International Journal of
Molecular Sciences

MDPI

Article

An Unsupervised Classifier for Whole-Genome Phylogenies, the Maxwell© Tool

Joël Gardes [1], Christophe Maldivi [1], Denis Boisset [1], Timothée Aubourg [2] and Jacques Demongeot [2,*]

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

Abstract: The development of phylogenetic trees based on RNA or DNA sequences generally requires a precise and limited choice of important RNAs, e.g., messenger RNAs of essential proteins or ribosomal RNAs (like 16S), but rarely complete genomes, making it possible to explain evolution and speciation. In this article, we propose revisiting a classic phylogeny of archaea from only the information on the succession of nucleotides of their entire genome. For this purpose, we use a new tool, the unsupervised classifier Maxwell, whose principle lies in the Burrows–Wheeler compression transform, and we show its efficiency in clustering whole archaeal genomes.

Keywords: unsupervised classifier; maxwell classifier; Burrows–Wheeler compression transform; normalized compression distance (NCD); Vitányi distance; phylogenetic trees

Citation: Gardes, J.; Maldivi, C.; Boisset, D.; Aubourg, T.; Demongeot, J. An Unsupervised Classifier for Whole-Genome Phylogenies, the Maxwell© Tool. *Int. J. Mol. Sci.* **2023**, *24*, 16278. https://doi.org/10.3390/ijms242216278

Academic Editors: Hari Shanker Sharma and Kunio Takeyasu

Received: 30 September 2023
Revised: 20 October 2023
Accepted: 2 November 2023
Published: 13 November 2023

1. Introduction

There are numerous algorithms of classification, supervised or otherwise. In the set of supervised algorithms, we can find several types of classifiers, like k-means (k-nearest neighbors [1–3]) and SVM (support vector machine [4,5]) methods, and examples include neural networks with the Hopfield [6–8] or Boltzmann [9] approach and deep learning [10] for the unsupervised type (see also Supplementary Material Table S1 for other classifiers). The Maxwell Algorithm of Clustering (MAC) belongs to the last category.

This MAC classification tool uses a totally reversible compression method (which is particularly suitable for the health sector), and applications have already been undertaken in genetics, for problems of storage or sequence recognition [11]. Here, the challenge is different: it consists of classifying nucleotide sequences of complete genomes, of distinct size and species, without any prior indication of co-evolution, to see if it is possible to obtain clusters in agreement with the existing phylogenies, which are, in general, based on specific RNAs (such as 16S ribosomal RNA) or proteins important for the biosynthesis or degradation of RNAs (such as RNase P1).

In Section 2, we will present the methodology related to the Burrows–Wheeler transform and Vitányi distance used in different MAC steps. Then, in Section 3, we describe some genomic applications, and in Sections 4 and 5, we discuss these results and we conclude by opening some perspectives for future works.

2. Biological Context

2.1. Gene Translation for Building Proteins

The tRNA-Gly^{GCC} secondary structures in different domains (eukaryotes, archaea and bacteria) contain invariant loops: D-loop, Anti-codon loop, Tψ-loop and articulation, whose sequence called AL (archetypal loop) is AAUGGUACUTGCCAUUCAAGAUG. AL is compatible with the loop sequence of tRNA-Gly^{GCC} of *Lupinus albus* (Figure 1).

