ARTICLE

“Protoribosome” as new game of life

Jacques Demongeot[1]

We show the existence of an RNA sequence which may have played a role in the origin of life to promote the first peptide assemblages such as a “protoribosome”. It is constructed using sequences from tRNA and rRNA (5S, 16S, 18S, 28S and 50S) of species coming from Archae, Bacteria and Eukaryota. Subsequences of the selected RNA sequence are present as frequent and stable repeat motifs in current genomes, and it possesses optimal biochemical as well as combinatorial properties.

[1]University of Grenoble Alpes, Faculty of Medicine, Laboratory AGEIS, Team AGIM (Autonomy, e-Gnosis Medical, Imaging & Modelling) 38700 La Tronche, France

email: Jacques.Demongeot@yahoo.fr

Fifty years ago, in 1967, S. Ulam simulated large automata networks and remarked that with simple growth rules, he obtained complicated patterns similar to those observed in biology[1]. Inspired by these results, J. Conway started in 1968 to simulate a new cellular automaton called “Game of Life” by M. Gardner in 1970, because it allowed to see discrete numerical structures which moved on the plane and were duplicated[2]. But the Conway’s algorithm did not incorporate realistic genetic considerations in the game, making it possible to test, for example, the plausibility of molecular events that may have led to the appearance of life. We seek in this article to go in this direction, proposing a simple RNA structure that could have served as a matrix for building the first peptides (a "protoribosome").

We will start the game by searching for species in which two types of RNA needed in the ribosomal protein building, *i.e.,* ribosomal RNAs and loops of transfer RNA, share at least heptameric sequences as witnesses of their same co-evolution. We found species satisfying this constraint, *e.g.*, a proteobacterium, *Rhodobacter sphaeroides* (Figs. 1A and 1B). By combining these sequences, we construct a circular RNA with 22 nucleotides, called AL (for Archetypal Loop) and we search for the most stable hairpin having the same sequence as AL. AL is partly observed in large genetic data bases with repeated motifs (tetramers, pentamers and hexamers) in non-coding DNA of numerous species and specific codons contributing to their stability.

Then, optimal combinatorial and biochemical properties of AL will be listed, especially by looking for AL relics in current RNAs (namely miRs, rRNAs, tRNAs and circRNAs). Eventually, AL is proposed as candidate for an early “protoribosome” facilitating during the evolution the construction of the first nucleopeptidic conjugates, a way to explain that molecules like rRNAs and tRNA involved in the protein translation share similar heptameric sequences.

Results

Construction of a circular RNA, called AL (for Archetypal Loop)

Comparing sequences C**ATGAATGGTACT**T**CCATTC**G**A** of Gly-tRNA^{TCC} loops[3],
AATGGTACTGCGTC**TCAAGA**C**G** from 5s rRNA[4],
GCCG**T**AA**A**C**GATGAATG**CC**A**G**T** from 16s rRNA[5],
AATT**G**A**ACTG**TGTG**TCAAGATG** from 23s rRNA[5],
ACG**G**CC**AC**CTT**C**G**TCCAAGATG** from 30s RNA[5],
TC**G**ACCGC**GCCATTC**TC**GAT**C**A** from 50sRNA[5],
GGTC**C**A**G**GTCA**TCAAGAT**C**AA** from rpsA1[5] and
G**ATGGT**G**C**CCA**C**G**TT**GGC**GATG** from CRISPR CAS1[5]
from the genome of *Rhodobacter sphaeroides*, we get a cyclic RNA called AL (Archetypal Loop): **GATGAATGGTACTGCCATTCAA**.

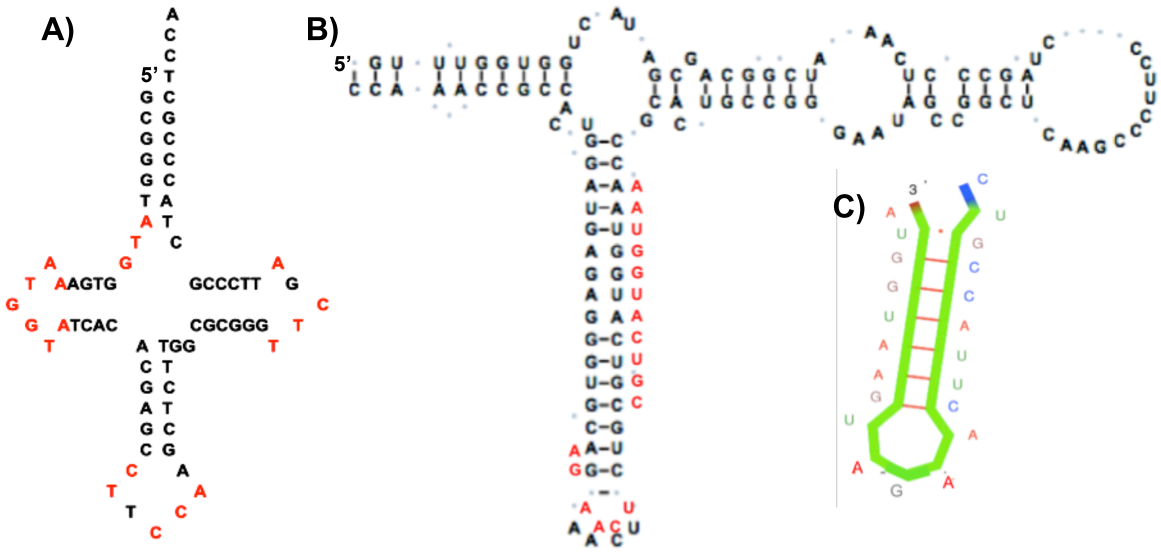


Fig. 1. Construction of the Archetypal Loop AL. (**A**) AL hemi-sequence ATGAATGGTACT and hexamer CCATTC from the loops of the Gly-tRNA^{TCC} of *Rhodobacter sphaeroides*[3]. (**B**) AL hemi-sequence AAUGGUACUGC and hexamer UCAAGA from the hairpin of the 5s rRNA[4] of *Rhodobacter sphaeroides*. (**C**) Optimal hairpin form for AL[6].

It is possible to build by using the Kinefold® algorithm[6], the most thermodynamically stable hairpin (-9.5 kcal/mol) among the 22 RNA chains obtained from the circular permutations of AL (Fig. 1C). This structure could explain, when the conditions of denaturation exist, a first loss of the hexamer CUGCCA (the anticodon loop of many current tRNA^{GLY}s), followed by a break between heptamers UUCAAGA (the Tψ-loop of many current tRNAs) and AAUGGUA (the D-loop of many current tRNAs). An argument in favor of this scenario is the distribution of the pentamers frequencies inside the current genome (from Rfam database[7]), which shows the two best survival probabilities for the pentamers coming from the most stable part of AL, also parts of the D-loop and Tψ-loop of many present tRNAS, *i.e.*, AAUGG, AUGGU, UGGUA, GGUAC, TTCAA, TCAAG and CAAGA as shown on Fig. 2. If we consider other subsequences of AL, we find many repeated motifs, such as AATGG[8] and GATG[9] existing in human microsatellites, AGAT in vertebrate repeated UTR motifs[10] and CCATTCA in Alpha Satellite of the Human Chromosome 17[11] and from the box HMG (High Mobility Group Box, a protein domain involved in DNA binding[12]), as well as optimal codons, *e.g.*, determinant for mRNA stability in yeast genome[13]. We can generalize the result obtained in Rhodobacter sphaeroides to other species in worlds like Archae, Bacteriae and Eukaryotae, by searching similar sequences:

Rhodobacter sphaeroides
AATGGTATTCCCATTCGATTTG tRNA-Gly
GATGGTGCCCACGTTGGCGATG CRISPR CAS1
AATGGTACTGCGTCTCAAGACG 5S rRNA
CCTGGAACTGCCATTGAAACTC 16S rRNA
AATTGAACTGTGTGTCAAGATG 23S rRNA
ACGGCCACCTTCGTCCAAGATG 30S rRNA
TCGACCGCGCCATTCTCGATCA 50S rRNA
AGGTCCAGGTCATCAAGATCAA Ribosomal protein S1 (rpsA1)
AATGGTACTGCCATTCAAGATG Consensus
Rhodobacter vinaykumarii
AATGGTACTCCCATTCGATCTG tRNA-Gly
CCTGCCACAGGCGTTCAAGATG CRISPR CAS4
AATGGTACTGCCATTCAAGATG Consensus
Rhodospirillum rubrum
TGAATGGTACTTCCAATTCGAA tRNA-Gly
CCAATGGTACTGCGTCTTAAGG 5S rRNA
CTCCAGGTACTGCCCTTGATAC 16S rRNA (https://www.arb-silva.de/browser/)
CGAATGGTACTGCCATTTAAAA Consensus
Rubellimicrobium thermophilum
AGTGGTACTTCCATTCGACATG tRNA-Gly
AATGGTACTGCGCCTCAAGACG 5S rRNA (http://combio.pl/rrna/)
GATGGTCCAGGCGCTGCCGCTC 16S rRNA (https://www.arb-silva.de/browser/)
AATGGTACTGCCACTCAAGATG Consensus
Haematobacter missouriensis
AGGGGTATTGCCATTCGAATTA tRNA-Gly
(http://trna.ie.niigata-u.ac.jp/cgi-bin/trnadb/whole_detail.cgi?SID=2138813)
TATGGTGCTTCCATTCCCGCTA tRNA-Gly (https://www.ncbi.nlm.nih.gov/nuccore/672903602?report=genbank)
AATGGTACTGCGTCTCAAGACG 5S rRNA
AATGGTAGTGACAATGGGTTAA 16S rRNA
AATGGTACTGCCATTCAAGATG Consensus
Paracoccus sp. S4493
AATGGTACTTCCCTTCGATTTA tRNA-Gly
(https://www.ncbi.nlm.nih.gov/nuccore/NZ_JXYF01000001.1?from=63131&to=63204&sat=19&sat_key=63645080&report=fasta&strand=2)
GATGGTACTGCGTCTTAAGACG 5S rRNA
AATGGTGGTGACAGTGGGTTAA 16S rRNA (http://www.ebi.ac.uk/ena/data/view/FJ457300&display=fasta)
AATGGTACTGCCATTCAATTTA Consensus
Flavobacteria bacterium MS024-2A
CTGGTATTGCCATTCGAATCGC tRNA-Gly
(http://gtrnadb.ucsc.edu/genomes/bacteria/Flav_bact_3519_10/flavBact_3519_10-tRNAs.fa)
ATGGTACTGCCATCCGGTGGGA 5S rRNA
ATGGTAACGGCATACCAAGGCA 16S RRNA (http://www.ebi.ac.uk/ena/data/view/AM931128&display=fasta)
ATGGTACTGCCATTCGAAGGGA Consensus
Methanococcus maripaludis
CTGGTACTTCCATTCAAATCGT tRNA-Gly (http://gtrnadb.ucsc.edu/genomes/archaea/Meth_mari_C5/methMari_C5_1-tRNAs.fa)
TAAGTACTGCCATCUGGUGGGA 5S rRNA (https://www.ncbi.nlm.nih.gov/nuccore)
(http://biobases.ibch.poznan.pl/htbins/getseq.cgi?name=Methanococcus%20maripaludis)
TCGGTACGGGCCTTGAGAGAGG 16S rRNA (http://www.ebi.ac.uk/ena/data/view/AB546258&display=fasta)
TTGGTACTGCCATTCAGAGAGA Consensus
Tremella mesenterica
GATCTGCGAAGTCAAGATGAAT 5S rRNA
GGTAATTCTAGAGCTAATACAT 18S rRNA (https://www.ncbi.nlm.nih.gov/nuccore/256600119?report=fasta)
GTACCGTGAGGGAAAGATGAAA 28S rRNA (https://www.ncbi.nlm.nih.gov/nuccore/46402656?report=fasta)
GGTCCGTGAAGTCAAGATGAAT Consensus
Drosophila melanogaster
TGGATGTGTGATTGTACTGCCA Toll (https://www.ncbi.nlm.nih.gov/nuccore)
AGGAACTGCTCGTGAAGCTGAA CRISPR CAS9 (https://www.ncbi.nlm.nih.gov/nuccore)
Homo sapiens
GTGGTACTCCCATTCAATTTGG tRNA (http://trna.bioinf.uni-leipzig.de/DataOutput/Result)
ATGGTAGTCGCCGTGCCTACCA 18S rRNA (https://www.ncbi.nlm.nih.gov/nuccore/225637497?report=fasta)
ATGGTAATCCTGCTCAGTACGA 28S rRNA (https://www.ncbi.nlm.nih.gov/nuccore/1154886866?report=fasta)
TCAAGCTGAATAATACAACTGT RagB mRNA (https://www.ncbi.nlm.nih.gov/nuccore)
TATCTATGACCATTCAATATTG RagC mRNA (https://www.ncbi.nlm.nih.gov/nuccore)
ATGGTACTCCCATTCAATACGA Consensus

From the above sequences, we can obtain the following consensus sequences:

AATGGTACTGCCATTCAATTTA Consensus Bacteria
ATTGGTACTGCCATTCAGAGAG Consensus Archae
AATGGTACTCCCATTCAATACG Consensus Eukaryote
AATGGTACTGCCATTCAAGATG Consensus consensorum = AL

Another way to obtain AL is by remarking that the most frequent sequence shared by all above sequences is GAAUGGU. By starting from it under the constraint to have as result a hairpin, we have to deal also with its symmetrized sequence: CUUACCG. Then, by adding the end codon, we get UGAAUGGU and the symmetrized ACUUACCG, and choosing among the most frequent pentamers in Rfam (Fig. 2), the sequences UGGUA and UUCAA, we obtain the double sequence: AUGAAUGGUA
AACUUACCGU.

To end, by adding G and C respectively on left and right side for obtaining the missing codons AGA and ACU (the most frequent in their codon class in many species like *E. coli*, *S. cerevisiae* and *C. elegans*[14]), and then, we get the AL hairpin.

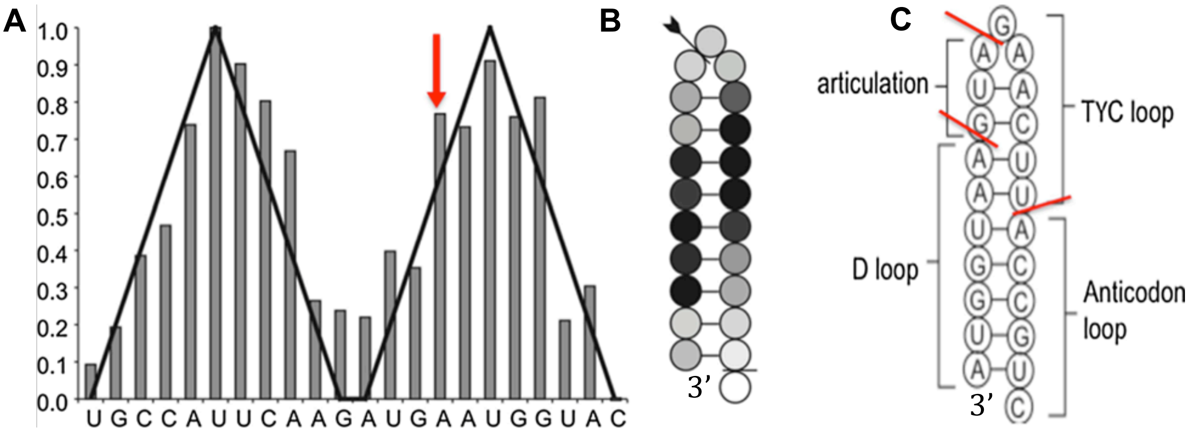


Fig. 2 Frequencies of the pentamers coming from AL and matching Rfam subsequences of length 5. (**A**) The bar indicating the frequency of a pentamer in Rfam[7] is located at the level of its first base, *e. g.*, the red arrow indicates the frequency of AAUGG[15]. The values of pentamer frequencies have been normalized to the value 1. (**B**) The positions marked with an arrow in the AL hairpin allow calculating the distances to each base, represented by the triangle lines in (A) and by values graphed in a gray scale inside circles in (B) with white and black representing respectively their minimum and maximum. (**C**) AL Hairpin is separated by red bars into segments aligned with tRNA conserved domains.

Discussion

Optimal combinatorial and biochemical properties of AL

Codons of the 20 amino-acids are present one and only one time in AL circular form. Anticodons of 16 amino-acids are present in AL: GAA, AAG, AAT, CAT, TAC, ACT, TGG, GGT, TGC, GTA, ATG, CTG, ATT, TTC, GCC, CCA. The 4 anticodons absent are those of lysine (affine to its codons in Table 1B), aspartic acid (quasi codon affine in Table 1B), cysteine (there is no result about its codon or anticodon affinity[16] in Table 1B and 1C, but it is the least frequent[17] in Table 1A and absent in the Miller experiment[18,19]), and arginine (quasi codon AGA affine in Table 1C). Then, we may consider AL as a “matrimonial agency” of shortest length favoring first local weak affinities of amino-acids to their codons and anticodons, second the constitution of strong peptide bonds between these amino-acids due to their vicinity to AL, and third, consequently, the construction of short peptides.

Amino-acid	**A** Mean frequency in 9 species[12]	**B** Relative codon enrichment	**C** Relative anticodon enrichment
Ala	0.0777	0.912	0.951
arg (AGR)	0.0627	0.929	0.998
Asn	0.0336	0.796	0.987
asp	0.0542	0.989	1.320
cys	0.0078	-	-
gln	0.0315	0.946	1.234
glu	0.0859	1.149	0.596
gly	0.0730	1.058	1.006
his	0.0192	0.904	1.078
ile	0.0666	1.437	1.283
leu (CUN)	0.0891	1.059	1.050
lys	0.0776	1.033	0.955
met	0.0241	0.731	1.152
phe	0.0361	0.862	1.384
pro	0.0435	0.980	0.993
ser (UCA)	0.0466	0.954	0.988
thr	0.0487	1.156	0.923
trp	0.0102	0.667	1.640
tyr	0.0300	1.130	1.274
val	0.0817	0.916	0.922

Table 1. Frequency and affinities of amino-acids. (**A**) Amino-acid frequencies in 9 species[17]. (**B**) Inside the ribosome of 4 different species and in presence of amino-acids, relative enrichment of their codons and (**C**) anticodons[16]. Red (blue) values correspond to (quasi-) significant enrichment ($p = 0.05$).

More, we have selected many other optimal combinatorial and biochemical properties of AL as summarized in Tables 2 and 3.

Optimal combinatorial properties[15]

1 All dinucleotides appear in AL except CG, which is very uncommon in most portions of vertebrate genomes

2 Among the rings satisfying the principle “to be as short as possible and contain at least one codon of each amino-acid”, there is no solution for a length < 22. For the length 22, all the 29520 solutions contain only one repeated codon AUN, N being G for 52% of the solutions

3 The codons present in AL are stable and its tetramers, pentamers and hexamers belong to the most frequent ones in many genomes[7-12]

4 From the 29520 solutions, there are only 24 rings forming a hairpin of maximal length 9. From them, 19 rings repeat AUG and only one, AL, reaches the thermodynamic optimum (Fig. 1C)

5 9 from the 19 rings have a succession of 22 overlapping codons from start (AUG) to end (UGA) codons

6 Ranking the average distances (circular Hamming, permutation and edit ones) from a ring to all the others, AL wins as having the minimum

7 The 4 conserved domains of the tRNAs (3 loops and one articulation pivot) are nicely distributed in the hairpin form of AL

8 The closest ring in mean edit distance to all tRNA loops from GtRNAdb is AL

9 The mean edit distances of AL to the real miRs from miRBase[20] and to the presently known microRNAs repeated in at least two different species (repmiRs) are significantly below the mean edit distance of AL to 20,000 random RNAs of length 22 having the same base frequencies than AL[15]

10 AL has 15 common bases with the barycenter of the repmiRs

11 There are more than 50% of the tRNAs of GtRNAdb[21] with an edit distance of their loops to AL ≤ 4 (more than 18 matches)

Table 2. Optimal combinatorial properties of the ring AL.

