Mathematical Biosciences and Engineering

MBE, 21 (1): 884–902.
DOI: 10.3934/mbe.2024037
Received: 15 October 2023
Revised: 26 November 2023
Accepted: 27 November 2023
Published: 21 December 2023

http://www.aimspress.com/journal/MBE

Research article

Combinatorial and frequency properties of the ribosome ancestors

Jacques Demongeot[1,*], Jules Waku[2] and Olivier Cohen[1]

[1] AGEIS, Faculty of Medicine, University Grenoble Alpes, 38700 La Tronche, France
[2] IRD UMI 209 UMMISCO and LIRIMA, University of Yaoundé I, P.O. Box 337, Yaoundé, Cameroon

*** Correspondence:** Email: Jacques.Demongeot@univ-grenoble-alpes.fr.

Abstract: *Background:* The current ribosome has evolved from the primitive stages of life on Earth. Its function is to build proteins and on the basis of this role, we are looking for a universal common ancestor to the ribosome which could: i) present optimal combinatorial properties, and ii) have left vestiges in the current molecules composing the ribosome (rRNA or r-proteins) or helping in its construction and functioning. *Methods:* Genomic public databases are used for finding the nucleotide sequences of rRNAs and mRNA of r-proteins and statistical calculations are performed on the occurrence in these genes of some pentamers belonging to the RNA proposed as optimal ribosome ancestor. *Results:* After having exhibited a possible solution to the problem of an RNA capable of catalyzing peptide genesis, traces of this RNA are found in many rRNAs and mRNA of r-proteins, as well as in factors contributing to the construction of the current ribosome. *Conclusions:* The existence of an optimal primordial RNA whose function is to facilitate the creation of peptide bonds between amino acids may have contributed to accelerate the emergence of the first vital processes. Its traces should be found in many living species inside structures structurally and functionally close to the ribosome, which is already the case in the species studied in this article.

Keywords: genome combinatorics; peptide genesis; primitive ribosome; AL-proximity; protein building; ribosome construction

1. Introduction

The numerous theories dealing with the problem of the origin of life differ in focusing each on a preponderant factor of emergence: i) the main primitive function (catalysis versus replication), ii) the initial location of the first living systems (black smokers versus ponds) or iii) the first molecule

involved in the origin of life (RNA versus DNA). In the last opposite theories, RNA first assumes that a primitive "RNA world" has existed, based on the collaboration between nucleic and amino-acids polymers, the organic synthesis of which being obtained in absence of any living process. Specialized RNA molecules functioning like enzymes suggest that RNA-based proto-cells have evolved before there were functional proteins. These proto-cells contained a large number of RNA molecules, with different catalytic functions, the peptide synthesis being catalyzed by some of these RNA molecules [1]. This RNA World could have appeared in the early history of life on Earth when RNA processes emerged from chemistry, carrying most of the information needed by the future metabolic transformations necessary for the emergence of life. This scenario could have taken place 3.5 billion years ago [2]. The present a day ribosome works to build the proteins needed by the cell that houses it and this in many vital metabolisms (energetic, secretory, sensory, etc.), depending on the specific function of the concerned cell. The ribosome was built gradually during evolution and its current components are the result of this evolution. If we examine the heart of its function, namely the peptide elongation, source of the final protein to be synthesized, we can think that primitive mechanisms were at the origin both i) bringing together amino-acids using weak binding to a primitive RNA and then ii) favoring the covalent binding of these amino-acids in small peptides. Following previous works [3–8], we exhibit in Section 2 Material and Methods a primitive RNA candidate, capable of catalyzing this synthesis. In Section 3 Results, we look for traces of it in molecules which currently either are part of the ribosome (rRNAs and r-proteins), or contribute to the construction of the ribosome in the cell. We discuss in Section 4 Discussion the limits of this approach and we conclude in Section 5 Conclusion.

2. Materials and methods

2.1. The genomic databases

The methods implemented in the present article involve the use of public databases relating to the genomes of numerous species. These databases are listed on Table 1.

Table 1. Web sites for data sources, accessed on 15 May 2023 [9–15].

tRNA databases	http://lowelab.ucsc.edu/GtRNAdb/Lafri3/Lafri3-align.html http://trna.bioinf.uni-leipzig.de/DataOutput/Result http://trna.ucsc.edu/tRNAviz/summary/
Codon frequency	https://www.genscript.com/tools/codon-frequency-table
Secondary structure	http://kinefold.curie.fr/cgi-bin/neorequest.pl?batch=0&sim=2&base=AL- 73086 https://en.vectorbuilder.com/tool/dna-secondary-structure/
Gene sequences	https://www.ncbi.nlm.nih.gov/nucleotide?cmd=search

2.2. The primitive ribosomal ancestor

In previous works [3–8], we have defined an RNA molecule of length 22 existing in ring or hairpin form and called AL (for Ancestral Loop), whose core comes from sequences of tRNA-loops common to multiple procaryote or eucaryote species: 5'-UGA(A)UGGUACUGCCAUUCAA(G)-3', where (A) and (G) may be missing (Figure 1, Table 2 below and Table S1 in Supplementary Material).

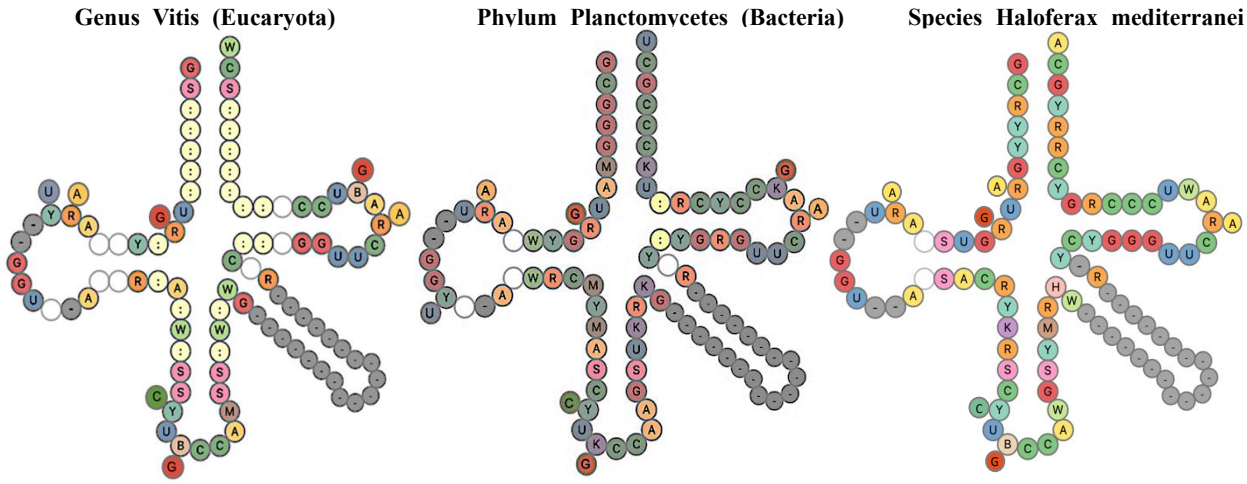

Figure 1. Secondary structure of tRNA-Gly^{GCC} of genus Vitis, phylum Planctomycetes and species Haloferax mediterranei (from tRNAviz [11]).

Table 2. tRNA loops (in red) from 11 different species (Archaea, Bacteria, Eucaryota) [9,10].

Species	tRNA-Gly^{GCC}
Methanococcus maripaludis	GCGGCTT**TG**ATGTAG**AC****TGGTA**TCATACGGCC**CTGCCA**CGGCCGACACCCGGG**TTCAA**ATCCCGGAGGCCGCA
Halorhabdus utahensis	GCGACGG**TG**GTGT**AG****TGGTA**TCACAGGACC**CTGCCA**CGGTCCTAACCCGAG**TTCAA**ATCTCGGCCGTCGCA
Termite group 1 bacterium	GCGGGTG**TA**G**T**TC**AG****TGGTA**GAACGTCTCGT**TGCCA**ACGAGAAGGtCGTGGG**TTCAAG**TCCCATCGCCCGCT
Vitis vinifera (grape)	GCGGAA**AT**A**G**CTT**AATGGTA**GAGCATAGCCT**TGCCA**AGGCTGAGGtTGAGGG**TTCAAG**TCCCTCCTTCCGCT
Arabidopsis thaliana (plant)	GCACCAG**TG**GTCT**AG****TGGTA**GAATAGTACC**CTGCCA**CGGTACAGACCCGGG**TTCAA**TTCCCGGCTGGTGCA
Medicago truncatula	GCACCAG**TG**GTCT**AG****TGGTA**GAATAGTACC**CTGCCA**CGCTACAGACCCGGG**TTCAA**TTCCTGGCTGGTGCA
Petromyzon marinus (lamprey)	GCATCGG**TG**GTTC**AG****TGGTA**GAAATCTCGC**CTGCCA**CGCGGGAGGCCCGGG**TTCAA**TTCCCGGCCGATGCA
Danio rerio (zebrafish)	ACATTGG**TG**GTTC**AG****TGGTA**GATTTCTCGC**CTGCCA**CGTGGGAGGCCCGGG**TTCAA**TTCCCGGCCAATGCA
Strongylocentrotus p. (sea urchin)	GCATTGG**TG**GTTC**AG****TGGTA**GAATTCTCGC**CTGCCA**CGCGGGGGACCCGGG**TTCAA**TTCCCGGCCAATGCA
Loxodonta africana (elephant)	GCATTGG**TG**GTTC**AG****TGGTA**GAATTCTCGC**CTGCCA**CGTGGGAGGCCTGGG**TTCAA**TTCCCAGCCAGTTCT
Callithrix jacchus (marmoset)	GCATGGG**TG**GTTC**AG****TGGTA**GAATTCTCGC**CTGCCA**CGCGGGAGTCCTGGG**TTCAA**TCCCCGGCCCACGCA

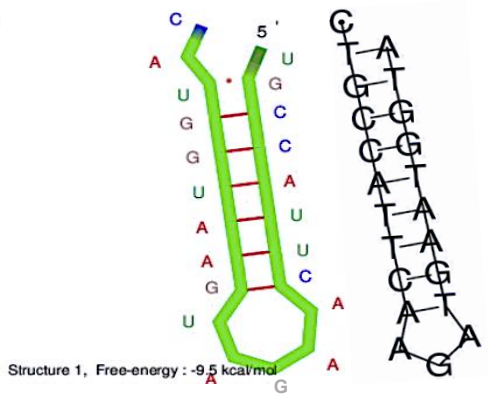

Figure 2. Most stable hairpin from tRNA loops obtained using Kinefold® [13] (left) and Vectorbuilder® (right) [14].

By searching for the most stable hairpin structure built from AL sequence, we find the following architecture, with head AGA, one branch made by end of AL and the other by start of AL (Figure 1).

2.3. Optimal combinatorial properties of AL

The ring form of AL presents one and only one representative of each synonymous codons class of genetic code and is barycenter of only 25 rings with this property, starting with AUG, ending with UGA and having the most stable hairpin structure. Steps of the proof of these optimal combinatorial properties are summarized in Figure 3, starting with 10^{22} possible solutions and ending with only one.

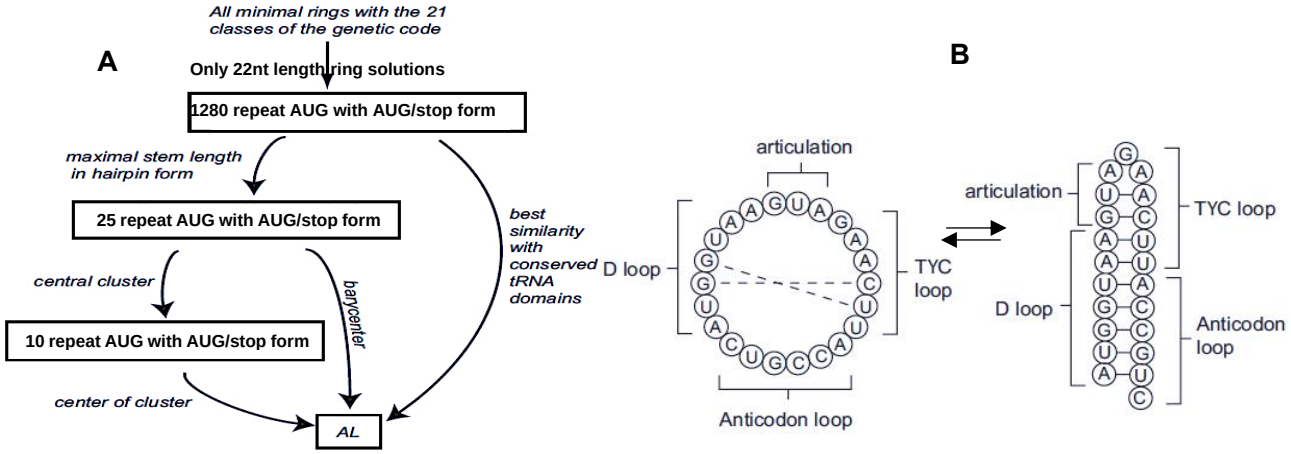

Figure 3. A) Schematic summary of the search for the RNA AL; B) ring and hairpin structure of AL.

2.4. Occurrence of AL codons in whole genomes

The majority of the most frequent codons in complete genome of many species are those belonging to the AL sequence (Table 3).

2.5. Pentamer proximity

The RNA sequence at the upper part of the hairpin form of AL is the easiest to be fragmented, because each of its 9 consecutive upper pentamers (a pentamer is a sequence of 5 successive nucleotides) has strictly less bonds than the previous and following pentamers (especially less GC bonds), then we will use the difference between observed and expected numbers of these 9 pentamers in a sequence S divided by the standard deviation of the expected number as the "AL-proximity" of S. The probability of observing by chance in sequence S a pentamer is p = 1/1024, then the expected number of pentamers from the set of the 9 upper pentamers of AL in S of length n = 2724 is equal to np = 2720 × (9/1024) = 23.9, with a standard deviation $\sigma = [np(1-p)]^{1/2} \sim 23.9^{1/2} \sim 4{,}9$. If the number of such observed pentamers in S equals 95, the difference X between expected and observed numbers equals 95–23,9=71,1 and X/σ= 14,5. Bernoulli distribution of X verifies the conditions for a Gaussian approximation: n=2720 ≥ 30, np ~ 23,9 ≥ 5 and n(1−p) ~ 2696 ≥ 5, so the probability of observing such a standardized difference X/σ is less than 1-F(14,5) < Proba(X/σ ≥ 14,5), where F is the standard Gaussian distribution function. Then, by using the Gaussian approximation proposed in [16], we get:

$$\text{Proba}(X/\sigma \geq t) \sim [1 - (1-\exp(-at^2))^{1/2}]/2, \text{ where } a = 0{,}647 - (0{,}021)\, t.$$

Hence, if t=14,5, a=0,3425, and Proba(X/σ ≥ 14,5) ~ exp(-0,3425x210,25)/4 ~ 1,3 10^{-32}. More generally, the value of the probability to observe, in a genetic sequence of length n, a number of

pentamers coming from the 9 top AL pentamers more than E + tσ (where E = (n–4)9/1024 is the expected number and $\sigma = [(n-4)9/1024(1-9/1024)]^{1/2}$ the standard deviation) is given in Table 4.

Table 3. Frequencies of all the 64 codons for two species, Saccharomyces cerevisiae (FSc) and Escherichia coli (Fco), with number (Nb) and percentage (%) in their whole genome [15]. Most frequent codons from AL or close to most frequent (≤ 20% less frequent) are in bold.

Co-don	Amino-acid	Frequency Saccharomyces cerevisiae	Frequency Escherichia coli	Co-don	Amino-acid	Frequency Saccharomyces cerevisiae	Frequency Escherichia coli
TTT	F	0.59	0.58	TCT	S	0.26	0.17
TTC	F	**0.41**	**0.42**	TCC	S	0.16	0.15
TTA	L	0.28	0.14	TCA	S	**0.21**	0.14
TTG	L	0.29	0.13	TCG	S	0.10	0.14
TAT	Y	0.56	0.59	TGT	C	0.63	0.46
TAC	Y	**0.44**	**0.41**	TGC	C	0.37	**0.54**
TAA	*	0.48	0.61	TGA	*	0.30	0.30
TAG	*	0.22	0.09	TGG	W	**1**	**1**
CTT	L	0.13	0.12	CCT	P	0.31	0.18
CTC	L	0.06	0.19	CCC	P	0.15	0.13
CTA	L	0.14	0.04	CCA	P	**0.42**	0.20
CTG	L	0.11	**0.47**	CCG	P	0.12	0.49
CAT	H	**0.64**	**0.57**	CGT	R	0.14	0.36
CAC	H	0.36	0.43	CGC	R	0.06	0.36
CAA	Q	**0.69**	0.34	CGA	R	0.07	0.07
CAG	Q	0.31	0.66	CGG	R	0.04	0.11
ATT	I	**0.46**	**0.49**	ACT	T	**0.35**	0.19
ATC	I	0.26	0.39	ACC	T	0.22	0.40
ATA	I	0.27	0.12	ACA	T	0.30	0.17
ATG	M	**1**	**1**	ACG	T	0.14	0.25
AAT	N	**0.59**	**0.49**	AGT	S	0.16	0.16
AAC	N	0.41	0.51	AGC	S	0.11	0.25
AAA	K	0.58	0.74	AGA	R	**0.48**	0.07
AAG	K	**0.42**	0.26	AGG	R	0.21	0.04
GTT	V	0.39	0.28	GCT	A	0.38	0.18
GTC	V	0.21	0.20	GCC	A	0.22	**0.26**
GTA	V	0.21	0.17	GCA	A	0.29	0.23
GTG	V	0.19	0.35	GCG	A	0.11	0.33
GAT	D	**0.65**	**0.63**	GGT	G	**0.47**	**0.35**
GAC	D	0.35	0.37	GGC	G	0.19	0.37
GAA	E	**0.70**	**0.68**	GGA	G	0.22	0.13
GAG	E	0.30	0.32	GGG	G	0.12	0.15

Table 4. Probability to observe a number of the 9 top AL pentamers more than E + tσ.

X/σ = t	2	4	6	10	13	14.5
a = 0,647 – (0,021) t	0.605	0.563	0.521	0.437	0.374	0.3425
$E=\exp(-at^2)$	0.089	0.000122	$7.2\ 10^{-9}$	$1.05\ 10^{-19}$	$3.55\ 10^{-28}$	$5.32\ 10^{-32}$
$P(X/\sigma \geq t) \sim [1-(1-E)^{1/2}]/2$	0.0227	$3\ 10^{-5}$	$1.8\ 10^{-9}$	$2.6\ 10^{-20}$	$8.9\ 10^{-29}$	$1.3\ 10^{-32}$

The value of the difference between the observed and expected numbers of pentamers from the 9 top pentamers of AL, expressed as a number t of standard deviations σ, is directly linked to the probability of observing this difference for a standard Gaussian variable. Then, we retain this quantity as measure of "AL-proximity". Because AL is proposed as a primitive RNA structure, this AL-proximity can be considered as a level of ancestrality.

3. Results

3.1. Ribosomal proteins and rRNAs components of the current ribosomes

For strengthening the hypothesis that AL ring is an ancient structure, we calculate the AL-proximity of ribosomal RNAs and proteins (RP) ordered following their anteriority. This anteriority has been defined in [17] through their relationship to the origin and evolution of the ribosome. Authors of [17] have used a phylogenetic comparative framework to study the ribosomal evolution and shown that contrarily to the previous observations, the primitive ribosome did not originate in the peptidyl transferase center of the large ribosomal subunit. By applying phylogenetic methods to RNA structures of thousands of molecules in hundreds of genomes, they found that components of the small subunit involved in ribosomal construction evolved earlier, starting with the oldest ribosomal proteins S12 and S17. This new ribosomal historical study has shown the existence of a gradual structural accretion of ribosomal proteins and RNA structures, suggesting that "functionally important and conserved regions of the ribosome were (progressively) recruited and could now be relics of an ancient ribonucleoprotein world" [17]. By seizing this new concept, we have investigated whether the hierarchy of riboproteins and ribosomal RNAs obtained by these authors were compatible with their content of pentamers coming from the terminal zone of the hairpin structure of AL. We therefore calculated their AL-proximity and compared it to the descending order from the most recent to the oldest proposed in [17]. This comparison is presented in the Figure 4.

On Table 5, the mean M_o (resp. M_n) of AL-proximities for the 33 oldest (resp. newest) ribosomal RNAs and proteins (after Gustavo Caetano-Anollés in [17]) is equal to 3.81 (resp. 2.4) and the corresponding standard deviation is equal to 89.28 (resp. 98.67). By applying a *t*-test of comparison of means, the most ancient rRNAs are closer to AL than the earliest ones (p=.001).

Indeed, the *t*-value is given by: $t = (M_o - M_n)/(\sigma_o^2 + \sigma_n^2)^{1/2} = 1.41/(0.18)^{1/2} = 3.34$, then the *p*-value equals 0.0007 [18].

HS	SC	MT	RP
3.1	2.7	2.0	S15
3.9	0.0	1.1	L23
2.6	1.5	0.9	L15
0.8	2.3	0.9	L30
3.7	1.6	4.3	L1
4.9	1.7	4.5	L4
0.7	4.2	0.5	L22
0.0	1.7	1.4	L11
2.2	7.5	0.8	S10
0.0	2.0	3.0	L13
5.1	3.3	4.2	L14
2.7	3.2	1.9	S19
3.3	2.8	5.3	S2
4.0	4.0	1.2	L5
1.0	6.9	2.5	S8
2.2	4.2	1.7	S7
3.6	7.8	1.8	S3
1.0	1.2	1.8	L29
1.7	3.8	0.8	L7/L12
1.0	7.4	7.0	S5
7.6	4.6	2.9	S11
3.1	7.0	7.6	L18
2.6	3.1	3.6	S14
2.8	2.5	3.7	L6
1.6	3.9	1.3	S13
3.4	2.3	3.5	S4
1.9	2.0	1.7	L24
2.5	3.6	4.9	L2
4.4	5.0	5.2	L3
3.2	6.4	2.7	S9
3.7	4.0	3.7	S17
5.6	5.8	7.3	S12

Figure 4. AL-proximity of ribosomal RNAs and proteins (RP) from Homo sapiens (HS in green), Saccharomyces cerevisiae (SC in red) and Marine Group I thaumarchaeote YK1309 (MT in brown) listed from the earliest (top) to the most ancient (bottom) RP during the evolution [17].

Table 5. Comparison between the AL-proximities of the 33 most ancient ribosomal proteins (old X_i) and the 33 earliest (new Y_i) in Figure 3.

33 old X_i	X_i-M_o	$(X_i$-$M_o)^2$	33 new Y_i	Y_i-M_n	$(Y_i$-$M_n)^2$
3.1	-0.71	0.50	3.1	0.70	0.49
7	3.19	10.20	2.7	0.30	0.09
7.6	3.79	14.39	2	-0.40	0.16
2.6	-1.21	1.45	3.9	1.50	2.26
3.1	-0.71	0.50	0	-2.40	5.75
3.6	-0.21	0.04	1.1	-1.30	1.68
2.8	-1.01	1.01	2.6	0.20	0.04
2.5	-1.31	1.71	1.5	-0.90	0.80
3.7	-0.11	0.01	0.9	-1.50	2.24
1.6	-2.21	4.87	0.8	-1.60	2.55
3.9	0.09	0.01	2.3	-0.10	0.01
1.3	-2.51	6.28	0.9	-1.50	2.24
3.4	-0.41	0.16	3.7	1.30	1.70
2.3	-1.51	2.27	1.6	-0.80	0.64
3.5	-0.31	0.09	4.3	1.90	3.62
1.9	-1.91	3.63	4.9	2.50	6.27
2	-1.81	3.26	1.7	-0.70	0.49
1.7	-2.11	4.44	4.5	2.10	4.42
2.5	-1.31	1.71	0.7	-1.70	2.88
3.6	-0.21	0.04	4.2	1.80	3.25
4.9	1.09	1.20	0.5	-1.90	3.60
4.4	0.59	0.35	0	-2.40	5.75
5	1.19	1.43	1.7	-0.70	0.49
5.2	1.39	1.94	1.4	-1.00	0.99
3.2	-0.61	0.37	2.2	-0.20	0.04
6.4	2.59	6.73	7.5	5.10	26.04
2.7	-1.11	1.22	0.8	-1.60	2.55
3.7	-0.11	0.01	0	-2.40	5.75
4	0.19	0.04	2	-0.40	0.16
3.7	-0.11	0.01	3	0.60	0.36
5.6	1.79	3.22	5.1	2.70	7.31
5.8	1.99	3.98	3.3	0.90	0.82
7.3	3.49	12.21	4.2	1.80	3.25
	M_o=3.81	σ_o=89.28		M_n=2.4	σ_n=98.67

3.2. The Nucleolin and the Nucleophosmin 1 (NPM1)

The nucleolin is a protein associated with intranucleolar chromatin and related to pre-rRNA molecules. It induces the chromatin decondensation by binding to another protein, the histone H1. It plays also a crucial role in pre-rRNA components transcription and in ribosome assembly from the ribosomal RNAs et proteins. We will study in the following AL-proximity between nucleolin and AL in 22 species representing different branches of the life tree: animals, plants, yeast and bacteria (Table 6). All the species have a high AL-proximity (more than 4.6, which is significant against the chance with $p < 10^{-6}$ in the Gaussian approximation of the binomial distribution). For example, the Gallus gallus nucleolin AL-proximity equals 13.2σ, which corresponds to $p < 10^{-29}$ (see Table 4).

Table 6. List of nucleolin AL-proximity for 22 species (from [15]).

22 Species with Nucleolin mRNA	AL-proximity
Solanum lycopersicum nucleolin (NCL) LOC101260453, mRNA NCBI Sequence: XM_010326160.3	18.1
Hydrotalea sandarakina nucleolin (NCL) str. DSM 23241 LX80DRAFT, GenBank: QKZV01000007.1	16.8
Gallus gallus nucleolin (NCL), transcript variant X1, mRNA NCBI Sequence: XM_046898333.1	13.2
Bauhinia variegata nucleolin (NCL) isolate BV-YZ2020 chromosome 1, GenBank: JAKRYI020000001.1	13
Homo sapiens nucleolin (NCL), mRNA NCBI Sequence: NM_005381.3	11.8
Lactobacillus lindneri nucleolin (NCL) DSM 20690 = JCM 11027, GenBank: FUXS01000002.1	11.7
Hydrobacter penzbergensis nucleolin (NCL) strain DSM 25353, GenBank: FNNO01000007.1	11.5
Ornithorhynchus anatinus nucleolin (NCL), mRNA NCBI Sequence: XM_029064618.2	11
Acanthochromis polyacanthus nucleolin (NCL) LOC110950732, mRNA NCBI Sequence: XM_022193488.2	10.8
Bactrocera dorsalis nucleolin (NCL) Fly_Bdor chr. 5 ASM2337382v1, mRNA NCBI Sequence: NC_064307.1	10.7
Monodelphis domestica nucleolin (NCL), variant X2, mRNA NCBI Sequence: XM_056819948.1	9.4
Xyrauchen texanus nucleolin (NCL) LOC127639553, mRNA NCBI Sequence: XM_052121621.1	9.3
Cannabis sativa nucleolin (NCL) chromosome 1, cs10, mRNA NCBI Sequence: NC_044371.1	9.3
Xenopus laevis nucleolin L homeolog (NCLL), mRNA NCBI Sequence: NM_001372137.1	8.8
Arabidopsis thaliana nucleolin like 2 (NUC-L2), mRNA NCBI Sequence: NM0_01338347.1	7.5
Dicentrarchus labrax nucleolin (NCL) LOC127349869, mRNA NCBI Sequence: XM_051375928.1	6.9
Helianthus annuus nucleolin (NCL) cultivar HA300 chromosome 17, GenBank: JANJOV010001181.1	6.9
Exophiala nucleolin (NCL) sp. JF 03-4F unplaced genomic scaffold EDD36 scaffold_3, GenBank: MU404352.1	6.8
Raphanus sativus nucleolin-like (NCL) cultivar WK10039 chromosome 4, GenBank: JRUI03000004.1	6.5
Cyprinus carpio nucleolin-like (NCL) LOC109082092, mRNA NCBI Sequence: XM_042712222.1	6.25
Carex littledalei nucleolin (NCL) isolate C.B.Clarke chromosome 3, GenBank: SWLB01000003.1	5.6
Saccharomyces cerevisiae nucleolin (NCL) S288C chr. VII, mRNA NCBI Sequence : NC_001139.9	4.6

Among the species having proteins the closest to AL (i.e., having the highest AL-proximity), procaryotes are found that present nucleolin-like proteins in their nucleoid. The primary role of the nucleolin is rRNA synthesis and ribosome biogenesis. One might speculate that bacterial nucleolin-like proteins induce rDNA transcription to exploit the host ribosome activity for their own survival. It could be the case in Table 5 for three bacteria: Hydrotalea sandarakina, Hydrobacter penzbergensis (family Chitinophagaceae from the phylum Bacteroidota) and Lactobacillus lindneri (family Lactobacillaceae from the phylum Bacillota).

Nucleolin inhibits the apoptosis and is overexpressed in numerous cancers [19–21] as well as its

associated proteins like nucleophosmin 1 (NPM1) (Table 7), and others like HSPD1, PAICS, CCT5, SERBP1 and GART (Table 8). Among the mRNAs of these proteins having the highest AL-proximity, there is the nucleophosmin 1 mRNA.

Table 7. List of nucleophosmin 1 AL-proximity for 33 species (from [15]).

33 Species with NPM1 mRNA	AL-proximity
Monodelphis domestica nucleophosmin 1 (NPM1), chr. 1, MonDom5, mRNA NCBI Sequence: NC_008801.1	16.4
Homo sapiens nucleophosmin 1 (NPM1), transcript variant 7, mRNA NCBI Sequence: NM_001355006.1	16.1
Rattus norvegicus nucleophosmin 1 (NPM1), mRNA NCBI Sequence: NM_012992.4	15.5
Mytilus coruscus strain nucleophosmin 1 (NPM1), contig: Mco4455, GenBank: CACVKT020004326.1	15.4
Bos taurus nucleophosmin 1 (NPM1), NPM1-GG allele, exon 1 and partial cds, GenBank: GQ144334.1	14.7
Mus musculus nucleophosmin 1 (NPM1), cDNA clone RZPDo836E0452D for gene Npm1, GenBank: CT010327.1	14
Xenopus tropicalis nucleophosmin (numatrin), mRNA NCBI Sequence : NM_20355.1	13.1
Lipotes vexillifer nucleophosmin 1 (NPM1) LOC103076865, misc_RNA NCBI Sequence: XR_456924.1	12.9
Bauhinia variegata nucleophosmin 1 (NPM1), isolate BV-YZ2020 chromosome 9, GenBank: JAKRYI020000009.1	12.2
Ictidomys tridecemlineatus nucleophosmin 1 (NPM1), isolate GS200 Itri18, GenBank: JAESOR010000030.1	12.1
Phodopus roborovskii nucleophosmin 1 (NPM1), contig: tig00001838, GenBank: CALSGD010001391.1	11.5
Eptesicus fuscus nucleophosmin 1 (NPM1), variant X2, mRNA NCBI Sequence: XM_054718139.1	11.4
Pan troglodytes nucleophosmin 1 (NPM1), variant X5, mRNA NCBI Sequence: XM_054684087.1	11.2
Pongo abelii nucleophosmin 1 (NPM1), variant X4, mRNA NCBI Sequence: XM_024246637.2	11.2
Agelaius phoeniceus nucleophosmin 1 (NPM1), variant X3, mRNA NCBI Sequence: XM_054642943.1	11.2
Capra hircus isolate 0256 nucleophosmin 1 (NPM1), partial cds, GenBank: HM006820.1	10.7
Mirounga angustirostris nucleophosmin 1 (NPM1), variant X2, mRNA NCBI Sequence: XM_045901451.2	10.3
Pteronotus parnellii mesoamericanus nucleophosmin 1 (NPM1), variant X2, mRNA NCBI Sequence : XM_054568056.1	10.3
Cervus elaphus hippelaphus nucleophosmin 1 (NPM1), isolate Hungarian chr. 25, GenBank: MKHE01000025.1	9.7
Talpa occidentalis nucleophosmin 1 (NPM1), mRNA NCBI Sequence : XM_037498589.2	9.3
Eublepharis macularius nucleophosmin 1 (NPM1), mRNA NCBI Sequence: XM_054985708.1	9.2
Falco cherrug nucleophosmin 1 (NPM1), partial mRNA NCBI Sequence: XM_014282768.2	9.1
Grus americana nucleophosmin 1 (NPM1), mRNA NCBI Sequence: XM_054841946.1	9
Taeniopygia guttata clone 0069P0004F05 nucleophosmin-like (NPM1), mRNA GenBank: EF191668.1	7.5
Caligus rogercresseyi tsa-crog-ngs-11089614 nucleophosmin 1 (NPM1), sequence GenBank: GAZX01006564.1	6.5
Rana catesbeiana nucleoplasmin 1 (NPM1) mRNA, complete cds, GenBank: DQ340656.1	6.2
Lycopersicon esculentum nuclear matrix protein 1 (NMP1) mRNA, GenBank: AF289255.1	5.9
Acyrthosiphon pisum nucleoplasmin 1 (Nlp), mRNA NCBI Sequence: NM_001161947.2	5.8
Labeo rohita nucleophosmin 1 (NPM1), strain BAU-BD-2019 chromosome 14, GenBank: CM040949.1	4.6
Monochamus saltuarius sensory neuron membrane protein 1 (NMP1) mRNA, GenBank : MT008451.1	4.5
Triticum aestivum nuclear matrix protein 1 (NMP1), mRNA partial cds, GenBank: AH011609.2	4.2
Hordeum vulgare nuclear matrix protein 1 (NMP1), mRNA partial cds, GenBank: AF289261.1	4.2
Salvelinus alpinus nucleoplasmin 1 (NPM1), LOC111963872, mRNA NCBI Sequence : XM_023987434.1	4

Table 8. AL-proximity for mRNAs of proteins associated to nucleolin in Homo sapiens and Monodelphis domestica.

Proteins	Homo sapiens AL-proximity	Monodelphis domestica AL-proximity
NCL	11.8	9.4
HSPD1	8.3	4.4
PAICS	6.2	4.0
CCT5	6.0	2.8
NPM1	20.1	16.4
SERBP1	4.0	3.4
GART	2.1	1.4

NPM1 is associated with nucleolar ribonucleoproteins like nucleolin (NCL) and is capable to bind both single and double-stranded nucleic acids. As for the nucleolin, it is involved in ribosomal biogenesis and helps small proteins transport to nucleolus. Among its multiple functions, we can also notice the functions of histone chaperone, ribosome biogenesis and transport, genomic stability and DNA repair, endoribonuclease activity and centrosome duplication during cell cycle. Its proximity to AL could be explained as for the nucleolin by their seniority, ubiquity and functionality.

3.3. The MutL-related proteins

The phylogenetic tree of Figure 5 has been obtained by comparing different MutL homologs, which are all mainly involved in DNA mismatch repair in different species [19]. The dendrogram has been generated using Megalign® from DNASTAR platform and is built from the similarities between mRNA nucleotide sequences and the proximity to AL partly reflects the distance to the root of the tree.

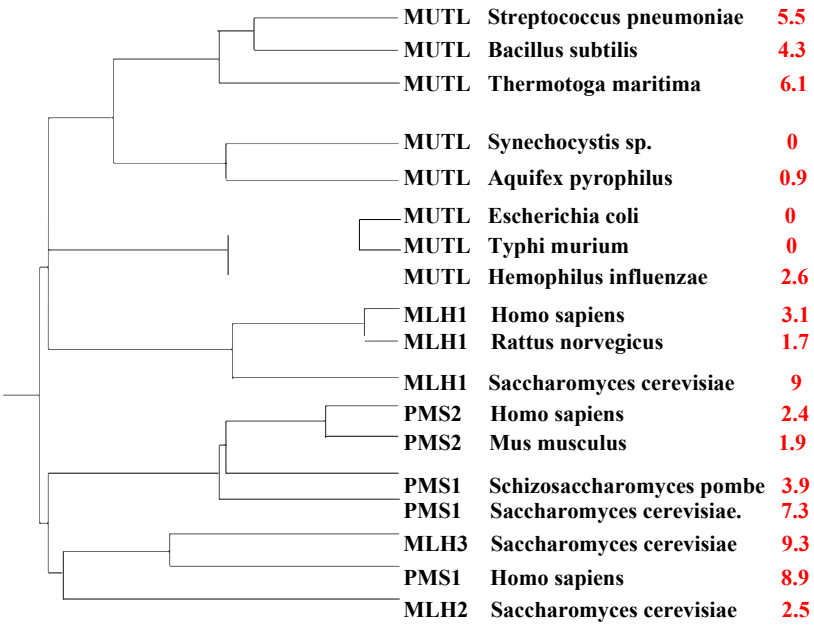

Figure 5. AL-proximity (right in red) of MutL-related proteins (after [19]).

3.4. Helicases for rRNAs building

The ribosomal RNAs (rRNAs) are necessary for the construction of the current ribosome composed of a large (resp. small) subunit LSU (resp. SSU) whose specific components called LSU RNAs (resp. SSU RNAs) are the largest (the smallest) of the major ribosomal RNA [22]. The associated RNA helicases contain a significant proportion of AL-pentamers and mean AL-proximity equals 9.2 for SSU and 8 for LSU helicases corresponding to their anteriority in evolution (Table 9).

Table 9. List of the RNA helicases needed for building components called LSU (resp. SSU) for the large (resp. small) ribosomal subunit, with their AL-proximity.

Protein	Ribosomal subunit	AL-proximity
Dbp8p	SSU	3.2
Dbp4p	SSU	10.4
Dhr1p	SSU	15.5
Dhr2p	SSU	7.4
Rrp3p	SSU	13.7
Rok1p	SSU	6.5
Fal1p	SSU	7.7
Prp43p	SSU & LSU	8.8
Has1p	SSU & LSU	8
Dbp3p	LSU	5.7
Dbp6p	LSU	7.4
Dbp7p	LSU	4.6
Dbp9p	LSU	7.2
Mak5p	LSU	10.7
Drs1p	LSU	10.4
Dbp2p	LSU	6.7
Spb4p	LSU	5.5
Dbp10p	LSU	14.4
Mtr4p	LSU	8.4
ECM16	LSU	7.7

3.4. Energy-releasing enzymes in the ribosome assembly process

Numerous enzymes (often small RNA-associated proteins) change the free energy of the pre-ribosomal precursors to reorganize the rRNA-protein complexes [22]. For example, Utp14, Bms1, Kre33 and Nog1 are respectively involved in nucleolar processing of pre-18S and 40S ribosomal subunits, as well as in nucleolar cleavages of precursor rRNAs during 18S and 25S rRNA synthesis and in biogenesis of the 60S ribosomal subunit. More generally, the energy-releasing enzymes participating in the construction of the ribosome are described in a cascade of syntheses and cleavages of rRNA molecules leading to the whole structure of the actual ribosome. On Figure 6, the numbers in blue represent the AL-proximity of the mRNAs of these energy-consuming enzymes (in red).

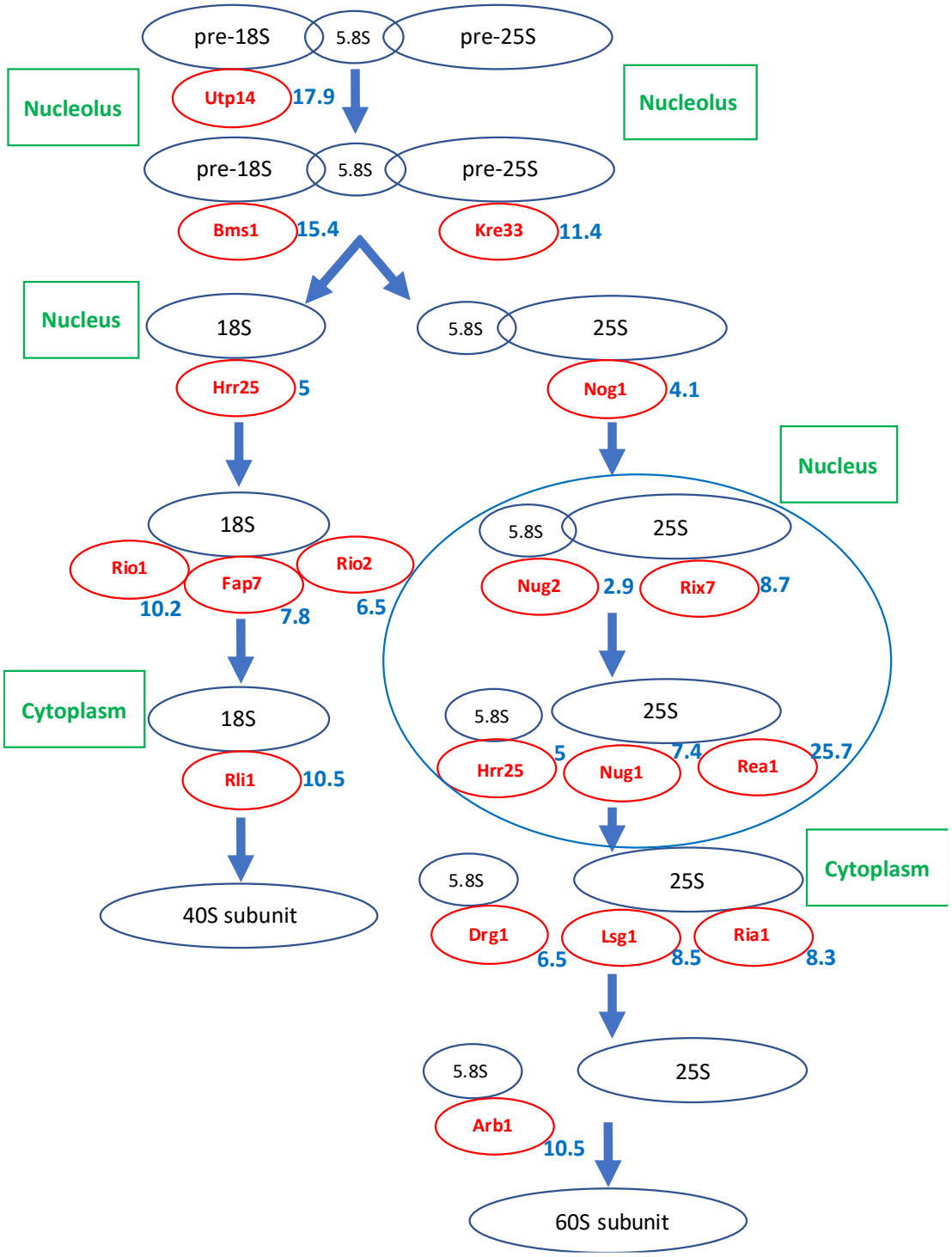

Figure 6. Steps (blue arrows) of the progressive maturation of ribosomal rRNAs with successive cleavages of the initial RNA involving energy-consuming enzymes (in red) for nucleolar, nucleic and cytoplasmic steps. The steps are labeled with the name of the nuclease enzyme involved in the cleavage. For simplicity, only main steps of the major 60S processing pathway are shown. Steps surrounded by a big blue ellipse will be detailed in Figure 7. The numbers in blue represent the AL-proximity of the mRNA of the energy-consuming enzymes (in red).

We see on Figure 6 that the first step of the progressive maturation of ribosomal rRNAs concerning the first cleavages of the initial RNA involves in nucleolar compartment some energy-consuming enzymes being the closest to AL, which is in favor of their anteriority in the process of evolution where they progressively appeared. We focus now on the nuclear steps surrounded by a blue circle in Figure 6, we see on Figure 7 the steps of the reorganization of the initial pre-60S rRNA during ribosomal maturation until the emergence of the RNA components 25S, 7S and 5S of the ribosome [22].

Following [23], the mechanism of cleavage starts with the binding of the ribosome biosynthesis protein Nsa1 to the pre-60S subunit containing the 27SA rRNA and other early 60S assembly factors (Rrp5, Noc1, and Nop4) are also bound to the 27SA. Then, the ATPase Rix7 interacts with Nsa1, to remove it from the 27SA causing the loss of Rrp5, Noc1, and Nop4, which allows the binding to 27SB of the later 60S assembly factors Rsa4, Nop53, Spb1, Sda1, Arx1, and the energy-consuming enzymes Nug1 and Nog2. The energy-consuming enzymes of the Figure 7 are represented in orange, other assembly factors in yellow, ribosome biosynthesis protein Nsa1 in blue and ATPase Rix7 in red. As in Figure 6, the early factors are those containing the most pentamers from AL, hence having the highest AL-proximity (in blue).

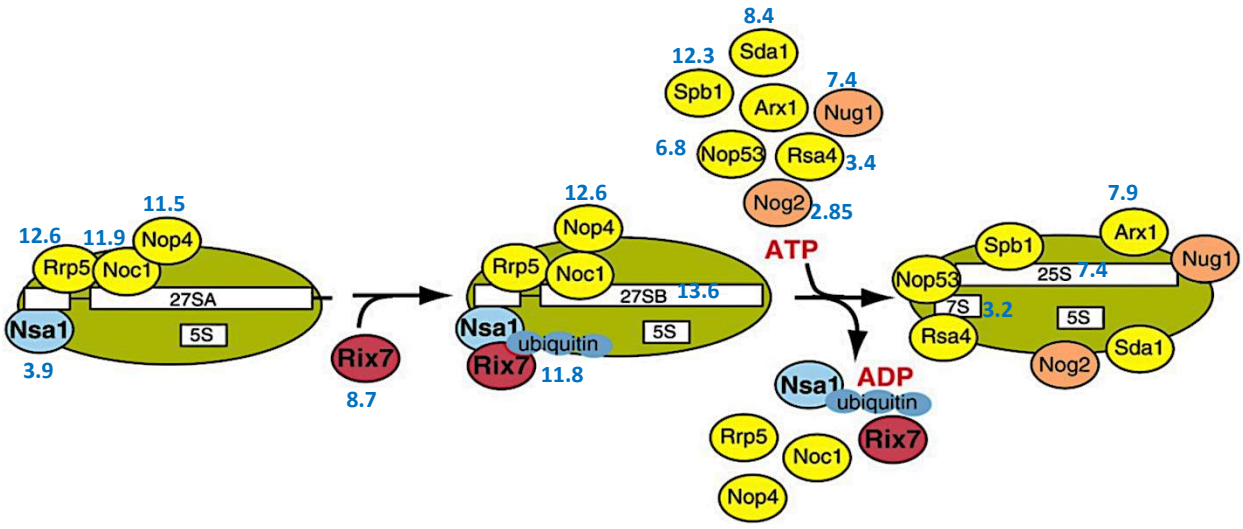

Figure 7. Nuclear steps of the remodeling of pre-60S subunits. The numbers in blue represent the AL-proximity of the mRNAs of the energy-consuming enzymes (in orange), assembly factors (in yellow), ribosome biosynthesis protein Nsa1 (in blue) and ATPase Rix7 (in red).

3.6. Ancestral character of the Archaea

The content of the genomes in AL top pentamers is much higher for Archaea genome than that for the genomes of more recent species. For example, the mean AL-proximity is equal to 93 for the complete genomes of Methanomada Archaea [15,24] (Figure 8), but AL-proximity of the whole chromosome 8 of a plant having both nucleolin and nucleophosmin mRNAs close to AL (Tables 6 and 7), Bauhinia variegata, equals only 38, which reinforces the idea that AL-proximity can serve as quantifier of the anteriority of species.

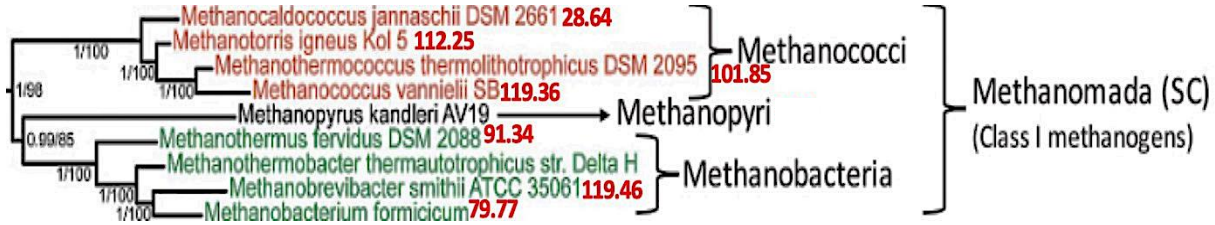

Figure 8. Phylogenetic tree of Methanomada Archea with indication (in red) of the AL-proximity of their complete genome (after [24]).

Some Archaea are also close to AL through the components of their ribosome, such as the 5S RNA (a component of the large ribosomal subunit) or through their tRNAs. It is the case of Methanococcus voltae A3, whose 5S RNA and tRNA-Gly^{GCC} primary and secondary structures are shown in Figure 9.

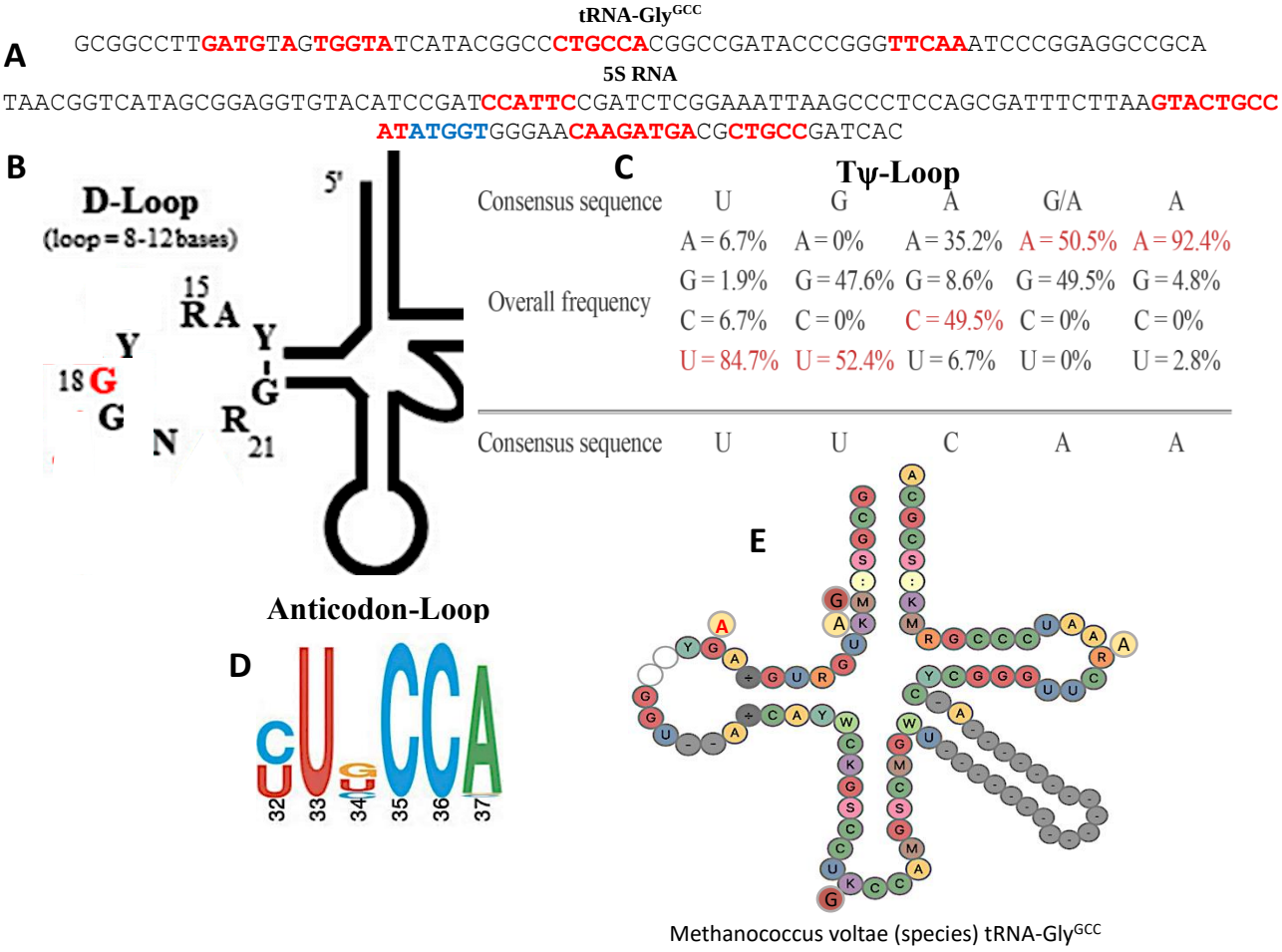

Figure 9. A) Primary structure of tRNA-Gly^{GCC} [9] and 5S RNA [25] of Methanococcus voltae; B) tRNA D-loop (after [26)]; C) Tψ-loop (after [27]); D) Anti-codon loop (after [28]); E) Secondary structure of tRNA-Gly^{GCC} of Methanococcus voltae [9].

4. Discussion

All the processes previously studied relate to the construction of the actual ribosome. If we hypothesize that the AL ring represented a primitive form of proto-ribosome, it is legitimate to consider the proximity to AL of all the actors (RNAs or proteins) involved in the maturation of the current ribosome. These actors appear either in graphs describing the stages of maturation, in phylogenetic trees comparing different species, or in comparison tables of the same actor (such as nucleolin or nucleophosmin 1) in different species.

By looking at the primary structure of tRNA-Gly^{GCC} [9] and 5S RNA [25] of Methanococcus voltae on Figure 9, the observed numbers of the 9 pentamers from the top of the hairpin form of AL are respectively 4 and 14. Because this 5S RNA contains 109 possible pentamers, the expected number of 9-pentamers from AL in 5S RNA is equal to 109x9/10024 ~ 1 pentamer, then the AL-proximity of this 5S RNA equals 13. This value corresponds to a probability of the order of 10^{-28} (cf. Table 4) to observe 14 times any of the 9 pentamers from the top of the hairpin form of AL in the 5S RNA. For the tRNA-Gly^{GCC}, the expected number of the 9 pentamers from AL is 67x9/1024 = 0.6, then the AL-proximity of the tRNA-Gly^{GCC} equals only 3.2, but if we restrict the occurrence of these 9 pentamers to the sequence made of the successive tRNA-loops, the observed number is 12, then the AL proximity becomes equal to 14.5 and the probability to observe 13 times the 9 pentamers from the top of the hairpin form of AL in the tRNA-Gly^{GCC} loops sequence is of the order of 10^{-28} (cf. Table 4). More generally, if we observe, in an RNA of length n+4, N pentamers among the 9 pentamers from the top of the hairpin form of AL, the observed frequency of these pentamers is equal to f=N/n and their expected number is equal to np, with p=9/1024, then with an AL-proximity equal to $(f-p)n/[np(1-p)]^{1/2}$, that is about $(f-p)(n/p)^{1/2}$, because p is small, the AL-proximity depends linearly on f=N/n and this dependency comes from both the observed number of pentamers and RNA length.

The probabilities above concerning the ancient character of Methanococcus voltae are going in the same direction as a number of evolutionary genomic studies supporting the ancestrally of superkingdom Archaea [29–31]. These works embody the last universal common ancestor of cellular life and are compatible with the existence of a primitive RNA as AL prior to this common ancestor. These studies propose also that ancestors of Euryarchaeota co-evolved with those of Bacteria prior to the diversification of Eukarya, which could be reinforced by the fact that the protein G1PDH (presented as a crucial marker of the evolution in [29]) has a decreasing value of its AL-proximity from Euryarchaeota and Bacteria to Eukaria, e.g., 4.25 for the Euryarchaea Methanothermobacter wolfeii G1PDH, 3 for the Bacterium Bacillus subtilis G1PDH and 2.8 for the Eukaryote fish Rachycentron canadum G3PDH (cf. Supplementary Material Table S2).

5. Conclusion

To conclude, we have presented in this article a certain number of molecular structures involved in the construction of the current ribosome and we found that their nucleotide sequences contain, much more frequently than expected by simple chance, pentamers resulting from the concatenation of an optimal combinatorial consensus sequence of the loops of many tRNA-Gly^{GCC}, namely AATGGTA for the D-loop [26], TTCAA for the Tψ-loop [27] and CTGCCA for the anticodon loop [28]. The RNA loop so obtained (called AL for Ancestral Loop) constitutes a ring structure which could: i) help the protein synthesis [32,33], ii) be capable of self-replication [34,35] and iii) have played the role of

peptide catalyzer at the origin of life on Earth. We will continue to explore systematically in the future this hypothesis, by looking for the trace of AL in even more species of Archaea, Bacteria and Eucaryotes, in order to strengthen the hypothesis of the emergence of an RNA world, defined by RNA molecules with catalytic and replicative properties [36].

Use of AI tools declaration

Authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

We would like to thank Michel Thellier for many helpful discussions.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. B. C. Alberts, The function of the hereditary materials: Biological catalyses reflect the cell's evolutionary history, *Amer. Zool*., **26** (1986), 781–796.
2. P. G. Higgs, N. Lehman, The RNA world: Molecular cooperation at the origins of life, *Nat. Rev. Genet.*, **16** (2015), 7–17. https://doi.org/10.1038/nrg3841
3. J. Demongeot, Sur la possibilité de considérer le code génétique comme un code à enchaînement, *Revue de Biomaths*, **62** (1978), 61–66.
4. J. Demongeot, J. Besson, Code génétique et codes à enchaînement, *C. R. Acad. Sc. Série III*, **296** (1983), 807–810.
5. J. Demongeot, A. Moreira, A circular RNA at the origin of life, *J. Theor. Biol*., **249** (2007), 314–324. https://doi.org/10.1016/j.jtbi.2007.07.010
6. J. Demongeot, V. Norris, Emergence of a "Cyclosome" in a primitive network capable of building "infinite" proteins, *Life*, **9** (2019), 51. https://doi.org/10.3390/life9020051
7. J. Demongeot, H. Seligmann, Evolution of tRNA sub-element accretion from small and large ribosomal RNAs, *Biosystems*, **193** (2022), 104796. https://doi.org/10.1016/j.biosystems.2022.104796
8. J. Demongeot, M. Thellier, Primitive oligomeric RNAs at the origins of life on Earth, *Int. J. Molecular Sci*., **24** (2023), 2274. https://doi.org/10.3390/ijms24032274
9. GtRNAdb. Available from: http://lowelab.ucsc.edu/GtRNAdb/
10. tRNA. Available from: http://trna.bioinf.uni-leipzig.de/DataOutput/Result
11. trNAviz. Available from: http://trna.ucsc.edu/tRNAviz/summary/.
12. Genscript. Available from: https://www.genscript.com/tools/codon-frequency-table/
13. Kinefold. Available from: http://kinefold.curie.fr/cgi-bin/neorequest.pl?batch=0&sim=2&base=AL-73086
14. Vectorbuilder. Available from: https://en.vectorbuilder.com/tool/dna-secondary-structure/
15. NCBI. Available from: https://www.ncbi.nlm.nih.gov/nucleotide?cmd=search
16. M. Edous, O. Eidous, A simple approximation for normal distribution function, *Math. Statist.*, **6** (2018), 47–49. https://doi.org/10.13189/ms.2018.060401

17. A. Harish, G. Caetano-Anollés, Ribosomal history reveals origins of modern protein synthesis, *PLoS One*, **7** (2012), e32776. https://doi.org/10.1371/journal.pone
18. Socscistatistics. Available from: http://www.socscistatistics.com/tests/studentttest/default2.aspx.
19. H. Flores-Rozas, R. D. Kolodner, The Saccharomyces cerevisiae MLH3 gene functions in MSH3-dependent suppression of frameshift mutations, *Proc. Natl. Acad. Sci USA*, **95** (1998), 12404–12409. https://doi.org/10.1073/pnas.95.21.12404
20. S. Yangngam, J. Prasopsiri, P. Hatthakarnkul, S. Thongchot, P. Thuwajit, P. Yenchitsomanus, et al., Cellular localization of nucleolin determines the prognosis in cancers: a meta-analysis, *J. Mol. Med*., **100** (2022), 1145–1157. https://doi.org/10.1007/s00109-022-02228-w
21. V. Firlej, P. Soyeux, M. Nourieh, E. Huet, F. Semprez, E. Huet, et al., Overexpression of nucleolin and associated genes in prostate cancer, *Int. J. Molecular Sci*., **23** (2022), 4491. https://doi.org/10.3390/ijms23094491
22. S. Melnikov, H. S. Kwok, K. Manakongtreecheep, A. van den Elzen, C. C. Thoreen, D. Söll, Archaeal ribosomal proteins possess nuclear localization signal-type motifs: Implications for the origin of the cell nucleus, *Mol. Biol. Evol*., **37** (2020), 124–133. https://doi.org/10.1093/molbev/msz207
23. B. S. Strunk, K. Karbstein, Powering through ribosome assembly, *RNA*, **15** (2009), 2083–2104. https://doi.org/10.1261/rna.1792109
24. P. Adam, G. Borrel, C. Brochier-Armanet, S. Gribaldo, The growing tree of Archaea: New perspectives on their diversity, evolution and ecology, *ISME J.*, **11** (2017), 2407–2425. https://doi.org/10.1038/ismej.2017.122
25. G. Wich, L. Sibold, A. Böck, Divergent Evolution of 5S rRNA Genes in Methanococcus, *Z. Naturforsch.*, **42** (1987), 373–380.
26. M. Roovers, L. Droogmans, H. Grosjean, Post-transcriptional modifications of conserved nucleotides in the T-Loop of tRNA: A tale of functional convergent evolution, *Genes*, **12** (2021), 140. https://doi.org/10.3390/genes12020140
27. C. W. Chan, B. Chetnani, A. Mondragón, Structure and function of the T-loop structural motif in noncoding RNAs, *RNA*, **4** (2013), 507–522. https://doi.org/10.1002/wrna.1175
28. H. S. Bernhardt, W. P. Tate, Evidence from glycine transfer RNA of a frozen accident at the dawn of the genetic code, *Biol. Direct*, **3** (2008), 53. https://doi.org/10.1186/1745-6150-3-53
29. J. T. Staley, G. Caetano-Anollés, Archaea-first and the co-evolutionary diversification of domains of life, *Bioessays*, **40** (2018), e1800036. https://doi.org/10.1002/bies.201800036
30. J. T. Staley, Domain Cell Theory supports the independent evolution of the Eukarya, Bacteria and Archaea and the Nuclear Compartment Commonality hypothesis, *Open Biol.*, **7** (2017), 170041. https://doi.org/10.1098/rsob.170041
31. G. Caetano-Anollés, A. Nasir, K. Zhou, D. Caetano-Anollés, J. E. Mittenthal, F. J. Sun, K. M. Kim, Archaea: The first domain of diversified life, *Archaea*, **2014** (2014), 590214. https://doi.org/10.1155/2014/590214
32. K. Tamura, P. R. Schimmel, Peptide synthesis with a template-like RNA guide and aminoacyl phosphate adaptors. *Proc. Natl. Acad. Sci. USA*, **100** (2003), 8666–8669. https://doi.org/10.1073/pnas.1432909100
33. K. Tamura, P. R. Schimmel, Chiral-selective aminoacylation of an RNA minihelix: Mechanistic features and chiral suppression, *Proc. Natl. Acad. Sci. USA*, **103** (2006), 13750–13752. https://doi.org/10.1073/pnas.0606070103

34. L. Zhou, D. Ding, J.W. Szostak, The virtual circular genome model for primordial RNA replication, *RNA*, **27** (2021), 1–11. https://doi.org/10.1261/rna.077693.120
35. K. Adamala, J. W. Szostak, Non-enzymatic template-directed RNA synthesis inside model protocells, *Science*, **342** (2013), 1098–1100. https://doi.org/10.1126/science.1241888
36. L. Demetrius, Directionality theory and the origin of life, *arXiv*, (2023), 2304.14873. https://doi.org/10.48550/arXiv.2304.14873

AIMS Press

©2024 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)