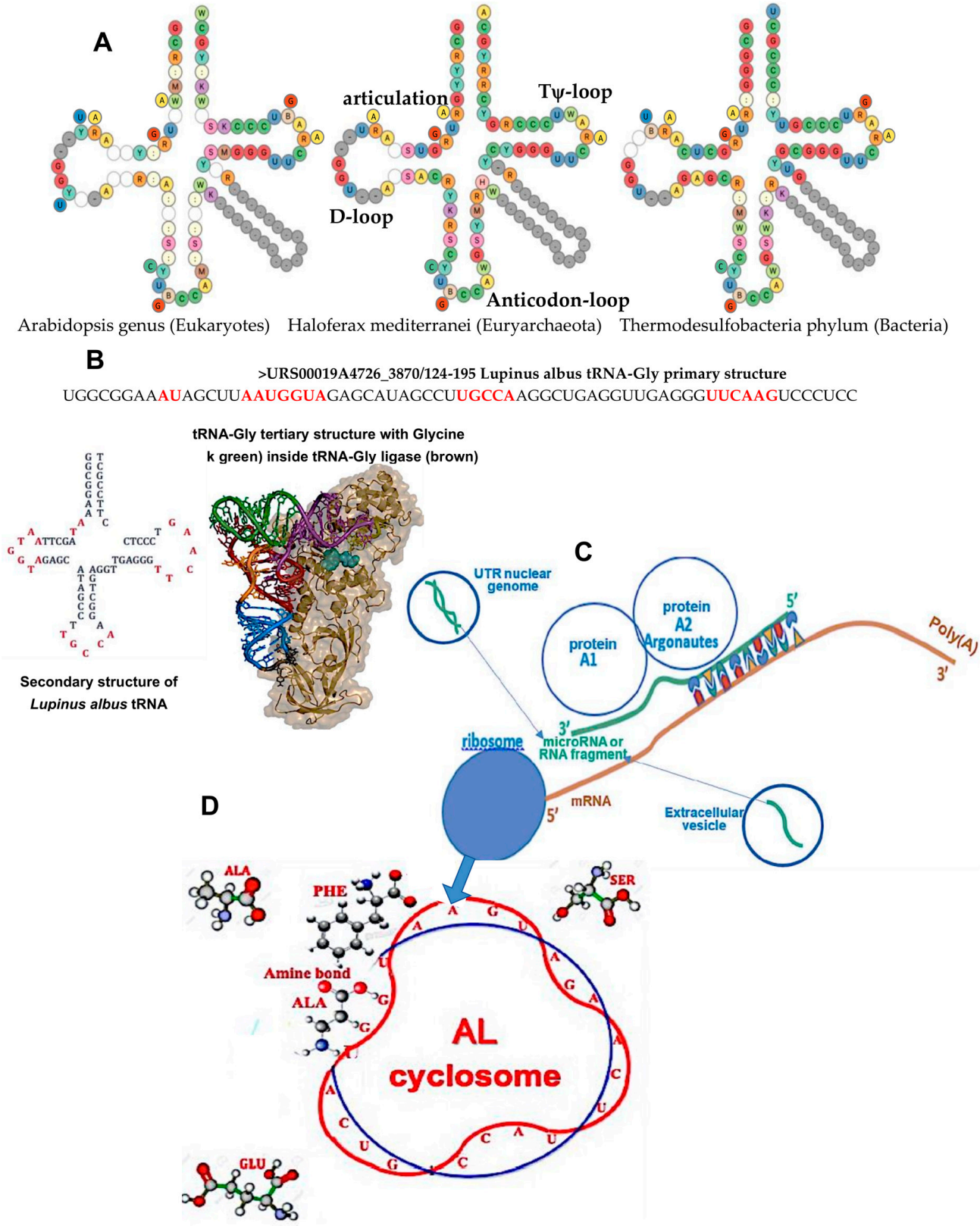

Figure 1. (**A**) tRNA-Gly^{GCC} secondary structure of eukaryotes (Arabidopsis), archaea (Haloferax) and bacteria (Thermodesulfobacterium) [12]. (**B**) The primary, secondary and tertiary structures of the tARN-Gly^{GCC} of *Lupinus albus* with its D-loop, Anti-codon loop, Ty-loop and articulation (in red). (**C**) Translation of an mRNA in the ribosome compared to (**D**) building of the peptidic dimer Phe-Ala catalyzed by the archetypal loop AL.