	Optimal biochemical properties[15]
12	AL fits well the loops of GlytRNA^{GCC} of *Oenothera coquimbensis*[22] and *Arabidopsis thaliana*[23] (Fig. 3). More generally, there are 18,196 occurrences of TGGTA and TTCAA or TTCGA respectively in D- and Ty-loops of the 111,385 tRNAs in GtRNAdb (supplementary material 1).
13	AL matches well with exon/intron boundaries[24]
14	One of the 19 rings of property 4, AB[25-28], fits well the circular AL (Fig. 3F)
15	AL matches with Hamming and edit distances ≤ 2 with at least 30 tRNA-Gly from GtRNAdb (supplementary material 2)
16	AL has GCC in anticodon position: GCC has already been suggested as the first anticodon, because it “anticodes” for the simplest amino-acid, the glycine
17	AL aligns with the articulation pivot of many tRNAs, AUG, and allows the pairing TGG-ΨCA needed between the D- and Ty-loops for the tRNA folding
18	AL matches well with many viral not coding genomes[29]
19	AL matches well with many microRNAs[30]
20	AL matches well with IRE and YUNR loops[31]
21	AL matches well with many circular RNAs[31]
22	AL contains twice all most unexpected dimers as defined by P. Slonimski[32]

Table 3. Optimal biochemical properties of the ring AL.

AL Relics in current genomes

Searching for AL relics consists in studying the survival of some AL sub-sequences in present genomes (Table 4) from different databases:

1) NCBI Sequences[33]

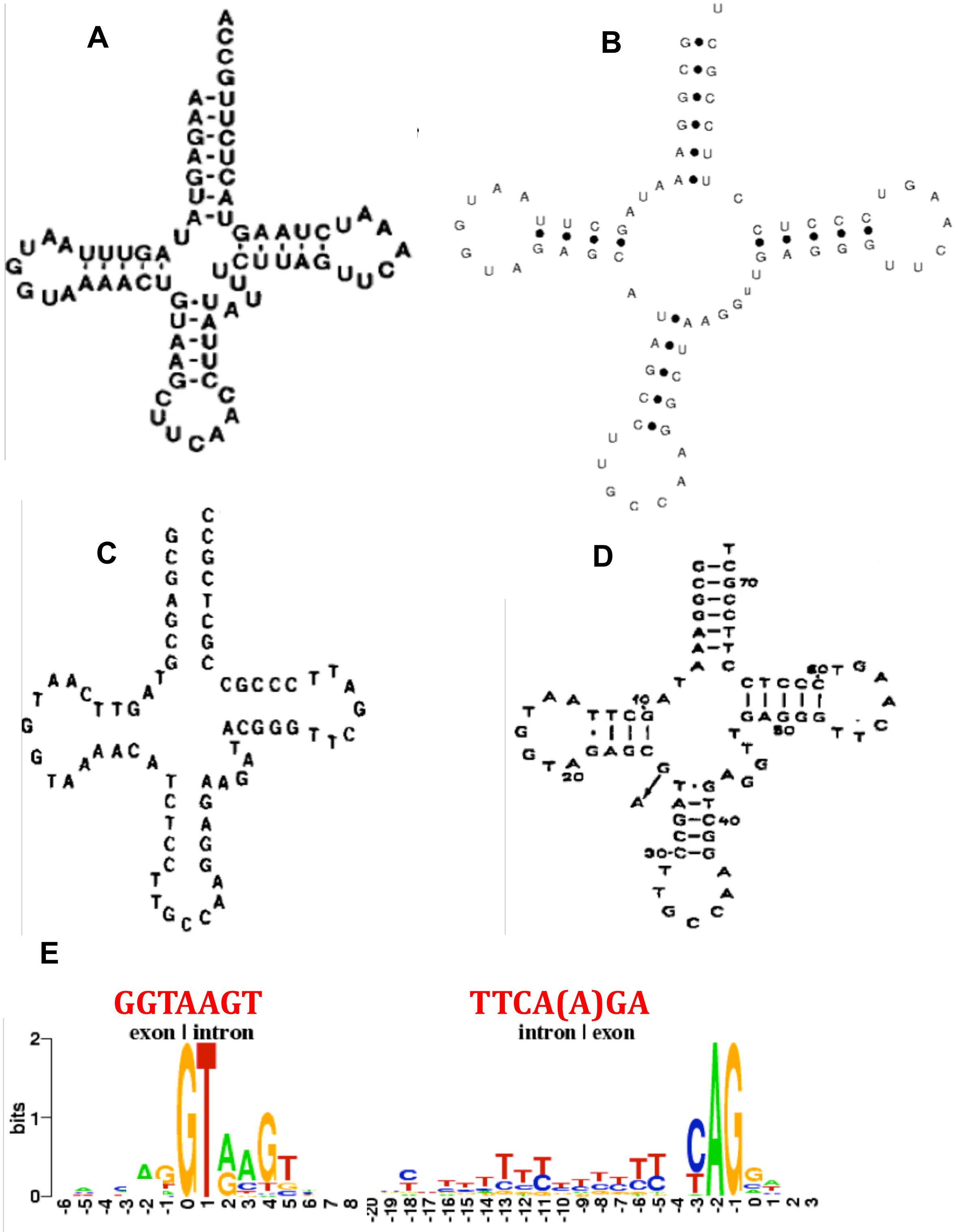
In Table 4, we have reported 140 quasi-perfect matches of length more than 18 for the sequence CTGCCATTCAAGATGATA in Reference Sequence Genomic Database from NCBI source[33]. The probability of occurrence of a sequence of length 18 equals $(1/4)^{18} \approx 1.5\ 10^{-11}$, with an expected number of 7.62 ± 4.56* (the symbol* corresponding to the 95%-confidence interval) among 13,777,995 sequences with a total number of 533,178,330,807 bases. We observed also such unexpected matches in other databases of NCBI source (Table 4).

2) Exon/intron boundaries sequences[24]

Consensus sequences at exon/intron boundaries in complex eukaryotes show a high similarity with AL subsequences GGTAAGT and TTCAAG (Fig. 3E).

