“Protoribosome” as new game of life

Jacques Demongeot

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 Archaea, 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.

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. 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. 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 CATGAATGGTACTTCCATTCA of Gly-tRNA\(^{TCC}\) loops\(^3\), AATGGTAAGTGCTCTCAAGAG from 5s rRNA\(^4\), GCCGTAAACATGAATGGCCAGT from 16s rRNA\(^5\), AATTGAACTGCTGTGCAAGATG from 23s rRNA\(^5\), ACGGCCACCTTCTGTCGAGATG from 30s rRNA\(^5\), TCGACCCGCGCCATTCTCGATCA from 50s rRNA, GGTCAGGTCATCAAGATCAA from rpsA1\(^5\) and GATGTTGCCCACTTTGCGCATTCA from CRISPR CAS1\(^5\) from the genome of *Rhodobacter sphaeroides*, we get a cyclic RNA called AL (Archetypal Loop): GATGAATGGTACTGCCATTCA.

![Construction of the Archetypal Loop AL](image)

Fig. 1. Construction of the Archetypal Loop AL. (A) AL hemi-sequence ATGAATGGTACT and hexamer CCATTTC 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\(^6\) algorithm, 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 T\(\psi\)-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 T\(\psi\)-loop of many present tRNAs, *i.e.*, AAUGG, AUGGU, UGGUA, GGUAC, TTCAAG, 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\textsuperscript{13}. We can generalize the result obtained in Rhodobacter sphaeroides to other species in worlds like Archaea, Bacteria and Eukaryotae, by searching similar sequences:

\textbf{Rhodobacter sphaeroides}

AATGTTACTCCATCGATTG \textbf{tRNA-Gly}
GATGTTGCCGACTGGGCGATG CRISPR CAS1
AATGGTACTGCTCTCAAGACG 5S rRNA
CTGGAACTCCATCGAAACTCT 16S rRNA
AATTTGACCTCGTCAAGATG 23S rRNA
ACGGCCACCTTGCAAAGATG 30S rRNA
TCAGCCGGCCCATCCTGCTA 50S rRNA
AATGTTACCCATCGATAAAGATCAA Ribosomal protein S1 (rpsA1)

\textbf{Rhodobacter vinaykumarii}

AATGTTACTCCATCGATCTG \textbf{tRNA-Gly}
CTGGAACTCCATCGAAACTCT 16S rRNA
AATGGTACTGAGCTAAGATG CRISPR CAS4
AATGTTACTGCGCAATCGATA 5S rRNA
CTGGAACTCCATCGAAACTCT 16S rRNA
AATGGTACTGCGCAATCGATA Consensus

\textbf{Haematobacter missouriensis}

AATGTTACCCATCGATAAAGATCAA Ribosomal protein S1 (rpsA1)

\textbf{Rubellimicrobium thermophilum}

AATGTTACCCATCGATAAAGATCAA Ribosomal protein S1 (rpsA1)

\textbf{Rhodosporillum rubrum}

AATGTTACCCATCGATAAAGATCAA Ribosomal protein S1 (rpsA1)

\textbf{Haematobacter missouriensis}

AATGTTACTGCCAAGATG CRISPR CAS9

\textbf{Bacteriae and Eukaryotae, by searching similar sequences:}

\textbf{Haematobacter missouriensis}

AATGTTACCCATCGATAAAGATCAA Ribosomal protein S1 (rpsA1)
From the above sequences, we can obtain the following consensus sequences:

**Consensus Bacteria**
AATGGTACTGCCAT

**Consensus Archae**
ATTGTACTGCCATCAGAG

**Consensus Eukaryote**
AATGGTACTCCCATCAATAG

Another way to obtain AL is by remarking that the most frequent sequence shared by all above sequences is GAAUGG GU. 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: AUGAAUGGU 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*), 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 is located at the level of its first base, e.g., the red arrow indicates the frequency of AAUGG. 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 in Table 1B and 1C, but it is the least frequent in Table 1A and absent in the Miller experiment), 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.