2.2. Combinatorial Properties of AL

The lupine tRNA-Gly^{GCC} of Figure 1B has the same loops as 1200 tRNAs of different species extracted from the tRNAviz database (cf. [12] and Supplementary Material Table S2) containing the same AL sequence of nucleotides. This remarkable invariance of tRNA loops had already been noted by Manfred Eigen [13] and is related to the following variational problem [14]. Is there a circular RNA of minimum length comprising codons from the 20 synonymy classes of the genetic code, with a possible overlap? The answer is negative for lengths 20 and 21. For length 22, there are 29,520 solutions among the 10^{22} possible, if we authorize a supplementary codon from the STOP class (Figure 2). Of these 29,520 solutions, 1280 start with AUG, end with a STOP codon and have a codon in all 20 amino acid classes, including twice AUG (Rule R). If we look for solutions having, in addition to the circular shape, a second hairpin shape of maximum enthalpy (requiring a complementarity of two halves of the loop), we find 25 solutions (Rule R*), including 10 from the 1280 previous solutions (Rule R**), and their barycenter is the sequence, which has either a functional circular form (capable of catalyzing peptide synthesis) or a stable hairpin form, with the two coexisting in equilibrium (Figure 3 and Supplementary Material Figure S1).

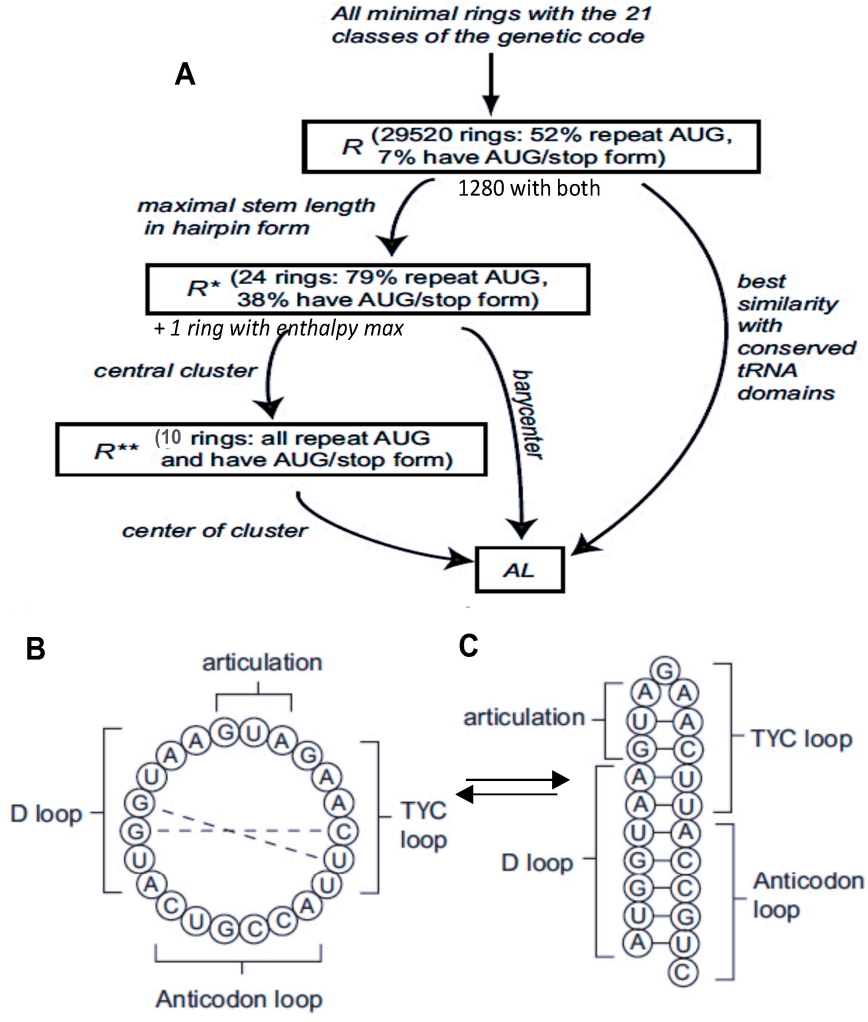

Figure 2. (**A**) Schematic summary of the search for the sequence AL, where rule R is "twice AUG", rule R* is "existence of a hairpin shape of maximum enthalpy (requiring a complementarity of two halves of the loop" and R** is "repeat AUG and have AUG-stop form"; (**B**) circular shape of AL; (**C**) hairpin shape of AL.