3) Frequent RNA or DNA sequences[34-39]

From tRNAs, rRNAs, miRNAs and circRNAs in genetic data bases (GtRNAdb, NCBI sources, miRBase and circBase), many sequences fit well AL (Fig. 3 and Table 4). AL-pentamers ATGGT and TTCAA (parts of the D-loop of many tRNAs) belongs to the most frequent tetradecamer sequences in the genome of respectively *Caenorhabditis elegans* and *Arabidopsis thaliana*[39].



hsa_circ_0015360 (reverse strand): GGCATTCAAGACTATGAATGTT
AB: GCCATTCAAGACTATGAATGGT

F Colius striatus isolate BGI_N325 21746-21766: AATGGTACTGCCATTCAAGATT
AL: AATGGTACTGCCATTCAAGATG

Fig. 3. AL Relics in current genomes. (**A**) TrptRNA^{TGA} of *Aspergillus nidulans*[38], (**B**) GlytRNA^{GCC} of *Arabidopsis thaliana*[23], (**C**) GlytRNA^{GCC} of *Hordeum vulgare*[40] and (**D**) GlytRNA^{GCC} of *Œnothera coquimbensis*[22], whose loops (D-, anti-codon and Tψ-loops) fit quasi-perfectly AL (**E**) Consensus sequences at exon/intron boundaries in eukaryotes[24]. (**F**) Quasi-perfect matches between AB and a sequence of a circular RNA[36], and between AL and a sequence of the genome of *Collius striatus*[33].

NCBI sources Expected nb of sequences $\mu \pm 2\sigma$	AL circular permutation TTCAAGATGAATGGTACTGCCA	AL circular permutation CTGCCATTCAAGATGAATGGTA	AL circular permutation AATGGTACTGCCATTCAAGATG
	Nb sequences observed of length ≥18	**Nb sequences observed of length ≥18**	**Nb sequences observed of length ≥18**
refseq_genomic 7.62±4.56*	70 ($p=10^{-112}$)	140 ($p=10^{-498}$)	98 ($p=10^{-234}$)
Human ALU repeat 0.15±0.64*	5 ($p=10^{-35}$)	5 ($p=10^{-35}$)	7 ($p=10^{-86}$)
Protein Data Bank (PDB) 2.82±2.77*	49 ($p=10^{-166}$)	50 ($p=10^{-173}$)	38 ($p=10^{-97}$)
	Nb sequences observed of length ≥10	**Nb sequences observed of length ≥10**	**Nb sequences observed of length ≥10**
16s ribosomal 25.5±8.3*	41 ($p=10^{-3}$)	112 ($p=10^{-66}$)	138 ($p=10^{-110}$)
7s mtDB 10^{-3}±0.05*		**Nb sequence TACTGCCAGCCACCATGAAT observed** 200 ($p=10^{-9450000}$)	

Table 4. Highly unlikely matches between sub-sequences from three circular permutations of AL and from different nucleic databases of NBCI sources[33]**.** These sources contain numerous reference sequences (*e.g.*, 13,787,534 for refseq_genomic) from multiple genomes (23,072 species for refseq_genomic) with a great number of bases (a total number of 533,178,330,807 bases for refseq_genomic) or a great number of individuals (*e.g.*, 164 countries or ethnies for 7s mtDB[41] and supplementary material 3). * indicates the 95% confidence interval and p the probability to expect the observed sequence, calculated from the normal approximation $N(\mu,\sigma)$ of the binomial variable X: $p=P(|X-\mu| \geq k\sigma) \leq \exp(-k^2/2)/k\sqrt{2\pi}=10^{-y}$, where $y \approx 0.4+\mathrm{Log}_{10}k+0.217k^2$.

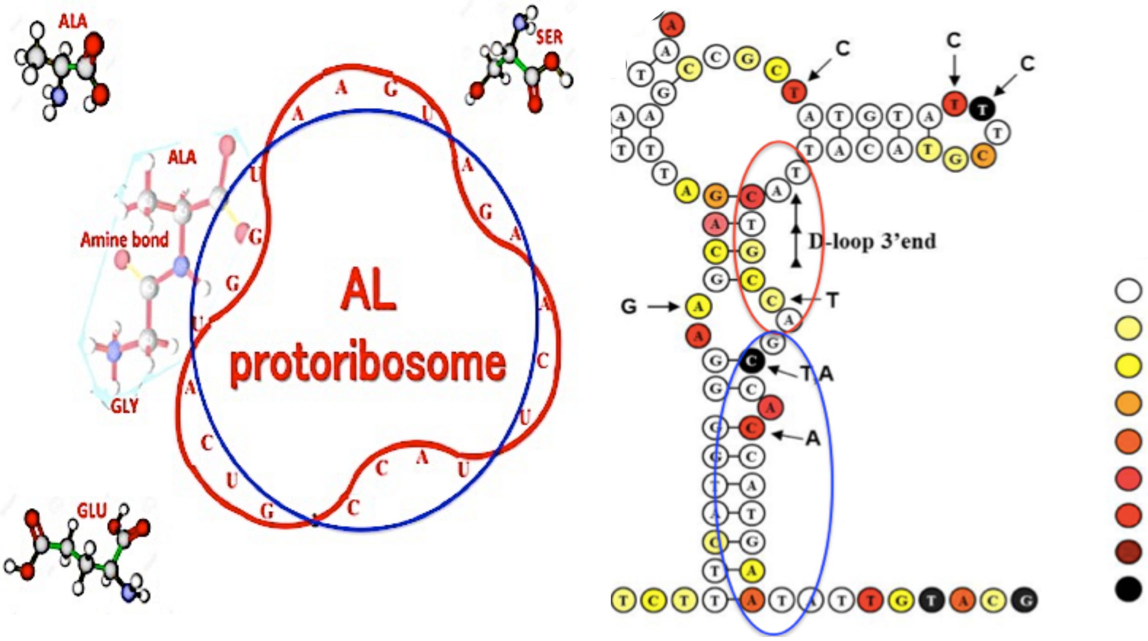


Fig. 4. AL as a protoribosome. (**A**) AL favors strong peptide bonds between amino-acids weakly fixed to their codons or anticodons in AL. (**B**) 2D-structure of the mitochondrial D-loop (7s mtDNA) with the central AL-dodecamer TACTGCCAGTCACCATGAAT[42].

Conclusion and Perspectives

To conclude, a small circular RNA, called AL, has been constructed from repeated sequences in the genome of *Rhodobacter sphaeroides*. AL presents the following features:

- its sub-sequences are observed as relics in many parts of the present genomes

- AL relics are present in tRNA loops and in many other RNA loops (IRE, YUNR, etc.)

- AL heptamers constitute the major part of the exon/intron boundaries

- AL has 22 nice combinatorial and biochemical properties in relation with amino-acids.