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

Table 1. Frequency and affinities of amino-acids. (A) Amino-acid frequencies in 9 species\(^\text{17}\). (B) Inside the ribosome of 4 different species and in presence of amino-acids, relative enrichment of their codons and (C) anticodons\(^\text{16}\). 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**

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.

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 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.

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

11. There are more than 50% of the tRNAs of GtRNAdb 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**

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).

AL matches well with exon/intron boundaries\(^{24}\)

One of the 19 rings of property 4, AB\(^{25-28}\), fits well the circular AL (Fig. 3F)

AL matches with Hamming and edit distances \(\leq 2\) with at least 30 tRNA-Gly from GtRNAdb (supplementary material 2)

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

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

AL matches well with many viral not coding genomes\(^{29}\)

AL matches well with many microRNAs\(^{30}\)

AL matches well with IRE and YUNR loops\(^{31}\)

AL matches well with many circular RNAs\(^{31}\)

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 \((1/4)^{18} \approx 1.5 \times 10^{-11}\), with an expected number of \(7.62 \pm 4.56\) (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*\(^{39}\).
Fig. 3. AL Relics in current genomes. (A) Trp tRNA\textsuperscript{TGA} of Aspergillus nidulans\textsuperscript{38}, (B) GlytRNA\textsubscript{GCC} of Arabidopsis thaliana\textsuperscript{23}, (C) GlytRNA\textsubscript{GCC} of Hordeum vulgare\textsuperscript{40} and (D) GlytRNA\textsubscript{GCC} of \OE nothera coquimbensis\textsuperscript{22}, whose loops (D-, anti-codon and T\textsubscript{\psi}-loops) fit quasi-perfectly AL (E) Consensus sequences at exon/intron boundaries in eukaryotes\textsuperscript{24}. (F) Quasi-perfect matches between AB and a sequence of a circular RNA\textsuperscript{36}, and between AL and a sequence of the genome of Collius stria\textsuperscript{33}. 

hsa_circ_0015360 (reverse strand): GGCAATTCAGACTATGAATGTT

AB:

GGCATTCAGACTATGAATGTT

GCCATTCAGACTATGAATGTT

Collius striatus isolate BG\textsubscript{I} N325 21746-21766: AATGGTACTGCCATTCAAGATG

AL:

AATGGTACTGCCATTCAAGATG

Fig. 3. AL Relics in current genomes. (A) Trp tRNA\textsuperscript{TGA} of Aspergillus nidulans\textsuperscript{38}, (B) GlytRNA\textsubscript{GCC} of Arabidopsis thaliana\textsuperscript{23}, (C) GlytRNA\textsubscript{GCC} of Hordeum vulgare\textsuperscript{40} and (D) GlytRNA\textsubscript{GCC} of \OE nothera coquimbensis\textsuperscript{22}, whose loops (D-, anti-codon and T\textsubscript{\psi}-loops) fit quasi-perfectly AL (E) Consensus sequences at exon/intron boundaries in eukaryotes\textsuperscript{24}. (F) Quasi-perfect matches between AB and a sequence of a circular RNA\textsuperscript{36}, and between AL and a sequence of the genome of Collius striatus\textsuperscript{33}.
**Table 4.** Highly unlikely matches between sub-sequences from three circular permutations of AL and from different nucleic databases of NCBI sources.\(^{33}\)

<table>
<thead>
<tr>
<th>Source</th>
<th>Nb sequences observed of length ≥18</th>
<th>Nb sequences observed of length ≥10</th>
<th>Nb sequence observed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NCBI sources</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>refseq_genomic</td>
<td>70 (p=10^{−112})</td>
<td>49 (p=10^{−168})</td>
<td>TACTGCCA GT CA CC ATGAAT</td>
</tr>
<tr>
<td>Human ALU repeat</td>
<td>140 (p=10^{−988})</td>
<td>50 (p=10^{−173})</td>
<td></td>
</tr>
<tr>
<td>Protein Data Bank (PDB)</td>
<td>98 (p=10^{−234})</td>
<td>7 (p=10^{−86})</td>
<td></td>
</tr>
<tr>
<td>16s ribosomal</td>
<td>4 (p=10^{−3})</td>
<td>112 (p=10^{−86})</td>
<td></td>
</tr>
<tr>
<td>7s mtDB</td>
<td>10^{−30.05}</td>
<td>138 (p=10^{−110})</td>
<td></td>
</tr>
</tbody>
</table>

*p* 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 \(X\): \(p=P(|X-\mu|≥\kappa\sigma)\leq\exp(-\kappa^2/2)k\sqrt{2\pi}=10^{-y}\), where \(y=0.4+\text{Log}_{10}k+0.217k^2\).

