computation

MDPI

Article

Genomic Phylogeny Using the Maxwell^{TM} Classifier Based on Burrows–Wheeler Transform

Jacques Demongeot [1,*], Joël Gardes [2], Christophe Maldivi [2], Denis Boisset [2], Kenza Boufama [1] and Imène Touzouti [1]

[1] Laboratory AGEIS EA 7407, Team Tools for e-Gnosis Medical, Faculty of Medicine, University Grenoble Alpes (UGA), 38700 La Tronche, France; boufama.kenza@gmail.com (K.B.); imenetouzouti20@gmail.com (I.T.)
[2] Orange Labs, 38229 Meylan, France; joel.gardes@orange.com (J.G.); christophe.maldivi@orange.com (C.M.); denis.boisset@orange.com (D.B.)
* Correspondence: jacques.demongeot@univ-grenoble-alpes.fr

Abstract: Background: In present genomes, current relics of a circular RNA appear which could have played a central role as a primitive catalyst of the peptide genesis. Methods: Using a proximity measure to this circular RNA and the distance, a new unsupervised classifier called Maxwell^{TM} has been constructed based on the Burrows–Wheeler transform algorithm. Results: By applying the classifier to numerous genomes from various realms (Bacteria, Archaea, Vegetables and Animals), we obtain phylogenetic trees that are coherent with biological trees based on pure evolutionary arguments. Discussion: We discuss the role of the combinatorial operators responsible for the evolution of the genome of many species. Conclusions: We opened up possibilities for understanding the mechanisms of a primitive factory of peptides represented by an RNA ring. We showed that this ring was able to transmit some of its sub-sequences in the sequences of genes involved in the mechanisms of the current ribosomal production of proteins.

Keywords: ribosome evolution; unsupervised clustering; Maxwell^{TM} classifier; Burrows–Wheeler transform; primitive peptide factory; Archetypal Loop ring

Citation: Demongeot, J.; Gardes, J.; Maldivi, C.; Boisset, D.; Boufama, K.; Touzouti, I. Genomic Phylogeny Using the Maxwell^{TM} Classifier Based on Burrows–Wheeler Transform. *Computation* **2023**, *11*, 158. https://doi.org/10.3390/computation11080158

Academic Editor: Shizuka Uchida

Received: 8 June 2023
Revised: 9 August 2023
Accepted: 9 August 2023
Published: 11 August 2023

1. Introduction

Among the molecules that have possibly played an important role in the origin of life on Earth, the first RNAs and peptides were formed by chance through a concatenation process among the nucleotides and amino acids pools, respectively, synthesized from the atoms (C, O, H, and N) of the primitive atmosphere due to sufficient electrical discharge [1,2]. They combined in same favorable sites (volcanic hot spring pools [3], clays like montmorillonite [4], alkaline hydrothermal vent/serpentinization [5], etc.) giving rise to large polymers, e.g., circular RNAs and proteins, whose interactions allowed their reproduction and isolation from the external environment. RNA core was made of rings or chains with catalytic properties helping amino acids to bind together. Peptides created via this peptide-bonding was later combined with lipids synthesized in the primitive atmosphere [6]. They could also assist with the synthesis of new RNA rings or chains that could serve further as ribozymes catalyzing the protein synthesis [7,8] as demonstrated in short segments of RNA [9,10]. By looking for the minimal circular RNA that first facilitated these interactions, we have previously identified an RNA structure [11,12], called AL (Archetypal Loop), capable of catalyzing peptide bonds between amino acids in its ring form (Figure 1A) and resisting denaturing environmental conditions in its hairpin form (Figure 1B) [13]. The AL sequence can be considered as the consensus sequence of tRNA loops of many species (only 4 species on Figure 1C and 242 others from GtRNAdB (see [14] and Supplementary Materials Table S1): *ATGGTACTGCCATTCAAGATGA* [15]. The RNA AL has interesting combinatorial properties: it comprises 22 nucleotides and offers 20 successive codons capable of binding transiently to the 20 amino acids of which they are the representatives in the genetic code

via overlapping [15]. This AL structure is unique for being the barycenter (for the circular Hamming distance) of a set of only 25 other possible solutions with a minimal of 22 nt length and with these combinatorial properties. Moreover, if AL starts with *AUG*, it ends with *UGA*, which is the punctuation codons of the genetic code.

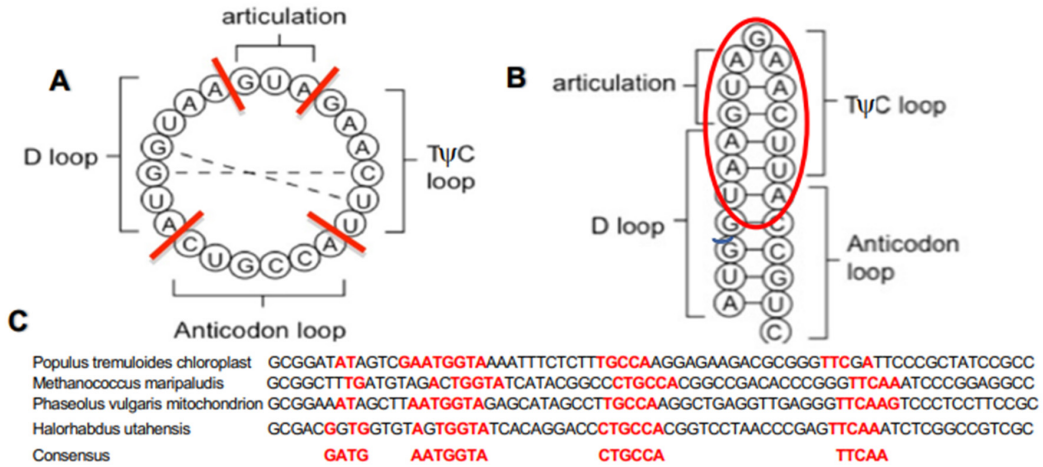

Figure 1. (**A**) Ring form of the Archetypal Loop (AL) with indication of the tRNA loops; (**B**) hairpin form of AL with indication of the upper part containing the nine P-pentamers; and (**C**) examples of tRNA-Gly of different species (from [14]).