Hence, AL could have played the role of an ancient "protoribosome" (Fig. 4A): this claim is central in the stereochemical hypothesis of the genetic code formulated by A. Katchalsky in 1973[43]: the existence of catalytic RNAs in clays such as the "montmorillonite" may have facilitated the appearance of small peptides, involved secondarily (as now the protein replicase) in the replication of RNA molecules, hence constituting a virtuous loop at the origin of life. The existence of a simple RNA structure capable to survive in a stable hairpin form or to be functional in a ring form has been postulated early after Katchalsky's hypothesis[44-46], and experimental works[47-48] try to reinforce the stereochemical hypothesis, despite criticisms[49], showing that the subject is still open experimentally and theoretically. A future work could concern the existence of AL relics on both nucleic and proteic sides, which could reinforce their role at the origin of life, for example:

- AL-heptamer TCAAGAT is part of the palindromes located upstream of replicase genes in *Rhodobacterales* repABC-9 replicons, replication units of the alphaproteobacterial plasmids[50]

- hexadecameric peptides corresponding to 12 sequences of 16 successive AL codons (without overlap) from MVLPFKMNGTAIQDEWYCHSR to IQDEWYCHSRMVLPFKMNGTA are observed in 332 proteins for an expected number of $4\ 10^{-12} \pm 3\ 10^{-6}$* from NCBI Blast[50] among 117,262,330 proteins having a total number of 42,988,570,095 amino-acids. Among these 332 proteins, many come from extremophiles of the *Rhodobacteraceae* family, like *Roseivivax marinus*, *Ponticoccus litoralis*, *Thiobacimonas profunda* and *Tropicibacter naphthalenivorans*

- the icosamer TACTGCCAGCCACCATGAAT has a central role in the 2D-structure of the D-loop of the mitochondrial DNA (Fig. 4B) and matches the corresponding AL sequence with 4 mutations in practically all D-loops of the mitochondrial DNA of individuals human from 164 countries and ethnic groups, and great apes[51]. The icosamer anti-matches (in red) like a microRNA, a sequence of the mitochondrial translocase precursor ENSG00000154174[52], an enzyme crucial for the cell energetics in human cells:
GGTGTGAAGTGGCATCTCATGGTGGTTTTGGTTTGCATTTCCCTAATG Translocase
CCATGACATGTTATA AGTACCACAGACCGTCATTACATGCTTTATGT D-loop,
giving an argument in favor of the ancient role of AL as a negative feedback regulator in the early energetic cell metabolism[30,53,54]

- the CRISPR-CAS system provides bacteria like *Rhodobacter sphaeroides* with adaptive immunity and we can notice that the AL-pentamers ATGGT and ATTCA, and hexamers TCAAGAT and AATGGT (corresponding respectively to the Tψ– and D-loops of many tRNAs) are often observed at many levels of this system (CAS proteins, Casposon TIR and CRISP repeats). For example, the typical repeat sequences for CRISPR1 and CRISPR3[55] contain AL-heptamers from tRNA loops:

GTTTTTGTACTC**TCAAGAT**TTAAGTAACTGTACAAC (CRISPR1)
GTTTTAGAGCTGTGTTGTTTC**GAATGGT**TCCAAAAC (CRISPR3)

as well as the sequences of TIR and CRISPR compared in[56], a consensus sequence from central part of the murine RSS VκL8, Jß2.6 and Jß2.2[57-59], and the human RSS spacer common for Vh, V328h2 and V328[60-62]:

3'-A**T**A**CATCCC**(C)**TCTTAAGTTC**CC**TT**-5' (TIR)
3'-**TTCCATCCC** -**TCTTAAGTTC**G**ATT**-5' (CRISPR)
5'-**ATGGTACTG** - **CCATTCAAGATGA**-3' (AL)
5'-G**TG**A**TAC**A**G** - **CC**C**TT**A**A**CA**A**AA**A**-3' (murine consensus RSS)
5'-**ATTCAA**C**ATGAA**-3' (human RSS spacer)

The probability $p=2\ 10^{-9}$ for 19 matches (with an insertion) between TIR and CRISPR using the binomial distribution B(1/4,22), $p=8\ 10^{-6}$ for 15 anti-matches between AL and CRISPR plus 1 quasi-anti-match G-T using the distribution B(1/4,21)xB(3/8,1), $p=7\ 10^{-4}$ for 13 matches between AL and consensus RSS using the binomial distribution B(1/4,22), $p=2\ 10^{-6}$ for 11 matches between AL and RSS spacer using the binomial distribution B(1/4,12)

- There exist short sequences with many AL hexamers between deeply conserved noncoding regions associated with developmental genes controlling the neuro-differentiation in Metazoans with homologous regulatory states, as for the immune cells[63,64]

Sequence name	AL heptamers	Sequence length
Bf1Id	5'---**TCAAGA**CGGG**GATGAATGG**G**TA**---3'	92
HsaHMX3	5'---**TGCCATT**---3'	30
SpuMax	5'---**ACTGCC**---**AATGGT**---3'	96
HsaSIX1	5'---**TGGTA**GA**GCCATT**(A)**CAAGA**---3'	82

The probability $p = 10^{-112}$ for 10 observed AL hexamers from the Metazoan regulatory sequences knowing that the expected number is 0.07±0.44* using the normal approximation of the binomial distribution B(2.5 10^{-4}, 300).

To draw some perspectives of the present work, we propose three main directions:

- searching for more AL relics in the present genomes at critical functional steps of the nuclear transcription/translation, mitochondrial energetic or cellular immune receptor machineries,

- understanding the evolution of the immune systems (from CRISPR to RAG systems passing through TOLL), taking into account the reuse of former AL RNA fragments already present in the "protoribosome"
- using sequences linked to AL in the "minimal cell" studies.

