's appearing in both lists.

Abbreviations:

aa- amino acid

aaRS - aminoacyl-tRNA synthetase

AC- anticodon

C/AC – codon and/or anticodon triplet

Pre-3'end – 10 mer string preceding the tRNA 3'end

Acknowledgements

We thank D. Agmon for generating the Matlab program, A. Cohen for the statistical consultancy and S. Agmon for his assistance. Support was provided by the MOD grant.

References

1. Rudinger J, Blechschmidt B, Ribeiro S and Sprinzl M. Minimalist aminoacylated RNAs as efficient substrates for elongation factor Tu. Biochemistry. 1994;33:5682-8.
2. Nissen P, Thirup S, Kjeldgaard M, Nyborg J. The crystal structure of Cys-tRNACys-EF-Tu-GDPNP reveals general and specific features in the ternary complex and in tRNA. Structure. 1999;7:143-56.
3. Schulman LH, Pelka H. Anticodon switching changes the identity of methionine and valine transfer RNAs. Science. 1988;242:765-768.

4. Sampson JR, DiRenzo A, Behlen L, Uhlenbeck OC. Nucleotides in yeast tRNAPhe required for the specific recognition by its cognate synthetase. Science. 1989;243:1363-1366.
5. Normanly J, Abelson J. tRNA identity. Annu Rev Biochem. 1989; 58: 1029-49.
6. Schimmel P, Giegé R, Moras D, Yokoyama S. An operational RNA code for amino acids and possible relationship to genetic code. Proc. Natl. Acad. Sci. USA. 1993;90:8763–8768.
7. Sampson R, Saks ME. Contributions of discrete tRNA(Ser) domains to aminoacylation by E.coli seryl-tRNA synthetase: a kinetic analysis using model RNA substrates. Nucleic Acids Res. 1993;21:4467-4475.
8. Hou YM. Discriminating among the discriminator bases of tRNAs. Chem. Biol. 1997;4:93-96.
9. Saks ME, Sampson JR. Variant minihelix RNAs reveal sequence-specific recognition of the helical tRNA(Ser) acceptor stem by E.coli seryl-tRNA synthetase. EMBO J. 1996;15:2843-9.
10. Giegé R, Sissler M, Florentz C. Universal rules and idiosyncratic features in tRNA identity. Nucleic Acids Res. 1998;26:5017–5035.
11. Musier-Forsyth K, Schimmel P. Atomic determinants for aminoacylation of RNA minihelices and relationship to genetic code. Acc. Chem. Res. 1999;32:368-375.
12. de Duve C. Transfer RNAs: the second genetic code. Nature. 1988;333:117-118.
13. Shimizu M. Specific aminoacylation of C4N hairpin RNAs with the cognate aminoacyl-adenylates in the presence of a dipeptide: origin of the genetic code. J Biochem. 1995;117:23-6.
14. Krzyzaniak A, Barciszewski J, Sałański P, Jurczak J. The non-enzymatic specific amino-acylation of transfer RNA at high pressure. Int J Biol Macromol. 1994;16:153-8.
15. Krzyzaniak A, Sałański P, Twardowski T, Jurczak J, Barciszewski J. tRNA aminoacylated at high pressure is a correct substrate for protein biosynthesis. Biochem Mol Biol Int. 1998;45:489-500.
16. Trifonov EN. The triplet code from first principles. J. Biomol. Struct. Dyn. 2004;22:1–11.
17. Miller S L, Urey H C, Oró J. Origin of organic compounds on the primitive earth and in meteorites. J. Mol. Evol. 1976;9:59–72.
18. Hopfield JJ. Origin of the genetic code: a testable hypothesis based on tRNA structure, sequence and kinetic proofreading. Proc. Natl. Acad. Sci. USA. 1978;75,4334–4338.