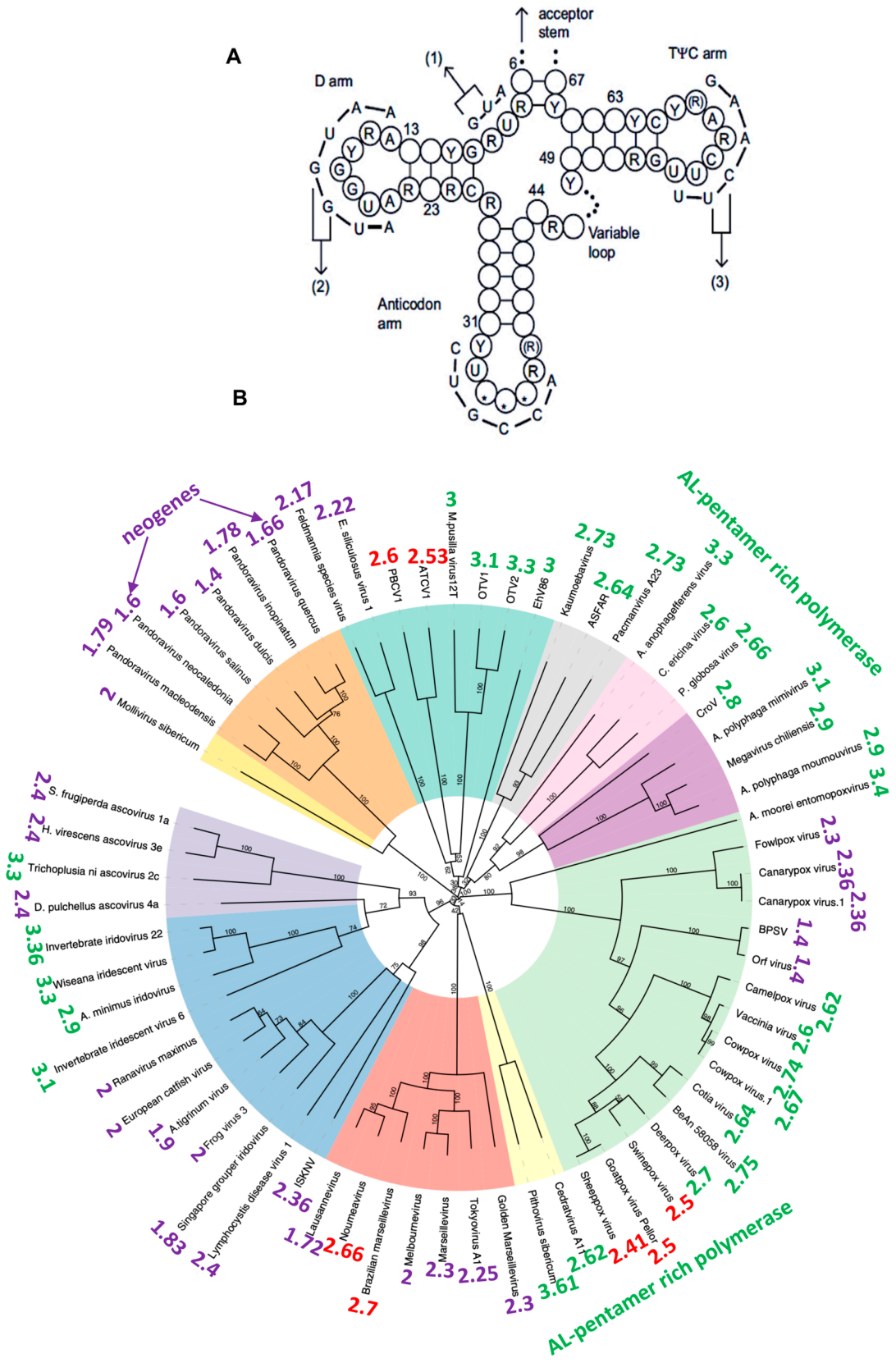

Figure 3. (**A**) tRNA general architecture with AL matching on the loops; (**B**) giant virus phylogeny with indication of their AL-proximity.

2.3. Traces of the Archetypal Loop AL in the Present Genomes

If AL could have played a role in the construction of the first peptides, analogous to the role of the ribosome in current protein synthesis, we should find traces of it in the RNAs linked to ribosomal function. We have already seen its involvement, through four fragments, in the loops of the tRNA-Gly linked to the GCC anticodon. More generally, we find the succession of AL nucleotides (or of their family, with Y designating puric acids C and U, and R pyrimidine acids A and G) in the loops of numerous tRNAs (see Supplementary Material Table S2) and in numerous genomes of giant viruses [15] or viroids [16] (Figure 3).

The proximity to AL (or AL proximity) noted in Figure 3 around the circular tree of giant viruses is calculated in the following way: we count the number O of common pentamers between the nine pentamers of the head of the hairpin form of AL, the easiest to fragment, called AL-pentamers {ATTCA, TTCAA, TCAAG, CAAGA, AAGAT, AGATG, GATGA, AATGA, ATGAA, TGAAT} and those of the studied RNA sequence of length n. Then, we calculate the expected number E, equal to the possible number P = n-4, multiplied by the probability p = 9/1024 of observing one of the nine AL-pentamers. The standard deviation of E is equal to $\sigma = (Pp(1-p))^{1/2}$, and the AL-proximity equals $(O-E)/\sigma$. Figure 4A shows such a calculation for the human nucleophosmin 1 mRNA, involved in the construction of the ribosome. Its AL-proximity (14.5σ) is very high, because the probability of such a difference between observed and expected pentamers is of the order of 10^{-14} [17–21]. Figure 4B shows an identical calculation for a proximity between pairs of the sequence of amino acids corresponding to the succession of AL codons at a distance of at most 12 nucleotides from a given codon (called AL-pairs) and pairs of successive amino acids in the studied protein. The number of AL-pairs is significant (the probability of the difference between observed and expected pairs equal to 10^{-3}) and the most frequent pairs are 9-12 nucleotides apart, corresponding to an efficient catalysis of the di-peptide biosynthesis by AL (Figure 4C).

The abundance of traces of AL in nucleophosmin 1 is not limited to human species, and Figure 5 shows 20 species of eukaryotes having a nucleophosmin 1 mRNA close to AL, in terms of their content of AL-pentamers. The same is true for many RNAs and proteins (like nucleolin, see Supplementary Material Table S3) involved in the construction and functioning of the ribosome in its current version in eukaryotes, reinforcing the hypothesis of an ancient protein construction mechanism involving AL.

2.4. Ribosomal Proteins and rRNA Components of the Current Ribosomes

In order to strengthen the hypothesis that the AL ring is an ancient structure, we will calculate the AL-proximity of the ribosomal proteins (RPs) ordered following their anteriority. In Figure 6, the mean M_o (resp. M_n) of AL-pentamer proximities for the 11 most ancient (respectively, recent) ribosomal proteins (after Gustavo Caetano-Anollés in [24]) is equal to 3.81 (respectively, 2.4). By applying a *t*-test of comparison of means, the 11 most ancient ribosomal proteins are closer to AL than the 11 earliest ones ($p = 0.0001$). This result is coherent with the hypothesis according to which AL is an ancient structure having had a role similar to that played by the current ribosome in multicellular eukaryotic species such as Homo sapiens and unicellular species such as Saccharomyces cerevisiae or in unicellular archaea such as Marine Group I thaumarchaeote YK1309.

A

Homo sapiens mRNA nucleophosmin 1 (NMP1), clone: FLJ08034AAAF GenBank: AB451361.1

ATGG**AAGAT**TCGATGGACATGGACATGAGCCCCCTGAGGCCCCAGAACTATCTTTTCGGTTGTGAACTAA

AGGCCGACA**AAGAT**TATCACTTTAAGGTGGATAAT**GATGAA**AATGAGCACCAGTTATCTTTAAGAACGGT

CAGTTTAGGGGCTGGTGCAAAG**GATGA**GTTGCACATTGTTGAAGCAGAGGCA**ATGAAT**TACGAAGGCAGT

CCAATTAAAGTAACACTGGCAACTTTGAAAATGTCTGTACAGCCAACGGTTTCCCTTGGGGGCTTTGAAA

TAACACCACCAGTGGTCTTAAGGTTGAAGTGTGGTTCAGGGCCAGTGCATATTAGTGGACAGCACTTAGT

AGCTGTGGAGG**AAGATG**CAGAGTCAG**AAGATGAA**GAGGAGGAGGATGTGAAACTCTTAAGTATATCTGGA

AAGCGGTCTGCCCCTGGAGGTGGTAGCAAGGTTCCACAGAAAAAAGTAAAACTTGCTGCT**GATGAAGATG**

ATGAC**GATGATGATGAA**GAG**GATGATGATGAAGATGATGATGATGATGA**TTTTGATAATGAGGAAGCTGA

AGAAAAAGCGCCAGTGAAGAAATCTATACGAGATACTCCAGCCAAAAATGCACAAAAGTCAAATCAGAAT

GGAAAAGACTCAAAACCATCATCAACAC**CAAGAT**CAAAAGGA**CAAGA**ATCC**TTCAAGA**AACAGGAAAAAA

CTCCTAAAACACCAAAAGGACCTAGTTCTGTAGAAGACATTAAAGCAAAAATGCAAGCAAGTATAGAAAA

AGGTGGTTCTCTTCCCAAAGTGGAAGCCAA**ATTCA**TCAATTATGTGAAGAATTGCTTCCG**GATGA**CTGAC

CAAGAGGCT**ATTCAAGAT**CTCTGGCAGTGGAGGAAGTCTCTT

Nb observed AL-pentamères O = 48, possible P = 878, expected E = 7.717 (σ=2.78), **O-E = 14.5σ**

B

Nucleophosmin isoform 1 [Homo sapiens] NCBI Reference Sequence: NP_002511.1

MEDSMDMDMSPLRPQNYLFGCELKADKDYHFKVDNDENEHQLSLRTVSLGAGAKDELHIVEAEAMNYEGS

PIKVTLATLKMSVQPTVSLGGFEITPPVVLRLKCGSGPVHISGQHLVAVEEDAESEDEEEEDVKLLSISG

KRSAPGGGSKVPQKKVKLAADEDDDDDDEEDDDEDDDDDDFDDEEAEEKAPVKKSIRDTPAKNAQKSNQN

GKDSKPSSTPRSKGQESFKKQEKTPKTPKGPSSVEDIKAKMQASIEKGGSLPKVEAKFINYVKNCFRMTD

QEAIQDLWQWRKSL

Nb observed AL-pairs O = 172, possible P = 293, expected E = 138.4 (σ=11.7), **O-E = 2.9σ**

C

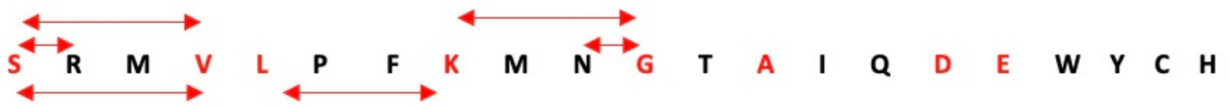

Figure 4. (**A**) AL-proximity of human nucleophosmin 1. (**B**) Same calculation for AL-pairs. (**C**) Most frequent AL-pairs observed (linked by two red arrows) and comparison with the oldest amino acids in the genetic code according to Koonin's [22] and Trifonov's [23] orders.