References

1. Wolfram, S. *A New Kind of Science* (Wolfram Media, Inc., Champaign, IL, 2002).
2. Gardner, M. Mathematical Games: The Fantastic Combinations of John Conway's New Solitaire Game 'Life'. *Scientific American* **22**, 120-123 (1970).
3. GtRNAdb (2017) at <http://lowelab.ucsc.edu/GtRNAdb/Rhod_spha_ATCC_17029/rhodSpha_ATCC17029-tRNAs.fa>.
4. 5SRNAdb (2017) at <http://www.combio.pl/rrna/alignment/>.
5. GenBank (2017) at <https://www.ncbi.nlm.nih.gov/nuccore/636559472?report=fasta>.
6. Kinefold (2017) at <http://kinefold.curie.fr/>.
7. Rfam (2017) at <http://www.sanger.ac.uk/science/tools/rfam>.
8. Fonville, N. C. Velmurugan, K. R., Tae, H., Vaksman, Z., McIver, L. J. & Garner, H. R. Genomic leftovers: identifying novel microsatellites, over-represented motifs and functional elements in the human genome. *Sci. Rep.* **6**, 27722 (2016).
9. Tóth, G., Gáspári, Z. & Jurka, J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* **10**, 967-981 (2000).
10. Pemberton, T. J., Sandefur, C. I., Jakobsson, M. & Rosenberg, N. A. Sequence determinants of human microsatellite variability.*BMC Genomics*. **10**, 612 (2009).
11. Willard, H. F. & Waye, J. S. Chromosome-specific subsets of human alpha satellite DNA: analysis of sequence divergence within and between chromosomal subsets and evidence for an ancestral pentameric repeat. *J. Mol. Evol.* **25**, 207-214 (1987).
12. Zhuma, T., Tyrrell, R., Sekkali, B., Skavdis, G., Saveliev, A., Tolaini, M., Roderick, K., Norton, T., Smerdon, S., Sedgwick, S., Festenstein, R. & Kioussis, D. Human HMG box transcription factor HBP1: a role in hCD2 LCR function. *The EMBO Journal*. 18, 6396-6406 (1999).
13. Presnyak, V., Alhusaini, N., Chen, Y. H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K. E., Graveley, B. R. & Coller, J. Codon optimality is a major determinant of mRNA stability. *Cell* **160**, 1111-1124 (2015).
14 GenScript (2017) at <http://www.genscript.com/tools/codon-frequency-table>.
15. Demongeot, J. & Moreira, A. A circular RNA at the origin of life. *J. Theor. Biol.* **249**, 314-324 (2007).
16. Johnson, D. B. & Wang, L. Imprints of the genetic code in the ribosome. *Proc. Natl. Acad. Sci. USA.* **107**, 8298-8303 (2010).
17. Brooks, D. J., Fresco, J.R., Lesk, A. M. & Singh, M. Evolution of amino acid frequencies in proteins over deep time: inferred order of introduction of amino acids into the genetic code. *Mol. Biol. Evol.* **19**, 1645-1655 (2002).
18. Miller, S. L. A production of amino acids under possible primitive earth conditions. *Science* **117**, 528 (1953).
19. Johnson, A. P., Cleaves, H. J., Dworkin, J. P., Glavin, D. P., Lazcano, A. & Bada, J. L. The Miller volcanic spark discharge experiment. *Science* **322**, 404 (2008).
20. miRBase (2017) at <http://www.mirbase.org/>.
21. GtRNAdb (2017) at <http://gtrnadb.ucsc.edu/>.
22. Michaud, M., Cognat, V., Duchêne, A. M. & Maréchal-Drouard, L. A global picture of tRNA genes in plant genomes. *Plant J*. 66, 80-93 (2011).
23. GtRNAdb (2017) at <http://lowelab.ucsc.edu/GtRNAdb/Athal/Athal-structs.html>.
24. B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, *Molecular biology of the cell*, New York: Garland Science (2002).
25. Demongeot, J. *Au sujet de quelques modèles stochastiques appliqués à la biologie* (PhD thesis Université J. Fourier, Grenoble, 1975) at <https://tel.archives-ouvertes.fr/tel-00286222>.
26. Demongeot, J. *Revue de Biomaths* **62**, 61-66 (1978).
27. Demongeot, J. & Besson, J. Genetic code and cyclic codes I. *C.R. Acad. Sc. Série III* **296**, 807-810 (1983)
28. Demongeot, J. & Besson, J. Genetic code and cyclic codes II. *C.R. Acad. Sc. Série III* **319**, 520-528 (1996).

29. Demongeot, J., Drouet, E., Moreira, A., Rechoum, Y. & Sené, S. Micro-RNAs: viral genome and robustness of the genes expression in host. *Phil. Trans. Royal Soc. A* **367**, 4941-4965 (2009).
30. Demongeot, J., Hazgui, H., Bandiera, S., Cohen, O. & Henrion-Caude, A. MitomiRs, ChloromiRs and general modelling of the microRNA inhibition. *Acta Biotheoretica* **61**, 367-383 (2013).
31. Demongeot, J. & Hazgui, H. The Poitiers school of mathematical and theoretical biology: Besson-Gavaudan-Schützenberger's conjectures on genetic code and RNA structures. *Acta Biotheoretica* **64**, 403-426 (2016).
32. Slonimski, P. P. Periodic oscillations of the genomic nucleotide sequences disclose major differences in the way of constructing homologous proteins from different procaryotic species. *Comptes Rendus Biologies* **330**, 13-32 (2007).
33. NCBI (2017) at <https://www.ncbi.nlm.nih.gov/nuccore/NW_010697787>.
34. Festuccia, N., Dubois, A., Vandormael-Pournin, S., Gallego Tejeda, E., Mouren, A., Bessonard, S., Mueller, F., Proux, C., Cohen-Tanoudji, M. & Navarro, P. Mitotic binding of Esrrb marks key regulatory regions of the pluripotency network. *Nature Cell Biology* **18**, 1139-1148 (2016).
35. Ivanov, A., Memczak, S., Wyler, E., Torti, F., Porath, H. T., Orejuela, M. R., Piechotta, M., Levanon, E. Y., Landthaler, M., Dieterich, C. & Rajewsky, N. Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals. *Cell Reports* **10**, 170-177 (2014).
36. Circbase (2017) at <http://*www.circbase.org/*>.
37. Liang, T., Yang, C., Li, P., Liu, C. & Guo, L. Genetic Analysis of Loop Sequences in the Let-7 Gene Family Reveal a Relationship between Loop Evolution and Multiple IsomiRs. *PLoS ONE* **9**, e113042 (2014).
38. Watanabe, K. & Osawa, S. In *D.* Söll, *U.R. Bhandary* (eds), *tRNA*: *Structure*, *Biosynthesis, and Function* (ASM Press, Washington, DC: ASM Press, 165-205, 1995).
39. Subirana, J.A. & Messeguer, X. The most frequent short sequences in non-coding DNA. *Nucleic Acid Res.* **38**, 1172-1181 (2009).
40. Oliver, R. P. & Poulsen, C. Carlsberg Structure of a heavily transcribed region of barley chloroplast DNA. Transfer RNA genes from serine (UGA), glycine (GCC, UCC), formyl-methionine and threonine (GGU). *Res. Commun.* **49**, 647-673 (1984).
41. Ingman, M. & Gyllensten, U. mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. *Nucleic Acids Res.* **34**, D749-D751 (2006).
42. Portugene (2017) at <http://fpereira.portugene.com/research1.html>.
43. Paecht-Horowitz, M. & Katchalsky, A. Synthesis of amino acyl-adenylates under prebiotic conditions. *J. Mol. Evol.* **2**, 91-98 (1973).
44. Hobish, M. K., Wickramasinghe, N. S. & Ponnamperuma, C. Direct interaction between amino acids and nucleotides as a possible physicochemical basis for the origin of the genetic code. *Adv. Space Res.* **15**, 365-382 (1995).
45. Yarus, M., Widmann, J. J. & Knight, R. J. RNA–Amino Acid Binding: A Stereochemical Era for the Genetic Code. *Mol. Evol.* 69, 406-429 (2009).
46. M. Yarus, The Genetic Code and RNA-Amino Acid Affinities. Life. 7, 13 (2017).
47. Polyansky, A.A. & Zagrovic, B. Evidence of direct complementary interactions between messenger RNAs and their cognate proteins. *Nucleic Acids Res.* **41**, 8434-8443 (2013).
48. Polyansky, A. A., Hlevnjak, M. & Zagrovic, B. Proteome-wide analysis reveals clues of complementary interactions between mRNAs and their cognate proteins as the physicochemical foundation of the genetic code. *RNA Biol.* **10**, 1248-1254 (2013).
49. Meyer, S. C. & Nelson, P. A. Can the origin of the genetic code be explained by direct RNA templating? *Bio-complexity* **2011**, 1-10 (2011).
50. Petersen, J., Brinkmann, H. & Pradella, S. Diversity and evolution of repABC type plasmids in Rhodobacterales. *Environ. Microbiol.* **11**, 2627-2638 (2009).
51. 7s MtDB (2017) at <http://www.mtdb.igp.uu.se/>.
52. NCBI (2017) at <https://blast.ncbi.nlm.nih.gov/Blast.cgi>.
53. Demongeot, J., Cohen, O. & Henrion-Caude, A. In M.A. Aon, V. Saks, U. Schlattner (eds), *Systems Biology of Metabolic and Signaling Networks* (Springer, Berlin, 63-114, 2013).
54. Demongeot, J., Hazgui, H. & Henrion-Caude, A. In Q.N. Tran & H.R. Arabnia (eds), *Computational Biology, Bioinformatics & Systems Biology* (Elsevier, New York, 135-165, 2015).
55. Horvath, P., Romero, D. A., Coûté-Monvoisin, A.C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C. & Barrangou, R. Diversity, Activity, and Evolution of CRISPR Loci in Streptococcus thermophilus. *J. Bacteriology* **190**, 1401-1412 (2008).