It has been proven that amino acids have an affinity with their cognate codons and anti-codons involving weak electromagnetic or van der Waals forces [16,17], which causes transient binding between amino acids and the AL ring containing the corresponding cognate triplets of nucleotides, and after being spatially close together, amino acids can bind to each other or to a neighboring peptide, with the mechanism being analog to that of the present protein synthesis in current cells. This mechanism was proposed by Katchalsky [18] and Eigen [19], which showed that RNA, in particular the ancestors of current transfer RNA, could have been involved in a primitive matrix capable of catalyzing the synthesis of both peptides and new RNAs, favoring the emergence of an RNA world made of RNA molecules with catalytic and replicative properties [20,21].

2. Materials and Methods

2.1. Calculation of the Archetypal Loop-Proximity

The methodology chosen starts from the calculation of a proximity called the AL proximity, which estimates the degree of possible heritability from the AL of an RNA sequence. The sequences are obtained from the RefSeq database of NCBI (National Center for Biotechnology Information) [22], which contains the genomes of many species. On 5 May 2023, the RefSeq Release 218 included the genome of 133,740 organisms, with 52,503,423 of mRNA transcripts of 260,776,371 proteins of which the gene contains 24,000 nucleotides and the mRNA transcript 1300 nucleotides on average in humans. The method used to compare an RNA sequence to the AL involves counting the number of common pentamers between those of the sequence and those located at the upper extremity of the hairpin form of the AL, which belongs to the following set, *P*, of 9 pentamers:

$$P = \{AUUCA, UUCAA, UCAAG, CAAGA, AAGAU, AGAUG, GAUGA, AUGAA, UGAAU\}.$$

The 9 elements of *P* are called *P*-pentamers. They are extracted from an AL sequence located near the head of the hairpin form of the AL. We use *P* for defining a criterion of proximity to the AL for any RNA sequence, that is, the number of standard deviations (SDs) between calculated and expected numbers of *P*-pentamers in the chosen sequence. For example, let us consider the nucleotide sequence of length $n = 2697$ observed for the mRNA of the nucleolin of Camelus dromedarius (Figure 2). Then, because the probability of ob-

serving a pentamer by chance is $p = 9/1024$, the average number of expected P-pentamers is $np = 2720 \times (9/1024) = 23.9$, with a standard deviation $\sigma = np(1-p) \sim 23.9^{1/2} \sim 4.9$. The number of calculated P-pentamers in the sequence is equal to 95; then, the difference between calculated and expected numbers is $95 - 23.9 = 71.1$, corresponding to $71.1/4.9 = 14.5\sigma$.