19. Eigen M, Lindemann BF, Tietze M, Winkler-Oswatitsch R, Dress A, von Haeseler A. How old is the genetic code? Statistical geometry of tRNA provides an answer. Science. 1989;244:673-9.
20. Sprinzl M and Vassilenko KS. Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res. 2005;33(database issue):D139-140.
21. Jühling F, Mörl M, Hartmann RK, Sprinzl M, Stadler PF, Pütz J. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. Nucleic Acids Res. 2009;37(database issue):D159-162.
22. Möller W, Janssen GM. Statistical evidence for remnants of the primordial code in the acceptor stem of prokaryotic transfer RNA. J. Mol. Evol. 1992;34:471-7.
23. Ribas de Pouplana L, Schimmel P. Two Classes of tRNA Synthetases Suggested by Sterically Compatible Dockings on tRNA Acceptor Stem. Cell. 2001;104:191–193.
24. Schrader JM, Chapman SJ, Uhlenbeck OC. Understanding the sequence specificity of tRNA binding to elongation factor Tu using tRNA mutagenesis. J. Mol. Biol. 2009;386:1255–1264.
25. Wright DJ, Martinis SA, Jahn M, Söll D, Schimmel P. Acceptor stem and anticodon RNA hairpin helix interactions with glutamine tRNA synthetase. Biochimie.1993;75:1041-9.
26. Cusack S, Härtlein M, Leberman R. Sequence, structural and evolutionary relationships between class 2 aminoacyl-tRNA synthetases. Nucleic Acids Res. 1991;19:3489-98.
27. Liu H, Peterson R, Kessler J, Musier-Forsyth K. Molecular recognition of tRNA(Pro) by Escherichia coli proline tRNA synthetase in vitro. Nucleic Acids Res. 1995;23:165–169.
28. Cusack S, Yaremchuk A, Tukalo M. The crystal structure of the ternary complex of T.thermophilus seryl-tRNA synthetase with tRNA(Ser) and a seryl-adenylate analogue reveals a conformational switch in the active site. EMBO J. 1996;15:2834-42.
29. Sankaranarayanan R, Dock-Bregeon AC, Romby P, Caillet J, Springer M, Rees B et al. The Structure of Threonyl-tRNA Synthetase-tRNA^{Thr} Complex Enlightens Its Repressor Activity and Reveals an Essential Zinc Ion in the Active Site. Cell. 1999;97:371–381.
30. Francklyn C, Shi JP, Schimmel P. Overlapping nucleotide determinants for specific aminoacylation of RNA microhelices. Science. 1992;255:1121-5.
31. Schimmel P, Ribas de Pouplana L. Transfer RNA: from minihelix to genetic code. Cell. 1995;81:983-986.
32. Hartman H. Speculations on the evolution of the genetic code IV. The evolution of the aminoacyl-tRNA synthetases. Orig Life Evol Biosph. 1995;25(1-3):265-9.

33. Woese CR. On the evolution of the genetic code. Proc. Natl. Acad. Sci. USA. 1965;54:1546–1552.

34. Koonin EV. Frozen accident pushing 50: Stereochemistry, expansion, and chance in the evolution of the genetic code. Life (basel). 2017;7,E22.

35. Mast CB, Schink S, Gerland U, Braun D. Escalation of polymerization in a thermal gradient. Proc. Natl. Acad. Sci. U S A. 2013;110:8030-5.

36. Briones C, Stich M, Manrubia SC. The dawn of the RNA World: toward functional complexity through ligation of random RNA oligomers. RNA. 2009;15:743-9.

37. Dick TP, Schamel WA. Molecular evolution of transfer RNA from two precursor hairpins: implications for the origin of protein synthesis. J Mol Evol. 1995;1:1-9.

38. Nagaswamy U, Fox GE. RNA ligation and the origin of tRNA. Orig Life Evol Biosph. 2003;33:199-209.

39. Holm NG, Andersson EM. Abiotic synthesis of organic compounds under the conditions of submarine hydrothermal systems: a perspective. A Planet Space Sci. 1995;43:153-9.

40. Maizels N, Weiner AM. Phylogeny from function: Evidence from the molecular fossil record that tRNA originated in replication, not translation. Proc. Natl. Acad. Sci. USA. 1994;91:6729-6734.

41. Agmon I. The dimeric proto-ribosome: structural details and implications on the origin of life. Int. J. Mol. Sci. 2009;10:2921–2934.

42. Agmon IC. Could a proto-ribosome emerge spontaneously in the prebiotic world? Molecules. 2016;21:E1701.

43. Nameki N, Tamura K, Asahara H, Hasegawa T. Recognition of tRNA(Gly) by three widely diverged glycyl-tRNA synthetases. J Mol Biol. 1997;268: 640-7.