56. Koonin, E. V. & Krupovic, M. Diversity, Activity, and Evolution of CRISPR Loci in Streptococcus thermophilus. *Nature Rev. Genetics* **16**, 184-192 (2015).
57. Lee, A. I., Fugmann, S. D., Cowell, L. G., Ptaszek, L. M., Kelsoe, G. & Schatz, D. G. A Functional Analysis of the Spacer of V(D)J Recombination Signal Sequences. *PLoS Biology* **1**, 56-69 (2003).
58. Pasqual, N., Gallagher, M., Aude-Garcia, C., Loiodice, M., Thuderoz, F., Demongeot, J., Ceredig, R., Marche, P. N. & Jouvin-Marche, E. Quantitative and Qualitative Changes in ADV-AJ Rearrangements During Mouse Thymocytes Differentiation: Implication For a Limited TCR ALPHA Chain Repertoire. *J. Exp. Medicine* **196**, 1163-1174 (2002).
59. Thuderoz, F., Simonet, M. A., Hansen, Dariz, O. A., Baum, T. P., Hierle, V., Demongeot, J., Marche, P. N. & Jouvin-Marche, E. From the TCRAD rearrangement quantification to the computational simulation of the locus behavior. *PloS Comp. Biol.* **6**, e1000682 (2010).
60. A.Ramsden, D. A., Baetz, K. & E.Wu, G. E. Conservation of sequence in recombination signal sequence spacers.*Nuclear Acids Research* **22**, 1789-1796 (1994).
61. Takeuchi, N., Ishiguro, N. & Shinagawa, M. Molecular cloning and sequence analysis of bovine T-cell receptor gamma and delta chain genes. *Immunogenetics* 35, 89-96 (1992).
62. Liu, M. F., Robbins, D. L., Crowley, J. J., Sinha, S, Kozin, F. Kipps, T. J., Carson, D. A. & Chen, P. P. Characterization of four homologous L chain variable region genes that are related to 6B6.6 idiotype positive human rheumatoid factor L chains. *J. Immunol.* **142**, 688-694 (1989).
63. Royo, J. L., Maeso, I., Irimia, M., Feng Gao, F., Peter, I. C., Lopes, C. S., Aniello, S. D., Casares, F., Davidson, E. H., Garcia-Fernández, J. & Gómez-Skarmeta, J. L. Transphyletic conservation of developmental regulatory state in animal evolution. *Proc. Natl. Acad. Sci. USA.* **108**, 14186-14191 (2011).
64. Royo, J. L., Maeso, I., Irimia, M., Feng Gao, F., Peter, I. C., Lopes, C. S., Aniello, S. D., Casares, F., Davidson, E. H., Garcia-Fernández, J. & Gómez-Skarmeta, J. L. Supporting Information. *Proc. Natl. Acad. Sci. USA.* **108 Suppl.**, 1-10 (2011) DOI: 10.1073/pnas.1109037108.

Acknowledgments:

We are very indebted to Alexandra Henrion Caude and Andres Moreira for many discussions about the molecular structures potentially involved at the origin of life.

The data reported in the paper are archived on the sites GtRNAdb, 5SRNAdb, GenBank, Protein Data Bank, Bacterial 16S rRNA Bioproject, Circbase, NCBI Blast and Kinefold or available in supplementary materials. The sites used are referenced as following:

Chan, P. P. & Lowe, T. M. GtRNAdb: A database of transfer RNA genes detected in genomic sequence. *Nucl. Acids Res.* **37**, 93-97 (2009).

Lowe, T. M. & Eddy, S. R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucl. Acids Res.* **25**, 955-964 (1997).

Szymanski M., Barciszewska M. Z., Erdmann V. A. & Barciszewski J. 5S Ribosomal RNA Database. *Nucl. Acids Res.* **30**, 176-178 (2002).

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucl. Acids Res.* **41**, 36-42 (2013).

Heinz, D. W., Baase, W. A., Dahlquist, F. W. & Matthews, B. W. How Amino-Acid Insertions are Allowed in an Alpha-Helix of T4 Lysozyme. *Nature* **361**, 561(1993).

Boudjeko, T., Tchinda, R. A. M., Zitouni, M., Nana, J. A. V. T., Lerat, S. & Beaulieu, C. Streptomyces cameroonensis sp. nov., a Geldanamycin Producer That Promotes Theobroma cacao Growth. *Microbes and Environments* **32**, 24-31 (2017).

Glažar P., Papavasileiou P. & Rajewsky N. circBase: a database for circular RNAs. *RNA* **20**, 1666-1670 (2014).

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410 (1990).

Gish, W. & States, D. J. Identification of protein coding regions by database similarity search. *Nature Genet.* **3**, 266-272 (1993).

Xayaphoummine A., Bucher T. & Isambert H. Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucl. Acid Res.* **33**, 605-610 (2005).