Nucleotide sequence of the Nucleolin gene

CTGGGCAGGCTCAGTCTTTCGCCTCAGTCCCGAGCTCTAGCTGGCAGCTAAGCGTACGCGCTCTGGGATCTTCAGTAC
CAGCGGCCGCCATCGACGTTGCTTGGGTTCCTTTGGGCTCATCTGCGCCACTCGCTCGCTATACACTCCGCTGTCATA
AACCTCCATCATGGTAAAGCTCGCAAAGGCTGGTAAAAA**TCAAG**GTGACTC**CAAGA**AAATGGCTCCTCCCCCAAAGGA
GGTAGAAG**AAGAT**AGCGAAGACGAGG**AGATG**TCAG**AAGATGAAGATGATGA**TAGCAGTGGGGAAGAGGTCATTATTCC
TCAGAAGAAAGGCAAAAAGGCTACCATAACTCCAGCAAAGAAGGTCATGGTTTCCCCAACAAAAAAGGTTGCAGTTGC
CACACCGGCAAAGAAAGCAGTTGTTACCCCTGGCAAAAAGGCAGCGGCTACGGCAGCTAAGAAGACAGTTACACCACC
TGCCAAAGCAGTAGCGACACCTGG**CAAGA**AGGGAGCCACACCAGGCAAAGCGTTGGTAGCAACCCCTGGTAAGAAGGG
AGCAGCCGCCCCAGCCAAGGGAGCAAAGAATGGTAAAAATGCCAAAAAGGAAGAGAGT**GATGA**GG**AAGATGATGA**CGA
CAGTGAAGAGGAG**GATGA**G**GATGA**CG**AAGATGA**G**GATGAAGATGAAT**TTGAGCCAGCAGT**GATGAA**AGCTGCTGCTGC
TGCTCCTGCCTC**AGATGA**TGAG**GATGATGA**G**GATGAAGATGA**T**GATGAA**GAC**GATGATGA**G**GATGA**T**GATGA**G**GATGA
AGAGGATTCTGAAGAAGAAGCTATGGAGACTGCACCACCCAAAGGAAAGAAAGCTCCAGCAAAAGCTGTTCCTGTGAA
GGC**CAAGA**GTGCTGCTGAG**GATGAAGATGAA**GAG**GATGA**C**GATGA**GGACGAGGAG**GATGAA**GAG**GATGA**CGACGAGGA
G**GATGA**TGAG**GATGA**C**GATGA**GGAGGAGGAGGAGGAC**GATGA**GGAGGAGGAGGAGGAGGAAGAGGAGGAAGAGCCTGT
CAAAGAAGCACCTGGAAAACGAAAGAAAGAAATGGCCAAACAGAAAGCAGCTCCTGAAGC**CAAGA**AGCAGAAAGTGGA
AGGCACAGAACCAACCACAGCTTTAATCTCTTTGTTGGAAACC**TGAAT**TTCAGCAAATCTGCTCC**TGAAT**TAAAAACG
GGTATCAGTGACCTTTTTGCTAAAAATGATCTTGCTGTTGTCGATGTCAGAATTGGTGTGTCTAGGAAGTTTGGCTAT
GTGGATTTTGAGTCTGCTGAAGACCTGGAAAAAGCCTTGGAACTCACTGGTTTAAAAGTCTTTGGCA**ATGAA**ATTAAA
CTAGAAAAACCAAAGGGAAAAGACAGTAAGAAGGATCG**AGATG**CGAGAACACTTTTGGCTAAAAATCTGCCTTATAAA
GTTACTCAG**GATGAAT**TAAAAGAAGTGTTTG**AAGATG**CTGTGGAGATCAGATTAATCAGCAAGGATGGAAAGAGTAAA
GGGATTGCTTTTAT**TGAAT**TTAAGACAGAAGCCGATGCAGAGAAAACCTTGGAAGAAAAGCAGGGAACAGAGAT**AGAT**
GGACGATCCATTTCCCTGTACTATACTGGAGAGAAAGG**TCAAG**GT**CAAGA**CCCTAGAGGTGGAAAGAATAGCACTTGG
AGTGG**TGAAT**CAAAAACTCTGGTTTTAAGCAACCTTTCCTTCAGTGCAACAGAAGAAACTCTTCAGGAAGTATTTGAG
AAGGCAACTTTTATCAAAGTGCCCCAGAGCCAAAATGGCAAATCTAAAGGGTATGCGTTTATAGAATTTGCTTCCTTT
G**AAGATG**CTAAAGAAGCTTTAAATTCTTGTAATAAAAGGGAAATTGAGGGCAGAGCAATCAGGCTGGAGCTGCAAGGA
CCCAGGGGATCACCTAATG**CAAGA**AGCCAGCCATCCAAAACTCTGTTTGTCAAAGGTCTGTCTGAGGATACCACAGAA
GAGACATTAAAGGAGTCGTTTGAGGGCTCTGTCCGTGCAAGGATAGTCACTGACCGGGAAACCGGGTCCTCCAAAGGG
TTTGGTTTTATAGAC**TTCAA**CAGCGAGG**AAGATG**CCAAAGCTGCCATGGAGGCCATGG**AAGATG**GTGAAATTGATGGA
AACAAAGTTACTTTGGACTGGGCCAAGCCTAAGGGTGAAGGTGGCTTTGGTGGCCGAGGAGGAGGCAGAGGTGGCTTT
GGAGGCCGAGGTGGTGGCAGAGGTGGCCGAGGTGGATTTGGTGGCAGAGGCCGGGGAGGCTTTGGAGGGCGAGGAGGC
TTCCGAGGAGGCAGAGGAGGAGGAGGAGACCACAAGCCACAAGGAAAGAAGACGAAGTT**TGAAT**AGTTTCTTCTATCC
CTGTCTTTCCCTCTTCCATTTGAAAGAAAGGACTCTGGGGTTTCTACTCTGTTACCAATGACAGAGCCTTCTGAGGAC
ATTCCAAGGCAGTATACAGCCCTGTGGTCTACTCGGAAATCCATATAGATAACATTTCACGGGAGATAACCCTGTTGG
TTTTGACTGGAT**ATTCA**GATAAACTTTTTAAAG**AGATGA**GTGATAGAGCTAACCCTTATCTGTAAGTTT**TGAAT**TTAT
ATTGTTTTTTCCCATGTACAAAACCATTTTTTTCCT**ATGAA**GTTTTTTGTTTGTTTTTTTTGCGTTGGGGGGTGTAAA
GGAAAGCAGAATGTTTTATCATGATTTTTGCTTCAGCAACTTTGGGACAAATTAAAAGTCCTAAACTCTGGT

Figure 2. mRNA sequence of nucleolin gene (NCL) of Camelus dromedarius breed African isolate Drom800 chromosome 5 (graphic extracted from NCBI Reference Sequence: XM_010985648,2 [22]). The P-pentamers are indicated in red bold (with possible overlaps).

Because the Bernoulli distribution of the P-pentamers checks the approximation conditions by the Gaussian distribution, $n = 2720 \geq 30$, $np \sim 23.9 \geq 5$, and $n(1-p) \sim 2696 \geq 5$, so the probability of observing such a difference is less than $1 - F(14.5) < \text{Proba}(\{X \geq 14.5\})$, where F is the Gaussian distribution function. Then, using the Gaussian distribution function approximation proposed in [23], we get: $\text{Proba}(\{X \geq t\}) \sim 0.5 - (1 - \exp(-at^2))^{1/2})/2$, where $a = 0.647 - 0.021t$. Hence, $t = 14.5$, $a = 0.3425$ and $\text{Proba}(\{X \geq 14.5\}) \sim \exp(-0.3425 \times 210.25)/4 \sim 1.33 \times 10^{-14}$. Because the value of the difference between the calculated and expected numbers of pentamers expressed as a number of standard deviations σ is directly linked to the probability of observing this difference, we retain this quantity as a measure of any RNA sequence's proximity to the AL, called P-proximity.

If the ring AL has played a role in building the first peptides, it is reasonable to search for the remnants of its nucleotidic sequence inside RNAs playing the same role in building the current proteins, e.g., ribosomal RNAs and mRNA of ribosomal proteins or of proteins favoring the accretion of the ribosomal components.

The number of P-pentamers calculated in Figure 2 is 95 and the expected number is 23.9, with a difference equal to 14.5σ, where σ is equal to the standard deviation of the Bernoulli empirical distribution corresponding to the P-pentamers observed by chance. Then, the probability to observe these 95 P-pentamers equals to about 10^{-14}. It is possible to search for relics such as the P-pentamers common to the AL and to rRNAs and mRNAs whose function is considered identical in the ribosomes of multiple species, like the mRNA of proteins nucleolin and NOL11 (see Supplementary Materials Tables S2 and S3). After calculating their P-proximity, we classify the corresponding mRNAs of various species

using the classifier Maxwell^{TM}, which is able to compare sequences of symbols [24], here the sequences of nucleotides, and conclude if the obtained clusters are coherent with the *P*-proximity values of their elements.