**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 TACTGCCAGTCA CCATGAAT\(^{42}\).
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 MVLPFKMNGTAIMQDEWYCHSR to IQDEWYCHSRMVLPFKMNGTA are observed in 332 proteins for an expected number of \(4 \times 10^{12} \pm 3 \times 10^{9}\) 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*, *Ponticococcus litoralis*, *Thiobacimonas profunda* and *Tropicibacter naphthalenivorans*

- the icosamer TACTGCCAGCCAACCATGAAT 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:

  **GGTGTGAGTGGCATCCTACGGTGGTTTGGTTTCATTTCCCTAATG** Translocase
  **CCATGACATGCTTTATGT** D-loop,

giving an argument in favor of the ancient role of AL as a negative feedback regulator in the early energetic cell metabolism\(^30\),\(^34\),\(^35\),\(^54\)
- 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 contain AL-heptamers from tRNA loops:

\[
\text{GTTTTGTACTCTCAAGATTTAAGTAACTGTACAAC (CRISPR1)}
\]

\[
\text{GTTTTAGAGCTGTGTTGTTTCGAATGGTTCCAAAC (CRISPR3)}
\]

as well as the sequences of TIR and CRISPR compared in^55^, a consensus sequence from central part of the murine RSS VκL, Jß2.6 and Jß2.2^57–59^, and the human RSS spacer common for Vh, V328h2 and V329^60–62^.

\[
3'\text{-ATACATCCCC(C)TCTTTAGTTCCCTT-5'} \text{ (TIR)}
\]

\[
3'\text{-TTCCATCCC-TCTTTAGTTCCATT-5'} \text{ (CRISPR)}
\]

\[
5'\text{-ATGGTACTG-CCATTCAGATGA-3'} \text{ (AL)}
\]

\[
5'\text{-GTGATAACGCCTTAAACAAA-3'} \text{ (murine consensus RSS)}
\]

\[
5'\text{-ATTCAACATGAA-3'} \text{ (human RSS spacer)}
\]

The probability \( p = 2 \times 10^{-9} \) for 19 matches (with an insertion) between TIR and CRISPR using the binomial distribution \( B(1/4,22) \), \( p = 8 \times 10^{-6} \) for 15 anti-matches between AL and CRISPR plus 1 quasi-anti-match G-T using the distribution \( B(1/4,21) \times B(3/8,1) \), \( p = 7 \times 10^{-4} \) for 13 matches between AL and consensus RSS using the binomial distribution \( B(1/4,22) \), \( p = 2 \times 10^{-6} \) 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 neurodifferentiation in Metazoans with homologous regulatory states, as for the immune cells^63,64^.

<table>
<thead>
<tr>
<th>Sequence name</th>
<th>AL heptamers</th>
<th>Sequence length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bf1ld</td>
<td>5'---TCAAGACGGGATGAATGGGTA---3'</td>
<td>92</td>
</tr>
<tr>
<td>HsaHMX3</td>
<td>5'---TGCCATT---3'</td>
<td>30</td>
</tr>
<tr>
<td>SpuMax</td>
<td>5'---ACTGAC---AATGGT---3'</td>
<td>96</td>
</tr>
<tr>
<td>HsaSIX1</td>
<td>5'---TGGTAGAGCCATT(A)CAAGA---3'</td>
<td>82</td>
</tr>
</tbody>
</table>

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

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).
51. 7s MtDB (2017) at <http://www.mtdb.igp.uu.se/>.

**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:


