Entropy, a tool for system biology applied to the genetic regulatory networks

Jacques Demongeot and Hana Hazgui

RNA molecules are often involved in the regulation of complex genetic networks as effectors, activators (small RNAs acting as transcription factors), inhibitors (microRNAs) or hybrids (circular RNAs). We give examples of such genetic networks and show that: i) RNA-RNA or RNA-peptide interactions could be explained by the presence of RNA «relics» having played an important role during the evolution and surviving among many present genomes, in which the heterogeneity of their probability distribution can be quantified by the static entropy, and ii) the dynamical entropy related to the dynamics of these networks can be used to characterize their robustness.*

1. Introduction

RiboNucleic Acid (RNA) is an important living molecule with the propensity to regulate many fundamental mechanisms of the cell. RNA molecules are indeed often involved in the regulation of many complex biological networks as effectors. They can be activators, like the small RNAs involved in transcription factors*, inhibitors like microRNAs involved in translation, or hybrids like circular RNAs, which are inhibiting microRNAs. The present genomes* are the result of a long evolution from the start of the life on the earth until the appearance of mammals and human. We will show that RNA relics persist coming from the beginning of life. These relics are small RNA sequences having fulfilled an important role for transmitting the living information already at the first steps of the evolution and playing still now a notable function in the control of important genetic networks involved in numerous living systems. In this paper, we will give some examples of complex biological networks regulated by RNAs and show that:

i) present RNA-RNA or RNA-peptide interactions could be explained by the presence of RNA «relics» having played an important role during the evolution and having survived in many genomes, their present probability distribution being quantified by the static entropy,

ii) the dynamical entropy related to the dynamics of the biological networks can be used to characterize the robustness of the RNA regulated systems.

We will describe the RNA relics in Section 2, then show in Section 3 their survival in present genomes and describe in Section 4 the role they play presently in the regulation of important genetic networks. Eventually, in Section 5, we will define the dynamical entropy and its role in the quantification of the network robustness.

2. RNA relics

Let us consider now the two RNA sequences called respectively AL (for Archetypical Loops) and AB (for Ancestral Bases):

5′-UGAAUGGUACUGCCAUUCAAGA-3′(AL)
5′-UGAAUGGUGCCAUUCAAGACUA-3′(AB)

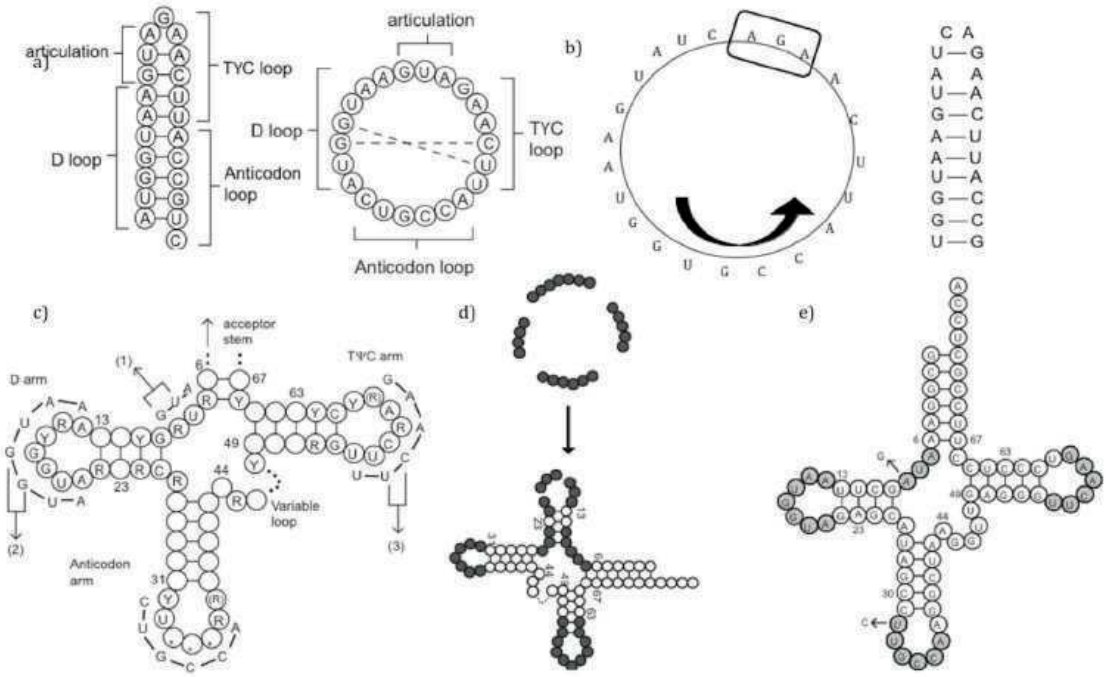


Fig. 1. (a) The relic sequence AL in hairpin and circular forms. (b) The relic sequence AL in hairpin and circular form. (c) AL matching the loops of the archetypal tRNA. (d) Correspondence between the circular and the clover-leave structure. The ring is cut into segments aligned with the conserved domains of the tRNA (loops), respecting the 5'-3' orientation. (e) AL matching the loops of the $tRNA^{Gly}$ of *Arabidopsis thaliana*. Only positions 9 and 32 differ in AL, as shown.

Their main features have been previously described,[1] hence we will describe only the most relevant for our argumentation about their survival until the present genomes.

The first small circular RNAs ("RNA rings") when life emerged could have been stabilized by a proto-membrane of amino acids* determined by the affinities to their different codons*, overlapping in a circular sequence.[2] A high diversity of codons in these rings, corresponding to different amino acids, would increase the chance that the rings attract enough amino acids to support their stability. In this way, the optimal RNA structure turns out to be a ring which codes for all the different available amino acids, i.e., having at least one codon in each of the synonymy classes of the genetic code.

At the same time, the ring should be as small as possible, since bigger rings will be more prone to accidental denaturation. This would lead to the determination of RNA rings of 20 nucleotides*, which contain codons in each of the coding classes.

However, two observations lead to study a different set of minimal rings. The first is combinatorial: exhaustive listing through backtracking algorithms shows that rings with the required property do not exist; hence, some latitude must be allowed. The second observation is based on the dynamics of the scenario: with structures being routinely denaturated, and new structures emerging from the pieces that float in the environment, there will be a selective advantage for those structures that are more likely to break in ways that facilitate a later reconstruction. Since the links inside peptides and RNA are likely to be stronger than the affinity-determined binding between them (based on weak forces like van der Waals* or electromagnetic, following the stereochemical* theory of the genetic code), a RNA-peptide ring structure is likely to be occasionally turned into a small strand of RNA and a peptide. The reconstruction of the same species of such a structure will be easier if it breaks preferentially in a certain point along the ring, i.e., if it contains an unprotected spot, one without an amino acid/codon complex. This, according to our assumptions, is expected to happen if the ring contains a stop codon*. It follows that we must

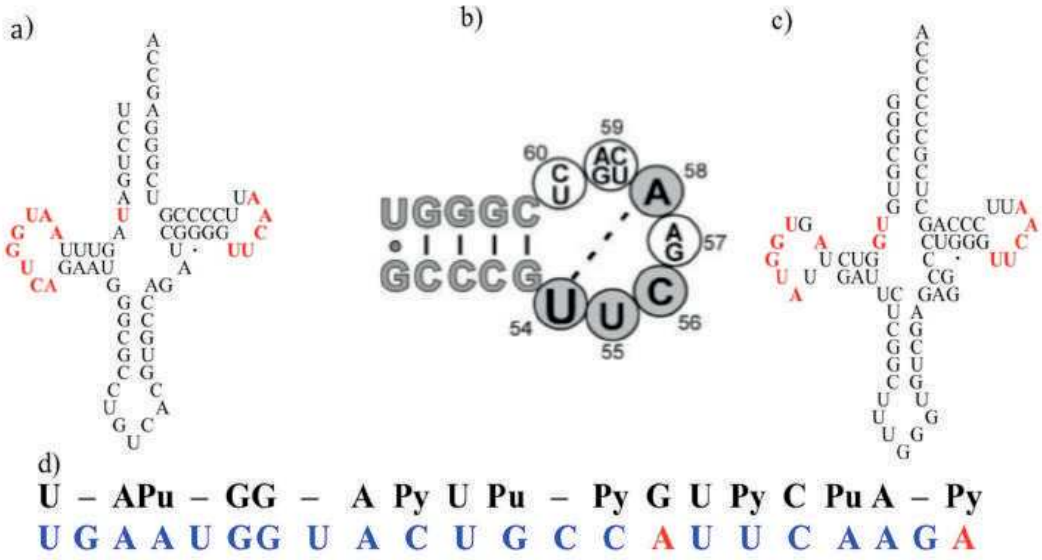


Fig. 2. (a) S. cerevisiae $tRNA^{Asp}$.[3] (b) Ty-loop consensus sequence derived from sequences of 11 species of T. thermophilus tRNAs contained in the tRNA Compilation Data base.[4] (c) S. cerevisiae $tRNA^{Pr}$.[5] (d) Consensus sequence proposed by B. Lewin for the tRNA loops.[6] The AL bases matching with Lewin's bases are in blue.

look at minimal rings able to code for the 21 synonymy classes of the genetic code (including the stop class). The combinatorics shows that the variational problem satisfying two opposite criteria, i.e., i) to have the minimal length and ii) to code for these 21 classes, has solutions only if the sequence begins with the start codon AUG. The set of corresponding solutions has a barycentre* for the circular Hamming distance*, called the AL ring. The AB ring is another solution whose hairpin form is more stable than the AL one (see Fig. 1).

The RNA ring AL matches well with the sequence obtained from the succession, in their order in the cloverleaf structure, of the tRNA loops (D-loop*, anti-codon loop and TΨ-loop*),[7] and that for many tRNAs (Fig. 2) like the consensus tRNA loops proposed by B. Lewin[8] and could be considered as part of an ancient proto-ribosomal system, in which AL have played two roles: i) it attracted amino acids with weak bonds like the present tRNAs, and ii) it favoured the creation of strong peptidic bonds between these amino acids like in the present ribosomes*.

3. Static entropy and diversity of the nucleotide motifs

3.1. Nucleotide sequences from RNA and DNA worlds

In order to compare AL with nucleotide sequences coming from genetic databases, we computed the number *W* of occurrences of AL subsequences of length equal or greater than 5 (called 5-ALsq) in RNA or DNA sequences. To assess the *W* significance, we ran 1000 computations, in which each nucleotide sequence was randomized by shuffling its bases, but keeping base content and length.

The mean number for these randomized cases was $\langle W_R \rangle = 91.665$, with a standard deviation σ_R of 286 (cf. Table 1). By computing W for all RNA sequences, we obtain a total of 120.441 significantly greater than $\langle W_R \rangle$, with $p=10^{-2200}$. The number W of matches between AL and RNA sequences was indeed well above $\langle W_R \rangle$ for any empirical distribution given on Table 1.[9]

The frequencies of the histogram given on Fig. 3 are consistent with the idea that AL fragments were the source material for the primitive RNAs, with higher probability for the fragments that were not disrupted in the AL hairpin form. The entropy of the distribution corresponding to the histogram given on Fig. 3 is equal to $H_1 = 2.93176$.

3.2. RNAs from human mitochondrial genome

Human mitochondria* contain small RNAs including microRNA, tRNA, etc. Analyzing these small RNAs from the cell lines HEK (730293 nucleotides[10]) and HeLa* (363056 nucleotides[11]). For HEK (resp. HeLa), the least frequent 7-ALsq AAGAUGA (the least frequent also on the histogram of Fig. 3) is observed 1760 (resp. 890) times, for an expected number of 713 ± 160 (6□) (resp. 355 ± 113 (6□)), showing a very significant occurrence of AL relics inside the human mitochondrial genome.

RNA type	S	W	$\langle W_R \rangle$	σ_R	$\sigma_R/\langle W_R \rangle$
tRNA conserved domains	2.77	3498	1262	32	0.025
Intron	1.46	65,119	44,599	202	0.005
snRNA (spliceosomal)	1.43	5421	3799	59	0.016
snRNA	1.39	5478	3941	59	0.015
rRNA	1.37	14,213	10,366	97	0.009
miRNA	1.33	716	540	22	0.041
tRNA	1.12	12,709	11,376	102	0.009
Ribozyme	1.11	2114	1902	41	0.022
sRNA	1.03	498	484	20	0.041
Cis-reg	0.97	9539	9809	93	0.009
Gene	0.96	4483	4678	65	0.014
Antisense	0.88	151	171	12	0.070
Whole Rfam	1.31	120,441	91,665	286	0.003

Table 1. Subsequences of length 5 common between AL and RNA families. The type snRNA corresponds to the aggregation of snRNA types in Rfam* database[12] (except for spliceosomal snRNAs listed apart) and contains almost only snoRNAs.

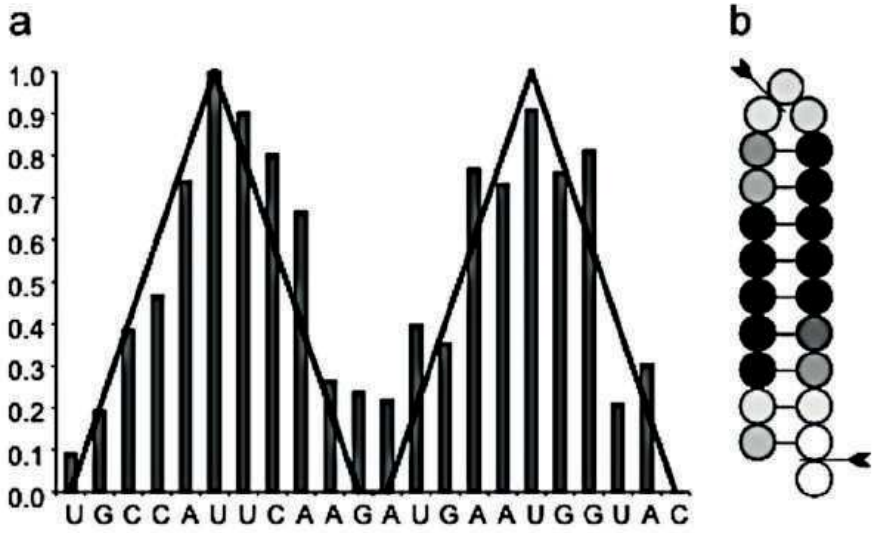


Fig. 3. Frequencies in AL (at their start nucleotide) of subsequences of length 5 (5-ALsq) matching with Rfam data base sequences.[13] For better comparison, frequencies have been normalized and dark colours in (b) indicates the position of the most frequent 5-ALsqs. The triangle lines in (a) correspond to the distances of each base with respect to the two inter-base positions marked by an arrow in the AL hairpin form (b). Values graphed in (a) are shaded in (b) (white and black representing the minimum and maximum values of this distance, respectively).

3.3. Circular RNA and DNA sequences

Circular RNAs correspond to possible inhibitors of the microRNAs ("microRNA sponges").[14] An example given on Fig. 4 shows the inhibitory power of the human circular RNA called CDR1as (resp. ciRs7) on microRNAs called miR-671 and miR-485, (resp. miR-671 and miR-485), and the inhibitory power of human microRNAs called miR-671 and miR-485 on their proteic targets, respectively Glucose-6-Phosphate Isomerase (involved in glycolysis) and FPN1 Ferroportin (involved in the iron regulation as shown in Section 4.3).

3′-ACGACCUUCGGGACCUCUACGACCUUCU-5′CDR1as

3′-GUUGGUGAGUUUUACUUGUUGU-5′ anti-AL 14 anti-matches vs hsa-miR-671-5p

5′-GGAAGCCCUGGAGGGGCUGGAGG-3′ hsa-miR-671-5p 21 anti-matches vs CDR1as

3′-CGAUCUGUACCUCCCCGACC-5′ GPI 16/20 anti-matches vs hsa-miR-671

3′-UGUUGGUGAGUUUUACUUGUUG-5′ anti-AL 14 anti-matches vs ciRs7

5′-AGAGAGGAUGGGGGAGUUGUGUAUUCUUCCAGGUUC-3′ciRs7

3′-CUGGAUCAGUGGAUCUA-5′ IRE-FPN1a 12/17 anti-matches vs ciRs7

5′-AUGGGGCAACAUAUUGUAUGAA-3 FPN1a 14 anti-matches vs hsa-miR-485

3′- UCUCUCCUCUCGGCACAUACUG-5′ hsa-miR-485 15 anti-matches vs ciRs 7

5′-AGAGAGGAUGGGGGAGUUGUGUAUUCUUCCAGGUUC-3′ciRs7

5′-CCUGUUGGUCUCUUCCAGGUAC-3′ IRP 14 matches with ciRs7

Fig. 4. Correspondence between AL, anti-AL, circular RNAs and their targets, the microRNAs.

The circular RNA data base[15] contains many circular RNAs with a significant frequency of occurrence of AL subsequences like in the following example where occurrences are indicated in red or blue (if overlap):

hsa_circ*_0071327 159076770 - 159076977, 208 bps*
AAAAAATCAAAGAGTGCCATCTTGGACCACTCATGATGATGTATTTCAGTACAACCCGATTCAGGCTTGGGTAC-
TCGGCCATTCTGCCAGCATTTCTGTTTCAGCAACTGCTGATAAGTTCCCCAGGTGAGCTTAACAGAAGAATGG-
GTGTCATTACTTGCTGAAGATAAAGATGCATCCCAAGAATGATGGGGCATGGGCGGCCAT

We observe on the sequence above 19 times 5-ALsqs from the sequence corresponding to Tψ– and D-loops: UUCAAGATGAATGGTAC, for 207x13/1024 = 2.6 ± 9.7 (6σ) expected ($p=10^{-24}$). Exploring in the same way the circular RNA data base,[16] we observe also many circular RNAs with a significant frequency of occurrence of AL subsequences (ALsq). From The Nucleotide database,[17] the distribution of 5-ALsqs (cf. Fig. 5) in Assphage circular DNA (97065 bp*) is equal to:

ttcaa 250 tcaag 154 caaga 146 aagat 163 agatg 163 gatga 122 atgaa 211 tgaat 238 gaatg 152 aatgg 145 atggt 156 tggta 120 ggtac 62 gtact 90 tactg 143 actgc 129 ctgcc 160 tgcca 155 gccat 121 ccatt 198 cattc 155 attca 206.

The total number of 5-ALsqs observed is equal to 3439, with an expected total equal to 22x97061/1024 = 2085 ± 73* (* = 95% confidence, $p=10^{-194}$). The entropy of the above distribution is equal to: $H_2 = 2.77723$.

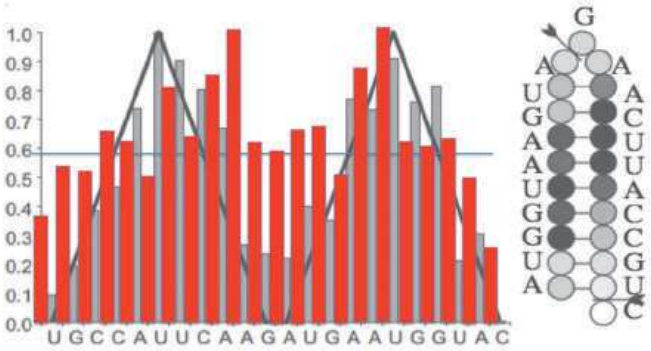


Fig. 5. Histogram of the 5-ALsqs observed in Assphage circular DNA.[18] Frequencies up to the blue bar correspond to the 0.95-significativity vs the uniformity.

By looking more precisely to the distribution of AL subsequences of length more than 6 (Fig. 6), we observe 17 9-ALsq, with 8.145 ± 4.57* expected ($p=10^{-3}$) and 69 8-ALsq, with 32.6 ± 9* expected ($p=10^{-11}$). The entropy of the distribution corresponding to the histogram given on Fig. 6 is equal to: H_3 =2.26074.The entropy of the distribution corresponding to the histogram of Fig. 6 is equal to: H3=2.26074.

9 & 10-ALsq		8-ALsq	
agatgaatg	1	agatgaat	2
gatgaatgg	0	gatgaatg	3
atgaatggt	1	atgaatgg	2
atgaatggta	1	tgaatggt	2
tgaatggta	1	gaatggta	3
gaatggtac	0	aatggtac	2
aatggtact	2	atggtact	4
atggtactg	0	tggtactg	1
tggtactgc	0	ggtactgc	0
ggtactgca	0	gtactgcc	0
tactgccat	1	tactgcca	2
tgccattca	1	actgccat	1
tgccattcaa	1	ctgccatt	1
gccattcaa	1	tgccattc	2
ccattcaat	0	gccattca	2
attcaagat	2	ccattcaa	4
ttcaagatg	0	cattcaag	1
attcaagatg	1	attcaaga	4
ttcaagatg	3	ttcaagat	10
tcaagatga	3	tcaagatg	9
aagatgaat	1	caagatga	4
aagatgaatg	1	aagatgaa	10

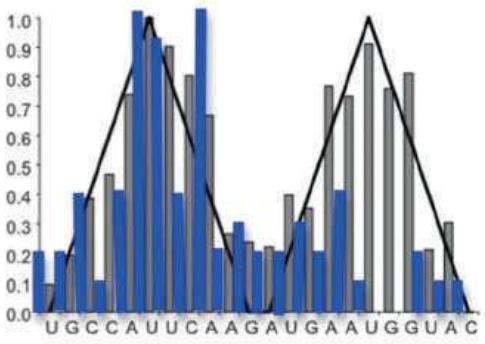
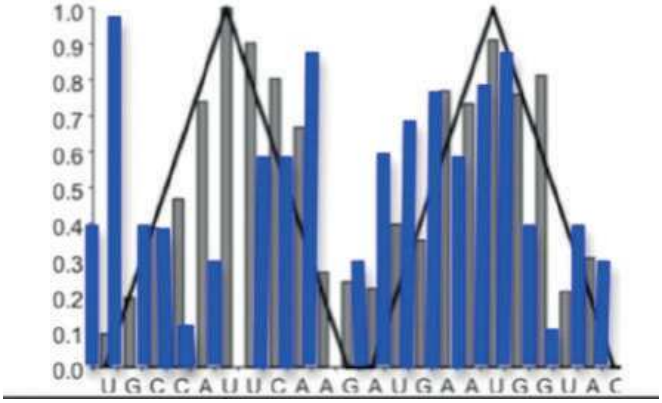


Fig. 6. Numbers (left) and Histogram (right) of the 8-AL sequences observed in Assphage circular DNA.[19]

3.4. DNA sequences from genomes of different species

By exploring the complete genomes of microbes,[20] which contain 15,358,075,464 bp for 4867 sequences, we observe the AL subsequence TTCAAGATGAATGGT (made of the Tψ- and D-loops indicated in red) 51 times for 14 ± 7.5* expected. We discover also quasi-perfect matchings:

Desaturase gene fragment, 5218 bp:[21] CAGCCCTCCAAGATGAATGGTA 19 AL-matches
Candida Ca20chr5, 1191531 bp:[22] TGGTACTGCCATTGAAGATAGA 19 AL-matches
Glyma14g34640.1, 3709 bp:[23] TGCTATTCAAGACTATGAAATG 19 AB-matches.



5-uple	Nb	frequency	-ΔG %	exp(-ΔG/RT)	exp(-ΔG/RT)/Z
		= Nb/106	(kJ/mol)		Z = 10794
aatgg	8	0.075	13.32	175	1.6%
atggt	9	0.085	14.92	325	3%
tggta	4	0.038	14.92	325	3%
ggtac	1	0.0095	14.27	253	2.3%
gtact	4	0.038	11.75	95	0.9%
tactg	3	0.028	11.75	95	0.9%
actgc	4	0.038	18.61	1361	12.6%
ctgcc	10	0.094	20.41	2735	25.3%
tgcca	4	0.038	20.58	2921	27%
gccat	4	0.038	18.45	1325	12.3%
ccatt	1	0.0095	12.32	119	1.1%
cattc	3	0.028	11.43	84	0.8%
attca	0	0	10.91	69	0.6%
ttcaa	6	0.057	11.53	87	0.8%
tcaag	6	0.057	12.93	150	1.4%
caaga	9	0.085	12.93	150	1.4%
aagat	0	0	10.33	55	0.5%
agatg	3	0.028	12.42	123	1.1%
gatga	6	0.057	12.45	125	1.1%
atgaa	7	0.066	10.91	69	0.6%
tgaat	8	0.075	10.91	69	0.6%
gaatg	6	0.057	11.43	84	0.8%

Fig. 7. Empirical and theoretical (thermodynamically predicted) histograms corresponding to 106 5-ALseq observed for 69 ± 13.3* expected in a sample of 100 circular RNAs of length less than 250 bp (Fig. 6) chosen by chance in[28].

The complete animal genomes[24] show frequently 12-ALsq centred on the D- or Tψ-loops of their tRNAs:

102 times CCATTCAAGATG in Mus musculus (Mouse) RefSeq RNA
Number of letters: 319,391,879. Number of sequences: 106,780
Expected: 19 ± 7*

38 times UGAAUGGUACUG in Oryctolagus cuniculus (Rabbit) RefSeq RNA
Number of letters: 118,511,990. Number of sequences: 42,502
Expected: 7 ± 4.2*

104 times CCATTCAAGATG in Pantroglodytes (Chimpanzee) RefSeq RNA
Number of letters: 206,359,824. Number of sequences: 68,378
Expected: 12 ± 5.5*

33 CCATTCAAGATG in Danio rerio (Zebrafish) RefSeq RNA
Number of letters:155,246,628. Number of sequences: 54,469
Expected: 9 ± 5*

We can also notice that we observe 191 times the 11-ALsq GCCATTCAAGA in the 370 M bp of plasmid genomes instead 88 ± 19* expected.[25]

Nearest-neighbor sequence (5'-3'/3'-5')	**ΔH kJ/mol**	**ΔS J/(mol·K)**	**ΔG37°C kJ/mol**
AA/TT	−33.1	−92.9	−4.26
AT/TA	−30.1	−85.4	−3.67
TA/AT	−30.1	−89.1	−2.50
CA/GT	−35.6	−95.0	−6.12
GT/CA	−35.1	−93.7	−6.09
CT/GA	−32.6	−87.9	−5.40
GA/CT	−34.3	−92.9	−5.51
CG/GC	−44.4	−113.8	−9.07
GC/CG	−41.0	−102.1	−9.36
GG/CC	−33.5	−83.3	−7.66
Terminal A-T base pair	9.6	17.2	4.31
Terminal G-C base pair	0.4	−11.7	4.05

Table 2. Nearest-neighbor thermodynamic parameters for DNA/DNA duplexes.[27]

3.5. The use of the static entropy

By exploiting the thermodynamic parameters given by T. Klingler et W. Rychlik[26] (cf. Table 2), we can calculate the theoretical probability of occurrence of the AL subsequences (cf. Fig. 6).

The static Shannon entropies of the empirical histogram (H_4) and of the predicted distribution of the AL sequences of length 5 (H_5) occurring in 100 circular RNAs of length less than 250 bases (cf. Fig. 7) chosen by chance in[27], are respectively equal to: $H_4 = 2.50232$ and $H_5 = 3.05508$. We can observe that the thermodynamic prediction H_5 is highly greater than all the other entropies H_i ($i = 1,...,4$), showing that the occurrence of AL relics in the present genomes is definitely not due to the chance nor to thermodynamic constraints during the evolution.

3.6. AL-decamers and AL-octamers

By looking to the NCBI data base dedicated to the 16S ribosomal RNAs (with 20143 sequences in The National Center for Biotechnology Information),[29] we see that species like homo sapiens are distributing the occurrences of the AL-decamers* following roughly the Rfam* histogram of Fig. 3, but by focusing only on Bacteria and Archaea*, we see a preferential use of AL-octamers* corresponding to the D-loop and the Ty-loop of the tRNAs, with which the 16S ribosomal RNAs cooperate for building proteins in the ribosome (Fig. 8 left). The distribution of the AL-octamers is more concentrated than that of the AL-decamers, with a Shannon entropy consecutively highly reduced (Fig. 8 right).

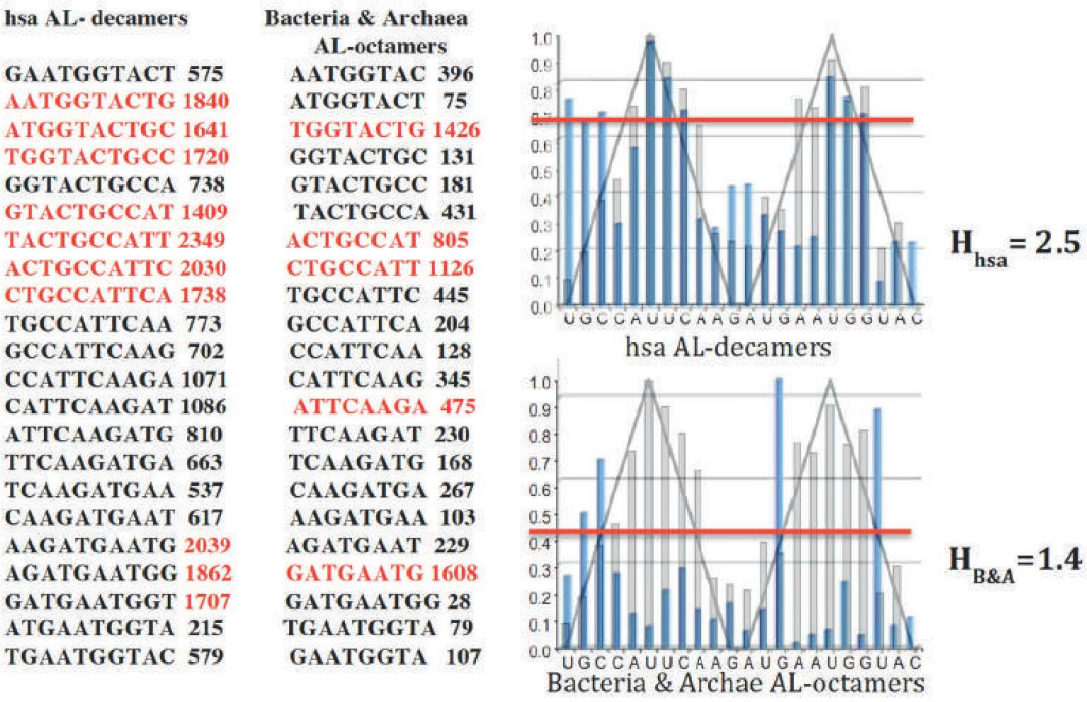


Fig. 8. Empirical histograms corresponding to the observation of AL-octamers in the NCBI data base called 16S ribosomal RNA sequences.[30] The occurrences of AL-decamers in 16S RNAs of homo sapiens (hsa) are given left, those of AL-octamers of Bacteria and Archaea on the middle and the corresponding histograms are represented on the right hand side. On left and middle, are indicated in red the numbers passing the upper limit of the 95%-confidence intervals corresponding to a purely random occurrence, and the red bars on the right hand side is related to the upper frequency corresponding to the 95%-confidence interval of the random occurrence frequencies.

A TTCATGCCA GAC TGGTATGAA **RNA 1**
CATGCCAGAA A TTCTGGTATG A **RNA 2**
C CACTATTCA AGA TGAATGGTG **RNA 3**
C TATTCACCA AGA TGGTGAATG **RNA 4**
TGAATGGTG C TACCATTCA AGA **RNA 5**
TGTGAATGG C CTATTCACA AGA **RNA 6**
TGAATGCCA C TGGTATTCA AGA **RNA 7**
TGAATGTGG C CTACATTCA AGA **RNA 8**
TGAATGCCA AGA TGGTATTCA C **RNA 9**
ATGTTTATGG AGA CCATAAGCA **RNA 10**
TATGTTTGG AGA CCAAGCATA A **RNA 11**
TTCATGCCAG AAA CTGGTATGA **RNA 12**
TGAATGGTA C TGCCATTCA AGA **AL**
TGTTTATGG A CCATAAGCA GAA **RNA 14**
TATGTTTGG A CCAAGCATA AAA **RNA 15**
TATGTTTGG AAG CCAGACATA A **RNA 16**
TACATTTGG AAG CCAGATGTA A **RNA 17**
TGTTTATGG AAG CCATAAATG T **RNA 18**
ATGGA A GCCAT TTACA A TGTAA **RNA 19**
CAGATGG AAG CCATTTG TA A TA **RNA 20**
CATGCCAGA T TCTGGTATG AAA **RNA 21**
TTCTGGTAT G ATGCCAGAA ACA **RNA 22**
AGAACCATG A TATGGTTCT GCA **RNA 23**
AGAATGGTG A TACCATTCT GCA **RNA 24**
TGAATGGT GCCATTCA AG A CT A **AB**

Table 3. List of the 25 RNA rings satisfying the property to have one and only one representative codon per synonymy class of the genetic code, with twice AUG and an internal hairpin or double hairpin structure involving at least 9 nucleotides in their branches.[31]

3.7. Perfect matches

The probability to observe the perfect matches with AL-22mers* by chance is very small ($22/4^{-22} \approx 10^{-12}$), which corresponds for example in the refseq*_representative_genomes[32] to 0.78±1.45* expected occurrences of AL-22mers. We will give some examples of such perfect matches for AL and for another RNA ring sharing with AL and AB the property to have one and only one representative codon per synonymy class of the genetic code, with twice AUG and with a maximal hairpin form[33] (cf. Table 3).

For RNA 24, the observation of 5 perfect matches in the NCBI data base refseq_representative_genomes has a probability $p=10^{-6}$ to be due to random occurrences (the expected number of them being equal to 0.8±1.5*):

Acinonyx* jubatus isolate AJU 981 Chewbacca
(Sequence ID: NW_015131191.1)
Query 17 CAAGATGAATGGTACTGCCATTCA 38
Sbjct 1211929 CAAGATGAATGGTACTGCCATTCA 1211950
AL CAAGATGAATGGTACTGCCATT

Varroa* jacobsoni isolate VJ856
(Sequence ID: NW_019214794.1)
Query 2 ACCATTCTGCAAGAATGGTGATA 24
Sbjct 712 ACCATTCTGCAAGAATGGTGATA 690
RNA24 ACCATTCTGCAAGAATGGTGAT

Varroa destructor
(Sequence ID: NW_019211457.1)
Query 2 ACCATTCTGCAAGAATGGTGATA 24
Sbjct 48915457 ACCATTCTGCAAGAATGGTGATA 48915479
RNA24 ACCATTCTGCAAGAATGGTGAT

Equus caballus isolate (Sequence ID: NC_009174.3)
A-kinase anchor protein 12 isoform X1
Query 16 ATGGTGATACCATTCTGCAAGA 37
Sbjct 15800704 ATGGTGATACCATTCTGCAAGA 15800725
RNA24 ATGGTGATACCATTCTGCAAGA

Equus przewalskii isolate
(Sequence ID: NW_007673354.1)
Query 16 ATGGTGATACCATTCTGCAAGA 37
Sbjct 986088 ATGGTGATACCATTCTGCAAGA 986067
RNA24 ATGGTGATACCATTCTGCAAGA

Mesocricetus auratus isolate Golden Hamster female
(Sequence ID: NW_004801 693.1)
Query 15 AATGGTGATACCATTCTGCAAG 36
Sbjct 4723863 AATGGTGATACCATTCTGCAAG 4723842
RNA24 AATGGTGATACCATTCTGCAAG

Another example is the RNA 19, where 22 21-mers are observed in the NCBI data base Nucleotide collection (nr/nt),[34] for 0.9±1.6* expected ($p=10^{-109}$). It partially fits the palindromic RNA built by a thermophilic DNA polymerase:[35]

ATACA C TGTAT ATACA G TGTAT
ATGGA A GCCAT TTACA A TGTAA **RNA 19**

3.7. Matches with microRNAs

AL and the RNA rings from Table 3 present also good matches with microRNAs having a notable importance in the regulation of the cell metabolism as those listed below, for which we give the localization (in red) and the number of matching nucleotides:

AUGGUACUGCCAUUCAAGAUGA
5'_CGAUGGACGUGACAUUCGUGAAAA_3':hsa-miR-106a AL 14

CAUGGUAAGUAGAACUUACCGU
3'_ AAGCCAAGGAUGACUUGCCGU_5':sb-miR-169 AL 14

UCAAGAUGAAUGGUACUGCCAU
5'-ACAUCCUGCAUAGUGCUGCCAG-3':hsa-miR-448 AL 14

UAGAACUUACCGUCAUGGUAAG
3'_UAGACGUGACAGUCGUGAAAU_5':hsa-miR-106b AL 13

AUGGUACUGCCAUUCAAGAUGA
5'_UGAUGGACGUGACAUUCGUGAAAC_3':hsa-miR-17_5p AL 13

AGUAUCAGAACUUACCGUGGUA
3'_UGUCGACCAACUUCCCCUGGUU_5':mmu-miR-133 AB 13

4. Small RNAs as regulators in genetic networks

4.1. The immunetwork responsible of the Toll Like Receptor (TLR) expression

The activation of Natural Killer (NK) cells, involved in innate immune response, is controlled by PU.1 gene (Fig. 8) and by proteins related the Toll Like Receptors* (TLR). The RP105 (CD180) is such a TLR-related protein identified in B cells (responsible of the humoral adaptive immune response)[36] and acting as regulator of B cell proliferation. B-cells lacking RP105 were shown to be

severely impaired in antibody production. The protein ICAM1 is a type of intercellular adhesion molecule continuously present in low concentrations in the membranes of leucocytes involved in humoral adaptive immune response.

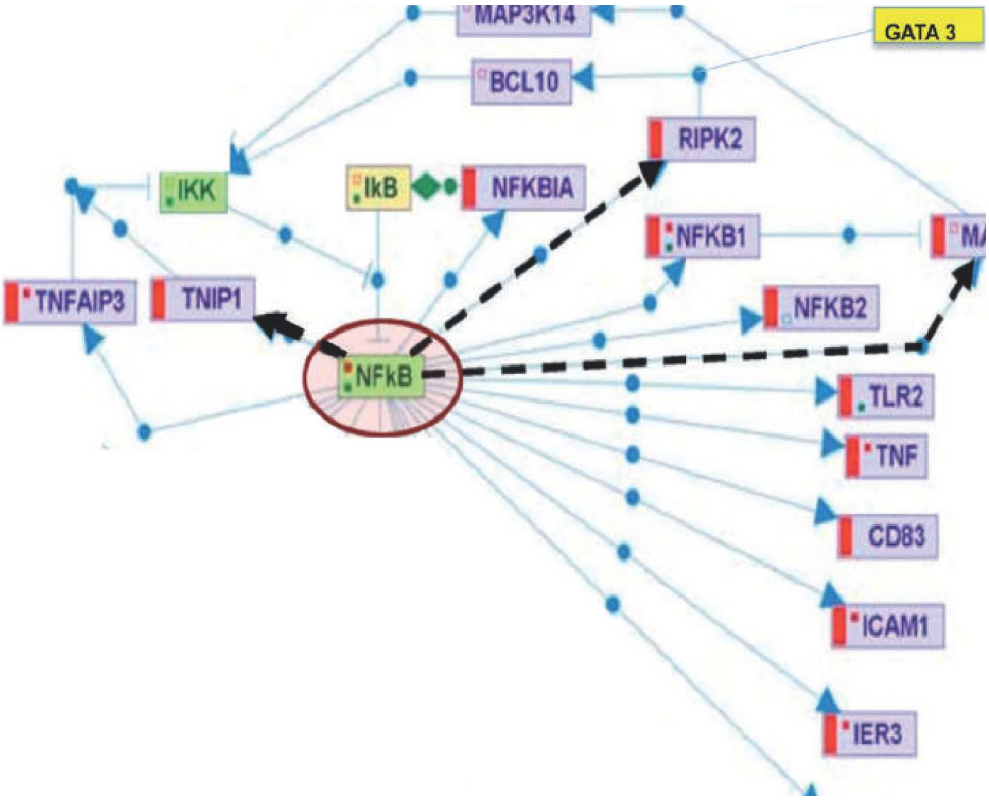


Fig. 9. The network controlling the production of the Toll Like Receptors (TLR) and ICAM1.

The network controlling the Toll Like Receptors (TLR) and ICAM1 expression contains a couple of circuits, one positive 4-circuit tangential to a negative 4-circuit (Fig. 9), giving only one attractor, which corresponds to the activation of TLR2*.

4.2. The links with the microRNAS

All the genes introduced above have links with the microRNAs exerting a negative control on them, then deciding what attractor will occur, by cancelling their target gene activity. For example, we have:

1) for the TLR protein pUNO-hRP105 sub-sequence (4937 bp), the hybridization is made by the microRNA miR 200, close to the reference sequence AL:

5′-CCAUUCAAGAUGAAUGGUACUG-3′ AL 14 anti-matches
3′-UGUAGCAAUGGUCUGUCACAAU-5′ hsa miR 200a 12 anti-matches
5′-UUGUGCUCAUUGAGAUGAAUGG-3′ pUNO-hRP105 sequence starting in position 531
5′-UACUGCCAUUCAAGAUGAAUGG-3′ AL 15 matches

2) for the GATA*-3 gene:

5′-GCCAUUCAAGAUGAA--UGGUACU -3′ AL 13 anti-matches
3′-AGGUAGUAAUGGGC--CGCAUAA-5′ has miR 200c 16 anti-matches
5′-UCUGCAUUUUUGCAGGAGCGUA-3′ GATA-3 sequence starting in position 57

3) for the ICAM1 sub-sequence CD54 cDNA (1615 bp):
5′-GGUGCCUAUUAACAAUAUGAAU-3′ AB 15 anti-matches
3′-GUACGUGUACGUGUG--UAUGUA-5′ has miR 297 15 anti-matches
5′-CCUCCCCA------CCCAC--AUACAU-3′ ICAM1 sequence starting in position 832

We can remark on the above matching that i) gene expressing the TLR hRP105 protein contains the sequence AGATGAA frequently observed in the not coding genome as part of the CAAGATGAA sequence, belonging both to the reference sequence AL and to the Tψ-loop of the tRNAs, which signifies their affiliation to an ancestral genome, and ii) the microRNAs hybridizing the genes TLR hR, GATA-3 and ICAM1 are also close to the reference sequence AL, showing the old origin of the innate immunologic system.

4.3. The IRP/IRE Network

The regulatory network controlling the iron (Fe) metabolism is given on Fig. 10. The arrows are defined by the interactions described in Section 3.3. We calculate the gene expression behavior (called attractor) obtained for any initial condition of the expression of the genes of the network, belonging to the attraction basin of this attractor.

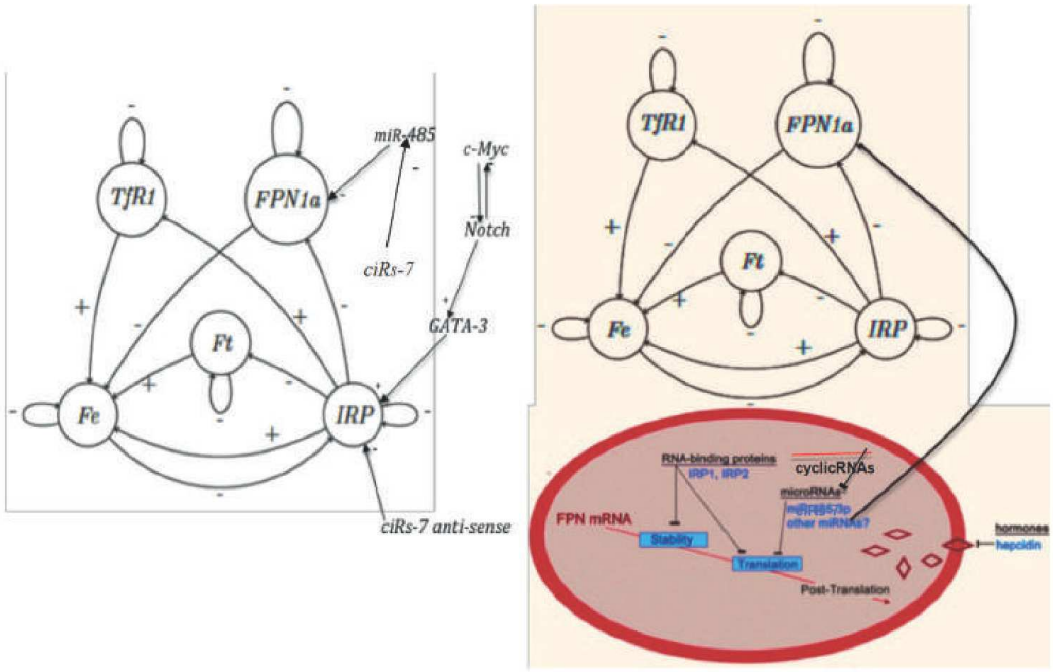


Fig. 10. The interaction graph of the iron regulatory network with inhibitory and activatory interactions (left) and the role[37] of the small RNAs (right).

5. Dynamical entropy and network robustness

We define[38] the energy U and frustration F of a genetic network N of size n by:

$$\forall\, x \in \Omega,\ U(x) = \Sigma_{i,j\in\{1,n\}}\ \alpha_{ij} x_i x_j = Q+(N) - F(x)$$

where x is a configuration of gene expression ($x_i = 1$, if the gene i is expressed and $x_i = 0$, if not), Q+(N) is the number of positive edges in the interaction graph G of the network N and F(x) the global frustration of x, i.e., the number of pairs (i,j) where the values of x_i and x_j are contradictory with the sign α_{ij} of the interaction between genes i and j:

Position	Gene	Fixed Point 1	Fixed Point 2	Limit Cycle 1				Limit Cycle 2	
1	TfR1	0	0	0	0	1	1	0	
2	FPN1a	0	0	0	0	0	0	0	
3	C-Myc	0	1	0	0	0	0	0	
4	Notch	0	0	1	1	1	1	1	
5	GATA-3	0	0	1	1	1	1	1	
6	IRP	0	0	0	1	0	0	1	
7	Ft	0	0	0	0	0	0	0	
8	Fe	0	0	0	0	1	0	1	
9	miR-485	0	0	0	0	0	0	0	
10	ciRs-7 anti-sense	0	0	0	0	0	0	0	
	Relative Attraction Basin Size (RABS)	512/1024	256/1024	216/1024			40/1024		

Table 4. Recapitulation of the attractors* of the iron metabolic system, for parallel updating mode, with the list of expressed (state 0) and not expressed (state 1) genes and the attraction basin* sizes.

$F(x) = \Sigma_{i,\, j\in\{1,n\}} F_{ij}(x)$, where F_{ij} is the local frustration of the pair (i,j) defined by:
$F_{ij}(x) = 1$, if $\alpha_{ij}=1$, xj=1 and $x_i=0$, or $x_j=0$ and $x_i=1$, and if $\alpha_{ij} = -1$, $x_j=1$ and $x_i=1$, or $x_j=0$ and $x_i=0$,
$F_{ij}(x) = 0$, elsewhere.

Then the dynamical entropy *E* of the network is defined by the Kolmogorov-Sinaï entropy* of the transition operator allowing the passage from a state of expression to the next state of expression.[39]

When this transition is stochastic, *E* can be approached by calculating the attractor entropy:

$$E \approx \log_2 2^n - E_{attractor},$$

where $E_{attractor}$ can be evaluated by the quantity: $E_{attractor} = -\Sigma_{k=1,m\leq 2^n} RABS(A_k) \log 2[RABS(A_k)]$, with $RABS(A_k)$ equal to the size of the attraction basin of the attractor A_k, divided by 2^n.

E serves as robustness parameter, being related to the capacity the genetic network has to return to the equilibrium after endogenous or exogenous perturbations.[40] The maximum value of *E* is indeed n and for the immunologic network, there is a unique attractor, hence $E_{attractor} = 0$ and *E* takes this maximum value equal to the number of the genes of the network, i.e., $E = 19$. In the iron network, we have: $E_{attractor} = 2.59412$, hence $E = 10 - 2.59412 = 7.40588$, which corresponds to a relatively low robustness. We can prove more generally[41] that the robustness of the network decreases if the frustration F of the network increases, because of the formula: $\partial E/\partial \log_2 w = -\mathrm{Var}F$, where w is the absolute value of the interaction (here supposed to be the same for all interactions) between the genes of the network.

6. Conclusion

The genetic networks regulated by the small RNAs show important properties necessary to control a biological function. The role of the microRNAs is to provide a partially unspecific inhibitory noise leaving only the control circuits having sufficiently strong interactions to be able to express the attractors. Circular RNAs are inhibiting the microRNAs in order to have, like in neural networks, the possibility to get a double reciprocal influence (inhibitory and anti-inhibitory) on

mRNAs and genes, i.e., on the genetic expression. The presence in a genome of RNA relics such as microRNAs or circular RNAs close to the primitive RNA rings contributes to its robustness by reducing the number of the circuits of the interaction graph (by cutting the links from or to the target genes they inhibit or activate). These circuits are indeed responsible (when positive) for the multiplication of the attractors of the genetic network dynamics.[42] In any case, the use of entropy (static or dynamical) can serve to quantify the role of these regulatory RNAs in maintaining during the evolution their ancestral resilient control of the robustness.[43]

Acknowledgments

The authors acknowledge the European project VPH (Virtual Physiological Human) and the National French project VHP (VisioHome Presence) for their support.

[1] J. DEMONGEOT, "Sur la possibilité de considérer le code génétique comme un code à enchaînement", *Revue de Biomaths* 62, 1978, 61-66; J. DEMONGEOT, N. GLADE, A. MOREIRA *et al.*, "RNA relics and origin of life", *Int. J. Molecular Sciences* 10, 2009, 3420-3441.

[2] J. DEMONGEOT, A. MOREIRA, "A circular RNA at the origin of life", *J. Theor. Biol.* 249, 2007, 314-324.

[3] H. CHOI, K. GABRIEL, J. SCHNEIDER *et al.*, "Recognition of acceptor-stem structure of tRNAasp by Escherichia coli aspartyl-tRNA synthetase", *RNA* 9(4), 2003, 386-393.

[4] T. TANAKA, Y. KIKUCHi, "Origin of the cloverleaf shape of transfer RNA – the double-hairpin model: Implication for the role of tRNA intro and the long extra loop", *Viva Origino* 29, 2001, 119-142; R. W. HOLLEY, J. APGAR, G. A. EVERETT *et al.*, "Structure of a ribonucleic acid", *op. cit*.

[5] B. Lewin, *Genes IX*, *op. cit*.

[6] S. GRIFFITHS-JONES, M. MARSHALL, A. KHANNA *et al.*, "Rfam: annotating non-coding RNAs in complete genomes", *Nucleic Acids Res.* 33, 2005, 121-124. http://rfam.xfam.org/

[7] T. KLINGLER, D. L. BRUTLAG, "Detection of Correlations in tRNA with structural implications", *Intelligent Systems for Molecular Biology* 1, 1993, 225-233; R. W. HOLLEY, J. APGAR, G. A. EVERETT *et al.*, "Structure of a ribonucleic acid", *Science* 147(3664), 1965,1462-1465.

[8] B. LEWIN, *Genes IX*, Boston, Jones & Bartlett, 2008.

[9] J. DEMONGEOT, A. MOREIRA, "A circular RNA at the origin of life", *op. cit*.

[10] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM797670

[11] T. B. HANSEN, E. D. WIKLUND, J. B. BRAMSEN *et al.*, "miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA", *The EMBO Journal* 30, 2011, 4414-4422.

[12] http://sra.dnanexus.com/experiments/SRX097598

[13] S. GRIFFITHS-JONES, M. MARSHALL, A. KHANNA *et al.*, "Rfam", *op. cit*.

[14] *Ibid.*; S. MEMCZAK, M. JENS, A. ELEFSINIOTI *et al.*, "Circular RNAs are a large class of animal RNAs with regulatory potency", *Nature* 495(7441), 2013, 333-338.

[15] http://www.circbase.org/cgi-bin/tablebrowser.cgi

[16] http://starbase.sysu.edu.cn/

[17] http://www.ncbi.nlm.nih.gov/nuccore/JQ

[18] *Ibidem*.

[19] *Ibid*.

[20] https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&BLAST_SPEC=MicrobialGenomes

[21] http://www.google.bj/patents/CN100591773C?cl=en

[22] http://www.candidagenome.org/download/sequence/C.../Ca20chr5.embl

[23] http://www.phytozome.net/cgibin/gbrowse_details/soybean?class=mRNA;name=Glyma14g34640.1

[24] http://blast.ncbi.nlm.nih.gov/

[25] *Ibidem*.

[26] T. KLINGLER, D. L. BRUTLAG, "Detection of Correlations in tRNA with structural implications", *op. cit.*; W. RYCHLIK, W. J. SPENCER, R. E. RHOADS, "Optimization of the annealing temperature for DNA amplification in vitro", *Nucleic Acids Res.* 18(21), 1990, 6409-6412.

[27] http://circbase.mdc-berlin.de

[28] http://www.circbase.org/cgi-bin/tablebrowser.cgi

[29] https://www.ncbi.nlm.nih.gov/nuccore?Db=nuccore&Cmd=DetailsSearch&Term=NR_00000%5BACCN%5D+%3A+NR_999999%5BACCN%5D+AND+%2216S+ribosomal+RNA%22%5Btitle%5D

[30] *Ibidem*.

[31] J. DEMONGEOT, A. MOREIRA, "A circular RNA at the origin of life", *op. cit*.

[32] https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/

[33] J. DEMONGEOT, A. MOREIRA, "A circular RNA at the origin of life", *op. cit*.

[34] http://blast.ncbi.nlm.nih.gov/

[35] X. LIANG, K. JENSEN, M. D. FRANK-KAMENETSKII, "Very Efficient Template/Primer-Independent DNA Synthesis by Thermophilic DNA Polymerase in the Presence of a Thermophilic Restriction Endonuclease", *Biochemistry* 43(42), 2004, 13459-13466.

[36] C. SANGOKOYA, J. F. DOSS, J. T. CHI, "Iron-Responsive miR-485-3p Regulates Cellular Iron Homeostasis by Targeting Ferroportin", *PLoS Genet* 9(4), 2013, e1003408; http://mirdb.org/miRDB/

[37] Ibid. ; http://mirnamap.mbc.nctu.edu.tw/

[38] C. SANGOKOYA, J. F. DOSS, J. T. CHI, "Iron-Responsive miR-485-3p Regulates Cellular Iron Homeostasis by Targeting Ferroportin", *op. cit.*; http://mirnamap.mbc.nctu.edu.tw/
[39] L. DEMETRIUS, "Statistical mechanics and population biology", *J. Stat. Physics* 30(3), 1983, 709-753; L. DEMETRIUS, "Directionality principles in thermodynamics and évolution", *PNAS* 94(8), 1997, 3491-3498.
[40] *Ibidem*.
[41] C. SANGOKOYA, J. F. DOSS, J. T. CHI, "Iron-Responsive miR-485-3p Regulates Cellular Iron Homeostasis by Targeting Ferroportin", *op. cit.*; http://mirnamap.mbc.nctu.edu.tw/
[42] J. DEMONGEOT, J. ARACENA, F. THUDEROZ, "Genetic regulation networks: circuits, regulons and attractors", *C.R. Biologies* 326(2), 2003, 171-188.
[43] J. DEMONGEOT, H. HAZGUI, "The Poitiers school of mathematical and theoretical biology: Besson-Gavaudan-Schützenberger's conjectures on genetic code and RNA structures", *Acta Biotheoretica* 64, 2016, 403-426; J. DEMONGEOT, M. JELASSI, H. HAZGUI *et al.*, "Biological Networks Entropies: Examples in Neural Memory Networks, Genetic Regulation Networks and Social Epidemic Networks", *Entropy* 20(1), 2018, 36.

Glossary

Acinonyx: cheetah.

AL-22mers: sequence of 22 nucleotides common to AL.

Amino acid: organic molecule containing two radicals, amine (-NH2) and carboxyl (-COOH), along with a side chain specific to each of the 22 amino acids existing in living systems.

Archaea: domain of ancient unicellular micro-organisms. They constitute the third kingdom of the life, the others being bacteria and eukaryota.

Attraction basin: set of all initial states visiting the same attractor when the time tends to infinity.

Attractor: set of all states of a system visited from an initial state when the time tends to infinity.

Barycenter: object being at the minimal sum of distances to the other objects of a set.

Bp: pair of bases (or nucleotides).

hsa circRNAs: homo sapiens circular RNAs located in the cell cytoplasm and inhibiting microRNAs.

Circular Hamming distance: distance between 2 RNA rings counting the number of transformations (rotation, transposition or mutation) necessary for passing from a ring to another.

Codon (or triplet, or trinucleotide): small nucleic chain consisting of three of the four DNA bases, and coding for an amino acid or for a punctuation signal used in the messenger RNA translation. Because there are 43 = 64 codons, the assignment of a codon to an amino acid is not equi-distributed, certain amino acid (like methionine) being represented by only one codon and others by 6 (like serine).

Decamer: sequence of 10 nucleotides.

D-loop: part of the tRNA representing the left leaf of the tRNA cloverleaf 2D secondary structure. It interacts with the Ty-loop for constituting the tRNA 3D tertiary structure in form of the letter t.

GATA gene: the corresponding protein mediates the differentiation of specific types of blood cells from their hematopoietic precursors.

Genome: set of all genes of an individual in a species.

HEK cell: Human embryonic kidney cell.

HELA: immortal cell line used in scientific research.

Kolmogorov-Sinaï entropy: quantity linked to the capacity a stochastic system has to return at its equilibrium measure after perturbation. If the stochastic dynamics is ruled by a Markovian matrix, this entropy is just the mean (for the invariant measure) of the entropies of each line (equal

to a distribution on the state space of the system) of the Markovian matrix.

Limit-cycle: attractor made of a set of states visited periodically when the time tends to infinity.

Mitochondrion: organelle ensuring cell respiration and energy production. It contains key enzymes like the translocase (transporting cellular ADP inside mitochondrion and mitochondrial ATP outside) and ATPase (transforming ADP in ATP inside mitochondrion).

microRNA: small RNA in chain or hairpin form inhibiting the translation by fixing mRNAs.

mRNA: messenger RNA is a nucleic molecule that contains the succession of the triplets coding for the amino acids of a protein.

Nucleotide (or base): organic molecule, which constitutes the simplest component of the deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). There are 5 nucleotides, adenine (A), thymine (T) replaced by uridine (U) in RNA, guanine (G)) and cytosine (C)

Octamer: sequence of 8 nucleotides

Peptide: a small protein made of several amino acids linked by peptide bonds, i.e., covalent chemical bonds linking two consecutive amino acid monomers along a peptide chain.

Rfam: database collecting RNA families.

Refseq: reference sequences, a data base of the NCBI (US National Center for Biotechnology Information).

Ribosome: intra-cellular structure in which occurs the translation of the messenger RNA into a protein.

Stereochemical theory of the genetic code: theory implying that the first amino acids to be incorporated into early peptides used a stereochemically-determined code, i.e., a direct specific interaction between amino acids and their codons and anti-codons.

Start codon: trinucleotide used as punctuation signal in the messenger RNA translation, indicating that the translation has to start. Start codon is AUG, coding also for the methionine.

Stop codon: trinucleotide used as punctuation signal in the messenger RNA translation, indicating that the translation has to stop. There are 3 stop codons: UAA, UAG, UGA.

Toll Like receptors (TLR): family of membrane proteins (like TLR2) acting as receptors, recognizing foreign substances and then, transmitting a signal to the cells of the immune system.

tRNA: transport RNA is a nucleic molecule transporting an amino acid in the site of protein elongation of the ribosome.

Ty-loop: part of the tRNA representing the right leaf of the tRNA cloverleaf 2D-secondary structure, interacting with the D-loop for building the tRNA 3D-tertiary structure in form of letter t.

Transcription factor: protein controlling (activating or inhibiting) the transcription of genetic information from DNA to messenger RNA, by binding to a specific DNA sequence.

van der Waals force: weak and short-range attractive chemical force between molecules., which increases with the total area of contact between molecules. It becomes repulsive at very short distance resulting from the Pauli exclusion principle, which prevents the collapse of molecules.

Varroa: mite parasite of the honey bees.