2.2. The Burrows–Wheeler Transform

The Burrows–Wheeler transform [25] is an algorithm used in lossless compression procedure which rearranges strings into runs of similar characters in a reversible way. Associated with a run-length algorithm, we obtain a function we use in "Normalized Compression Distance" (NCD) or Vitányi distance, in order to find similarities between them, like same repetition of motifs, same deletion or insertions, etc. The reason for the implementation of this "simplified" compression algorithm was to retrieve the symmetry of NCD. It is particularly convenient to compare genomic sequences independently of their length if they have coevolved under the action of the same operators. In evolution, there are 11 different genomic operators: Crossing-over, Mutation, Translocation, Insertion, Deletion, Transposition, Inversion, Repetition, Symmetrization, Palindromization, and Permutation. When these operators are used with the same frequency during evolution, Burrows–Wheeler transform serves to compress the sequences of the same origin which have similar evolutionary history.

First, Burrows–Wheeler transform involves organizing the circular permutation of a word following the lexicographic order, then taking the last letter of these permuted words and calculate the run-length encoding (RLE) of this new word formed by the rank of the permutation identical to the initial word followed by the sequence of the last letters of permuted words, by indicating before the number of repeated letters (Figure 3). This coding constitutes a lossless compression method and during decompression, the initial word can be reconstructed exactly from this information in a reversible (or adiabatic) way.

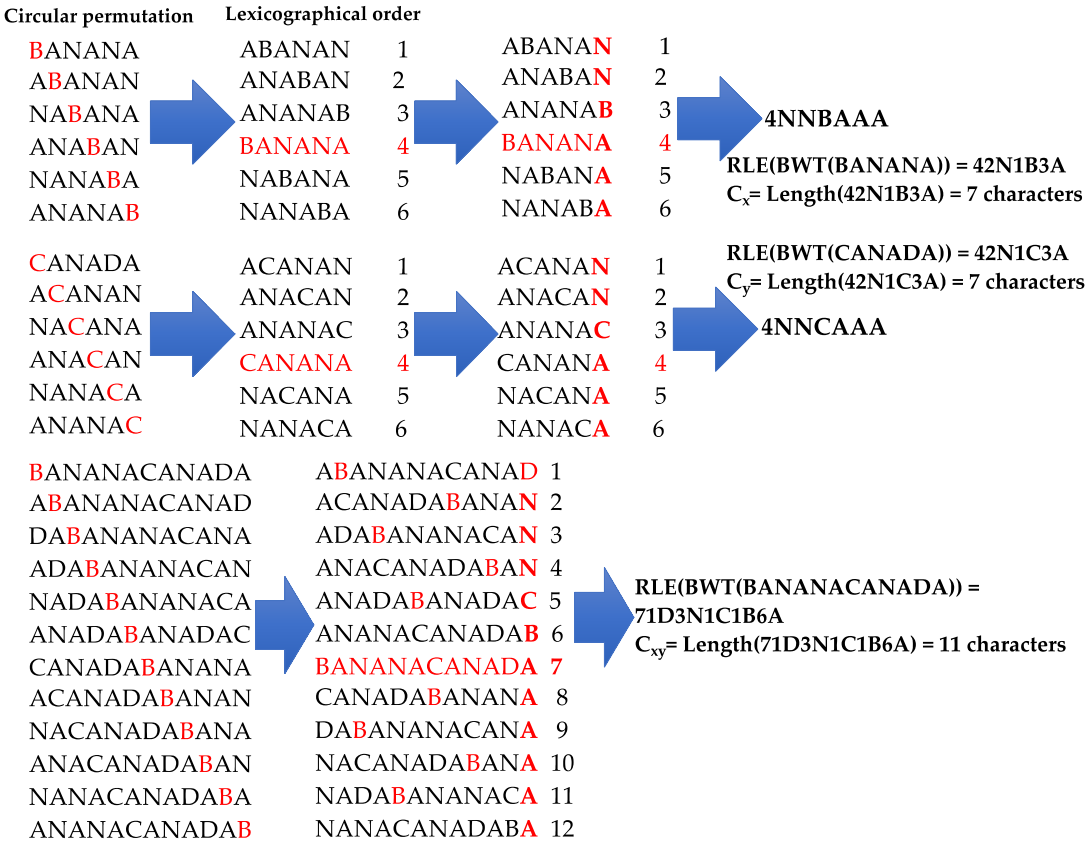

Figure 3. Burrows–Wheeler transform (BWT) of two words BANANA and CANADA, with two mutations B/C and N/D. Lengths of run-lengths (RLE) of BWT transforms of BANANA, CANADA and concatenation BANANACANADA are, respectively, 7, 7 and 11 characters. The red words represent the initial words changed during the Burrows-Wheeler transform.

2.3. The Vitányi Distance

The Vitányi distance between two sequences x and y [26,27] involves calculating the length of the RLE version of their Burrows–Wheeler transform (BWT), that is, the values of the coefficients C_x = Length[RLE(BWT(x))] and C_y = Length[RLE(BWT(y))], respectively, and then the value of the coefficient for the concatenated word xy, C_{xy} = Length[RLE(BWT(xy))] and calculating the ratio (Figure 3): $d(x,y) = [C_{xy} - \min(C_x,C_y)]/\max(C_x,C_y)$. Vitányi distance using Burrows–Wheeler transform and run-length between words BANANA and CANADA is equal to 0,57 (Figure 3). Vitányi distance is a real mathematical distance, with $d(x,x) = 0$, $d(x,y) = d(y,x)$ (symmetry) and $d(x,z) \leq d(x,y) + d(y,z)$ (triangular inequality).

2.4. The Maxwell^{TM} Classifier

The principle of the Maxwell^{TM} classifier [26] is to constitute clusters of words belonging to the set $\{x_i\}_{i=1,n}$, from which the distance matrix $D_{ij} = d(x_i,x_j)$ has been calculated. Then, each triplet of words constitutes a triangle in the graph associated with D and the area of this triangle is calculated using the classical Héron formula, and the original algorithm of Maxwell^{TM} has the following steps:

- Calculating the mean and standard deviation on histograms of triangle areas for filtering "large and deformed triangles" considered as outliers of the empirical distribution according to the number of standard deviations retained;
- Examining sub-graphs whose "useless" (respectively "best") representative edges are identified as attached to the least (respectively the most) connected nodes and removing them (respectively keep them as cluster central node);
- Processing sub-graphs with several local minima (i.e., nodes whose neighborhood does not contain another node that is closer to the sub-graph than the node itself) using Voronoï networking with the software Graphviz [28] for detecting internal boundaries;
- Testing at the end for sub-graphs whose mean and standard deviation are varied until Graphviz no longer detects any boundaries;
- Storing elements rejected by this statistics calculation in the form of "singleton clusters";
- Final recalling by clustering the population of singletons to detect new clusters.

3. Results

Table 1 shows that RNA classes whose content is homogeneous in AL-proximity, i.e., in evolutionary age (if the hypothesis on the primitivity of AL is true), are marked both by a large AL-proximity and by an upstream position in the Maxwell^{TM} classification tree (Figure 4). They correspond to ancient species in purely biological phylogenetic trees, calculated without reference to an ancestral RNA, and resulting only from comparisons between the genomic sequences of the compared species (Figure 5). The Maxwell^{TM} classification tree proposes a series of clusters organized from the root of the tree until its leaves and the content of each cluster is as presented in Table 1, which shows that one of the rules explaining the grouping in a class is the proximity to the AL of its members. It should be noted that at the root of the tree, where the hypothetical LUCA (the Last Universal Common Ancestor, defined first by C. Woese and G. Fox as the first living system [29,30]) is often placed, the ancient species of Salinarchaeum appears, which belongs to the very ancient classes of Archaea and Halobacteria from the Euryarchaeota branch (Figure 5).

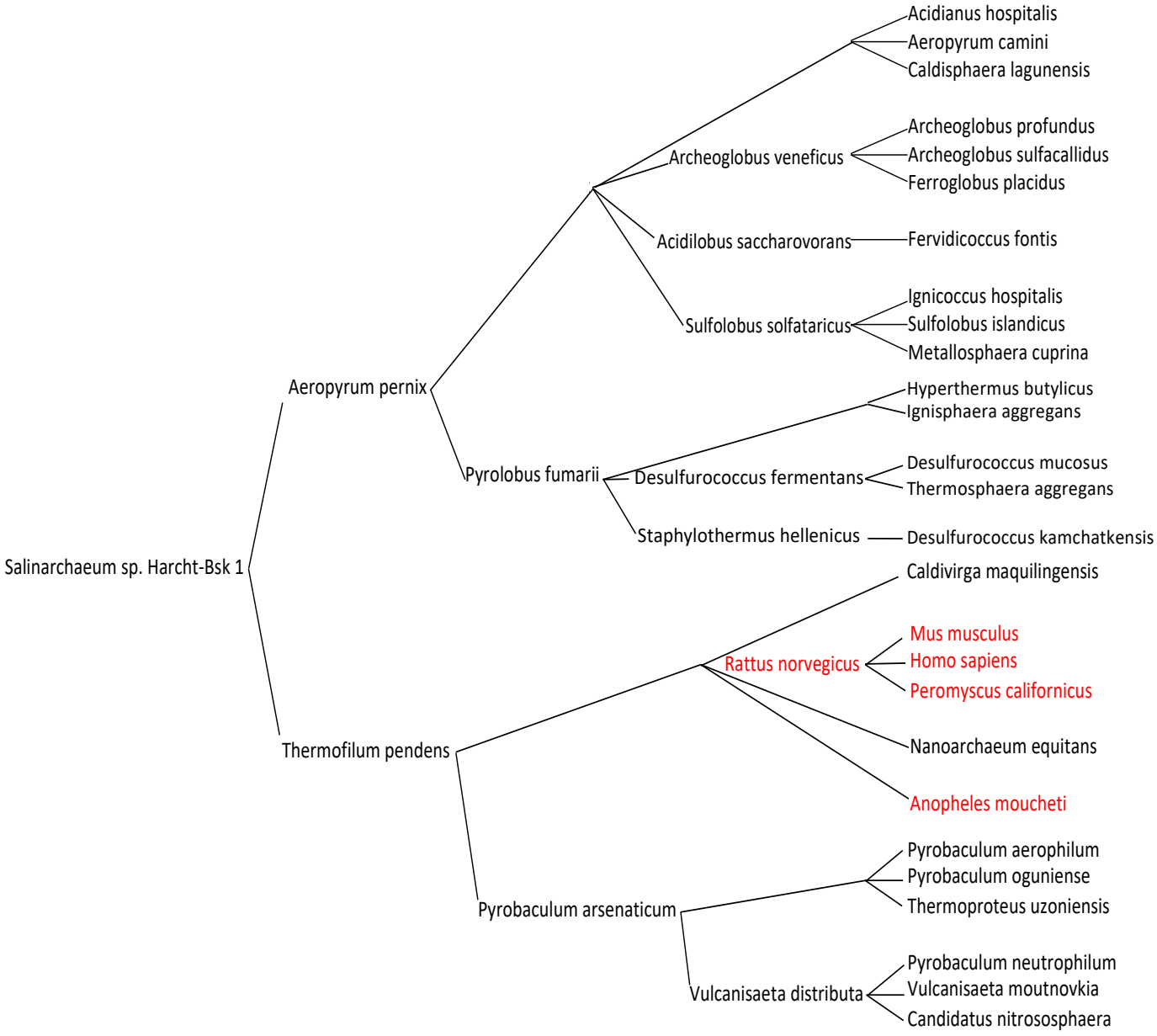

Figure 4. Representation of a part of the Maxwell^{TM} classification tree from the 5S ribosomal RNA (in black) and nucleolin mRNA (in red) of different species (see Supplementary Material).

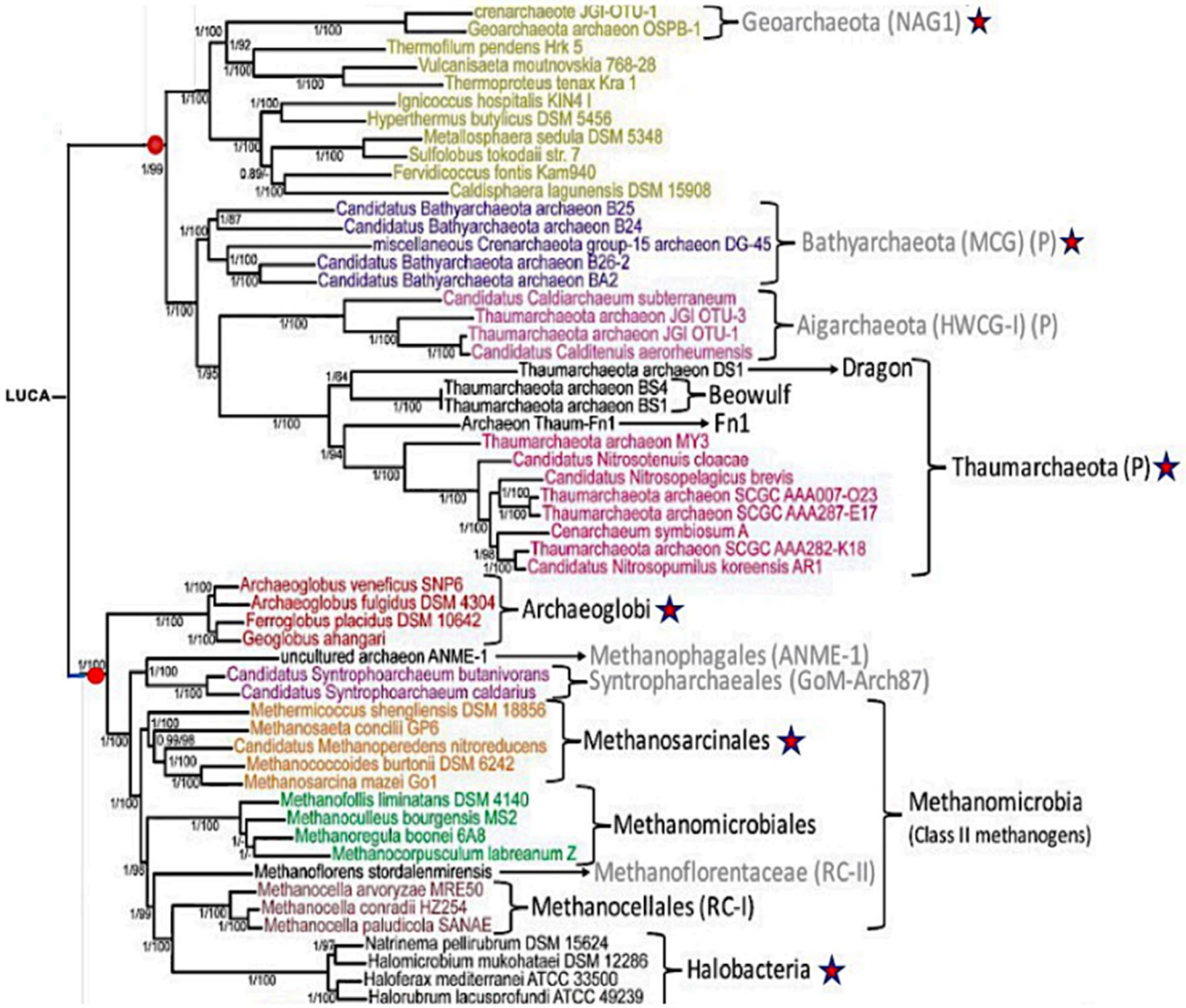

Figure 5. From [30], phylogeny of Archaea. The red stars correspond to the clusters of the Maxwell^{TM} classification. Red dots correspond to the two subtrees.

Table 1. Maxwell^{TM} classification clusters. Background color (white or orange clear) differentiates the clusters.

Name Gene or RNA	Distance to Barycenter	% Distance Total	AL-Prox	Mean AL-Prox
Lynx rufus nucleolin	0		13.9	14.66
Suncus etruscus nucleolin	623,310	24.2%	14.8	
Rhinolophus ferrumequinum nucleolin	445,316	17.3%	16.4	
Elephas maximus indicus nucleolin	569,205	22.1%	14.2	
Sciurus carolinensis nucleolin	496,040	19.2%	16.3	
Equus quagga nucleolin	392,613	15.2%	13	
Prionailurus viverrinus nucleolin	53,318	2%	14	

Table 1. *Cont.*

Name Gene or RNA	Distance to Barycenter	% Distance Total	AL-Prox	Mean AL-Prox
Halorhabdus utahensis DSM 12,940 strain DSM 12,940 5S ribosomal RNA	0		1.7	1.28
1 *Halovivax ruber* XH-70 strain XH-70 5S ribosomal RNA	571,428	19%	1.24	
Nitrosopumilus maritimus 5S	742,857	24.6%	0.32	
Sulfolobus solfataricus P2 strain 5S	734,693	24.4%	0	
Halomicrobium mukohataei DSM 12,286 5S	571,428	19%	2.8	
Halorubrum lacus profundi ATCC 49,239 strain ATCC 49,239 5S ribosomal RNA	393,939	13%	1.63	
Methanolobus psychrophilus R15 strain 5S	0		2.8	4.95
Hydrobacter penzbergensis nucleolin	981,132	33.5%	10.6	
Ogataea polymorpha strain nucleolin	969,924	33.1%	3.6	
Stackebrandtia nassauensis DSM 44,728 nucleolin	977,086	33.4%	2.8	
Archaeoglobus veneficus SNP6 strain SNP6 5S ribosomal RNA	0		0.9	1.32
Hyperthermus butylicus DSM 5456 strain DSM 5456 5S	670,103	31.6%	0	
Ferroglobus placidus DSM 10,642 strain DSM 10,642 5S ribosomal RNA	371,134	17.5%	1.54	
Candidatus Korarchaeum cryptofilum 5S	587,628	27.7%	1.7	
Archaeoglobus sulfaticallidus PM70-1 strain PM70 5S-1	190,000	9%	1.6	
Archaeoglobus profundus DSM 5631 strain DSM 5631 5S ribosomal RNA	300,000	14.2%	2.2	

The first four clusters of the classification tree successively represent Archaea (class 1), Archaea with ancient Bacteria and Fungi (class 2), Bacteria with ancient Archaea (class 3) and Mammals (class 4). This classification respects the known hierarchies of successive clades, obtained by comparing genomes of the same nature (see the Supplementary Materials Table S4 for the whole clustering) and the Maxwell^{TM} clustering with cladistic ranking can be described as a list, whose four first clusters are:

(1) Cluster 1 Archaea

Kingdom:	Archaea
Division:	Euryarchaeota
Class:	Halobacteria
Order:	Halobacteriales
Family:	Halobacteriaceae
Genus:	Halobacteriaceae halorabdus, Halovivax, Halomicrobium, and Halorubrum
Division:	Thaumarchaeota
Class:	Incertae sedis
Order:	Nitrosopumilales
Family:	Nitrosopumilaceae
Genus:	Nitrosopumilus *nitrospumilus maritimus*
Division:	Crenarchaeota
Class:	Thermoprotei
Order:	Sulfolobales
Family:	Sulfolobaceae *sulfolobus solfataricus*

(2) Cluster 2 Archaea and Bacteria

Division:	Euryarchaeota
Class:	Methanomicrobia
Order:	Methanosarcinales
Family:	Methanosarcinaceae *Methanolobus psychrophilus*
Domain:	Bacteria
Phylum:	Bacteroidota
Class:	Chitinophagia

Order:	Chitinophagales
Family:	Chitinophagaceae *Hydrobacter penzbergensis*
Kingdom:	Fungi
Division:	Ascomycota
Class:	Saccharomycetes
Order:	Saccharomycetales
Family:	Saccharomycetaceae
Genus:	Ogataea Ogataea polymorpha
Domain:	Bacteria
Phylum:	Actinomycetota
Class:	Actinomycetia
Order:	Glycomycetales
Family:	Glycomycetaceae
Genus:	Stackebrandtia stackebrandtia nassauensis

(3) Cluster 3 Bacteria and Archaea

Domain:	Bacteria
Phylum:	Bacteroidota
Class:	Chitinophagia
Order:	Chitinophagales
Family:	Chitinophagaceae *hyperthermus butylicus*
Phylum:	Euryarchaeota
Class:	Archaeoglobi
Order:	Archaeoglobales
Family:	Archaeoglobaceae ferroglobus placidus, Archaeoglobus sulfaticallidus, and Archaeoglobus profundus

(4) Cluster 4 Mammals

lynx, shrew, bat, elephant, squirrel, horse, and cat

4. Discussion

In the classification obtained using the classifier Maxwell^{TM}, there exists no information about the species, except the succession of nucleotides of some of their RNAs or mRNAS (5S ribosomal RNAs, nucleolin (NCL) and nucleolar protein (NOL11) mRNAs).

In Figure 5, the Archaea phylogeny [31] shows an organization compatible with the Maxwell^{TM} classification tree in Figure 4. In particular, all the classes marked with a red star correspond to classes of the Maxwell^{TM} tree, even though all their contents have not been systematically explored in the present study. This consistency between the classes discovered using the Maxwell^{TM} algorithm, only from the nucleotide sequence of some RNAs and the classes of an Archaea phylogeny, is an important argument validating the new Maxwell^{TM} classification method.

5. Conclusions and Perspectives

The challenging problem of finding an ancestor to RNAs related to the ribosomal protein factory can be partially solved by looking at the nucleotide sequence of some ribosomal RNAs and mRNAs of proteins involved in the building of the ribosome itself. Some invariant parts of these nucleotide sequences are detected via Maxwell^{TM} and future work will be dedicated to the classification of random sequences, using Maxwell^{TM}, respecting some evolutionary rules based on precise operators among the eleven acting in genome evolution and used via the genetic algorithms: Crossing-over, Mutation, Translocation, Insertion, Deletion, Transposition, Inversion, Repetition, Symmetrization, Palindromization, and Permutation [32,33]. This will allow us to extensively understand the hidden mechanisms of the Maxwell^{TM} algorithm in detecting common motifs in the nucleotide sequences of ribosomal and messenger RNAs. As the Maxwell^{TM} classifier mainly detects repeats, insertions, mutations and palindromizations common to multiple genomes that we wish to compare, the clustering trees obtained via it will have biological significance.

These trees will complement the classical phylogenetic trees from the primitive molecular structures of the current species in order to refine our current knowledge on evolution.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/computation11080158/s1. Table S1: List of tRNA-GlyGCC from 246 species extracted from GtRNAdB; Table S2: AL-pentamer content in nucleolin (NCL)) of species of Figure 2C. Red color represents P-pentamers, blue color corresponds to overlaps; Table S3: Examples of P-pentamer content in nucleophosmine (NPM1) of 8 species from Table S2. Red color represents P-pentamers, blue color corresponds to overlaps; Table S4: Maxwell^{TM} classification clusters.

Author Contributions: Conceptualization, J.D. and J.G.; methodology, J.D., J.G., C.M. and D.B.; K.B. and I.T. have performed the calculations; all authors have equally participated in the other steps of article elaboration. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All data are coming from public data bases and are given in Supplementary material.

Acknowledgments: The authors hereby acknowledge the support of the Orange Labs and the MIASH master (Mathematics and Informatics Applied to Human Sciences) of the University of Grenoble Alpes.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Miller, S.L. A Production of amino acids under possible primitive Earth conditions. *Science* **1953**, *117*, 528–529. [CrossRef]
2. Bada, J.L.; Lazcano, A. Prebiotic soup—Revisiting the Miller experiment. *Science* **2003**, *300*, 745–746. [CrossRef] [PubMed]
3. Damer, B.; Deamer, D. The Hot Spring Hypothesis for an Origin of Life. *Astrobiology* **2020**, *20*, 429–452. [CrossRef] [PubMed]
4. Katchalsky, A. Prebiotic synthesis of biopolymers on inorganic templates. *Naturwiss* **1973**, *60*, 215–220. [CrossRef]
5. Martin, W.; Russell, M.J. On the origin of biochemistry at an alkaline hydrothermal vent. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2007**, *362*, 1887–1925. [CrossRef] [PubMed]
6. Deamer, D. The Role of Lipid Membranes in Life's Origin. *Life* **2017**, *7*, 5. [CrossRef]
7. Turk-MacLeod, R.M.; Puthenvedu, D.; Majerfeld, I.; Yarus, M. The Plausibility of RNA-Templated Peptides: Simultaneous RNA Affinity for Adjacent Peptide Side Chains. *J. Mol. Evol.* **2012**, *74*, 217–225. [CrossRef]
8. Xiao, H.; Murakami, H.; Suga, H.; Ferré-D'Amaré, A.R. Structural basis of specific tRNA aminoacylation by a small in vitro selected ribozyme. *Nature* **2008**, *454*, 358–361. [CrossRef]
9. Deng, J.; Wilson, T.J.; Wang, J.; Peng, X.; Li, M.; Lin, X.; Liao, W.; Lilley, D.M.J.; Huang, L. Structure and mechanism of a methyltransferase ribozyme. *Nat. Chem. Biol.* **2022**, *18*, 556–564. [CrossRef] [PubMed]
10. Grum-Tokars, V.; Milovanovic, M.; Wedekind, J.E. Crystallization and X-ray diffraction analysis of an all-RNA U39C mutant of the minimal hairpin ribozyme. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2003**, *59*, 142–145. [CrossRef] [PubMed]
11. Demongeot, J. *Au Sujet de Quelques Modèles Stochastiques Appliqués à la Biologie. Modélisation et Simulation*; Université Joseph-Fourier: Grenoble, France, 1975.
12. Demongeot, J. Sur la possibilité de considérer le code génétique comme un code à enchaînement. *Rev. Biomaths* **1978**, *62*, 61–66.
13. Demongeot, J.; Besson, J. Code génétique et codes à enchaînement. *C. R. Seances L'Acad. Sci. Ser. III* **1983**, *296*, 807–810.
14. GtRNAdB. Available online: http://gtrnadb.ucsc.edu/ (accessed on 23 May 2023).
15. Demongeot, J.; Moreira, A. A circular RNA at the origin of life. *J. Theor. Biol.* **2007**, *249*, 314–324. [CrossRef]
16. Hobish, M.K.; Wickramasinghe, N.S.M.D.; Ponnamperuma, C. Direct interaction between amino-acids and nucleotides as a possible physico-chemical basis for the origin of the genetic code. *Adv. Space Res.* **1995**, *15*, 365–375. [CrossRef]
17. Tamura, K.; Schimmel, P. Oligonucleotide-directed peptide synthesis in a ribosome- and ribozyme-free system. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 1393–1397. [CrossRef]
18. Paecht-Horowitz, M.; Berger, J.; Katchalsky, A. Prebiotic synthesis of polypeptides by heterogeneous polycondensation of amino-acid adenylates. *Nature* **1970**, *228*, 636–639. [CrossRef]
19. Eigen, M. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **1971**, *58*, 465–523. [CrossRef]
20. Gilbert, W. Origin of life: The RNA world. *Nature* **1986**, *319*, 618. [CrossRef]
21. Kauffman, S.A. Approaches to the origin of life on Earth. *Life* **2011**, *1*, 34–48. [CrossRef]
22. NCBI. Available online: https://www.ncbi.nlm.nih.gov/refseq/ (accessed on 23 May 2023).
23. Edous, M.; Eidous, O. A Simple Approximation for Normal Distribution Function. *Math. Stat.* **2018**, *6*, 47–49. [CrossRef]
24. Gardes, J.; Maldivi, C.; Boisset, D.; Aubourg, T.; Vuillerme, N.; Demongeot, J. Maxwell®: An unsupervised learning approach for 5P medicine. *Stud. Health Technol. Inform.* **2019**, *264*, 1464–1465.

25. Burrows, M.; Wheeler, D.J. A block-sorting lossless data compression algorithm. *Digit. SRC Res. Rep.* **1994**, *124*, 10009821328.
26. Cilibrasi, R.; Vitanyi, P.M.B. Clustering by compression. *IEEE Trans. Inf. Theory* **2005**, *51*, 1523–1545. [CrossRef]
27. Cohen, A.R.; Vitányi, P.M.B. Normalized Compression Distance of Multisets with Applications. *IEEE Trans. PAMI* **2015**, *37*, 1602–1614. [CrossRef] [PubMed]
28. Graphviz. Available online: https://graphviz.org/ (accessed on 23 May 2023).
29. Woese, C.; Fox, G. The concept of cellular evolution. *J. Mol. Evol.* **1977**, *10*, 1–6. [CrossRef] [PubMed]
30. Gogarten, J.P.; Deamer, D. Is LUCA a thermophilic progenote? *Nat. Microbiol.* **2016**, *1*, 16229. [CrossRef]
31. Adam, P.S.; Borrel, G.; Brochier-Armanet, C.; Gribaldo, S. The growing tree of Archaea: New perspectives on their diversity, evolution and ecology. *ISME J.* **2017**, *11*, 2407–2425. [CrossRef]
32. Schmitt, L.M. Theory of Genetic Algorithms. *Theor. Comput. Sci.* **2001**, *259*, 1–61. [CrossRef]
33. Ighalo, J.O.; Marques, G. *Current Trends and Advances in Computer-Aided Intelligent Environmental Data Engineering*; Elsevier: Amsterdam, The Netherlands, 2022.