20 Species with nucleophosmin 1 mRNA	AL-proxy
Monodelphis domestica nucleophosmin 1 (NPM1), chr. 1, MonDom5, mRNA NCBI Sequence: NC_008801.1	16.4σ
Rattus norvegicus nucleophosmin 1 (NPM1), mRNA NCBI Sequence: NM_012992.4	15.5σ
Mytilus coruscus strain nucleophosmin 1 (NPM1), contig: Mco4455, GenBank: CACVKT020004326.1	15.4σ
Bos taurus nucleophosmin 1 (NPM1), NPM1-GG allele, exon 1 and partial cds, GenBank: GQ144334.1	14.7σ
Homo sapiens NPM1 nucleophosmin 1 isoform 1 (NPM1), partial cds, clone: FLJ08034AAAF GenBank: AB451361.1	14.5σ
Mus musculus nucleophosmin 1 (NPM1), cDNA clone RZPDo836E0452D for gene Npm1, GenBank: CT010327.1	14σ
Jaculus jaculus nucleophosmin 1 (NPM1), transcript variant X2, mRNA NCBI Reference Sequence: XM_004664999.2	13.4σ
Xenopus tropicalis nucleophosmin 1 (NPM1), mRNA NCBI Sequence : NM_20355.1	13.1σ
Lipotes vexillifer nucleophosmin 1 (NPM1) LOC103076865, misc_RNA NCBI Sequence: XR_456924.1	12.9σ
Bauhinia variegate nucleophosmin 1 (NPM1), isolate BV-YZ2020 chromosome 9, GenBank: JAKRYI020000009.1	12.2σ
Ictidomys tridecemlineatus nucleophosmin 1 (NPM1), isolate GS200 Itri18, GenBank: JAESOR010000030.1	12.1σ
Phodopus roborovskii nucleophosmin 1 (NPM1), contig: tig00001838, GenBank: CALSGD010001391.1	11.5σ
Eptesicus fuscus nucleophosmin 1 (NPM1), variant X2, mRNA NCBI Sequence: XM_054718139.1	11.4σ
Pan troglodytes nucleophosmin 1 (NPM1), variant X5, mRNA NCBI Sequence: XM_054684087.1	11.2σ
Pongo abelii nucleophosmin 1 (NPM1), variant X4, mRNA NCBI Sequence: XM_024246637.2	11.2σ
Agelaius phoeniceus nucleophosmin 1 (NPM1), variant X3, mRNA NCBI Sequence : XM_054642943.1	11.2σ
Capra hircus isolate 0256 nucleophosmin 1 (NPM1), partial cds, GenBank: HM006820.1	10.7σ
Mirounga angustirostris nucleophosmin 1 (NPM1), variant X2, mRNA NCBI Sequence: XM_045901451.2	10.3σ
Pteronotus parnellii mesoamericanus nucleophosmin 1 (NPM1), variant X2, mRNA NCBI Sequence : XM_054568056.1	10.3σ
Cervus elaphus hippelaphus nucleophosmin 1 (NPM1), isolate Hungarian chr. 25, GenBank: MKHE01000025.1	9.7σ

Figure 5. AL-proximity of nucleophosmin 1 (NPM1) mRNAs in 20 species of eukaryotes.

HS	SC	MT	RP
3.1	2.7	2	S15
3.9	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	1.7	1.4	L11
2.2	7.5	0.8	S10
0	2	3	L13
5.1	3.3	4.2	L14
2.7	3.2	1.9	S19
3.3	2.8	5.3	S2
4	4	1.2	L5
1	6.9	2.5	S8
2.2	4.2	1.7	S7
3.6	7.8	1.8	S3
1	1.2	1.8	L29
1.7	3.8	0.8	L7/L12
1	7.4	7	S5
7.6	4.6	2.9	S11
3.1	7	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	1.7	L24
2.5	3.6	4.9	L2
4.4	5	5.2	L3
3.2	6.4	2.7	S9
3.7	4	3.7	S17
5.6	5.8	7.3	S12

Figure 6. AL-proximity of ribosomal RNA 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 oldest (bottom) during evolution [24].

3. Results

3.1. Phylogenetic Tree of Archaea

From the Archaeota phylum, 85 species ([25,26]) were classified through a classic phylogenetic tree using the IQTree algorithm by comparing 41 genes, including 3 coding for mitochondrial proteins, 2 for RNA-polymerase sub-units and 36 belonging to a gene list involved in the ribosomal architecture (S2, S3, S5, S7, S8, S9, S10, S11, S12/S23, S13e, S15, S19, S17, L1, L2, L3, L4/L1, L5, L6, L10, L11, L13, L14b/L23e, L15, L16/L10e, L18/L5e, L22, L24, L25/L23, L29). We added to this tree values (in red) of the AL-proximity of the entire genomes of the species concerned. These values are often in agreement with seniority in the tree. This is the case for the example of the Euryarchaeota phylum, which generally has a high value of AL-proximity (see Figure 7).

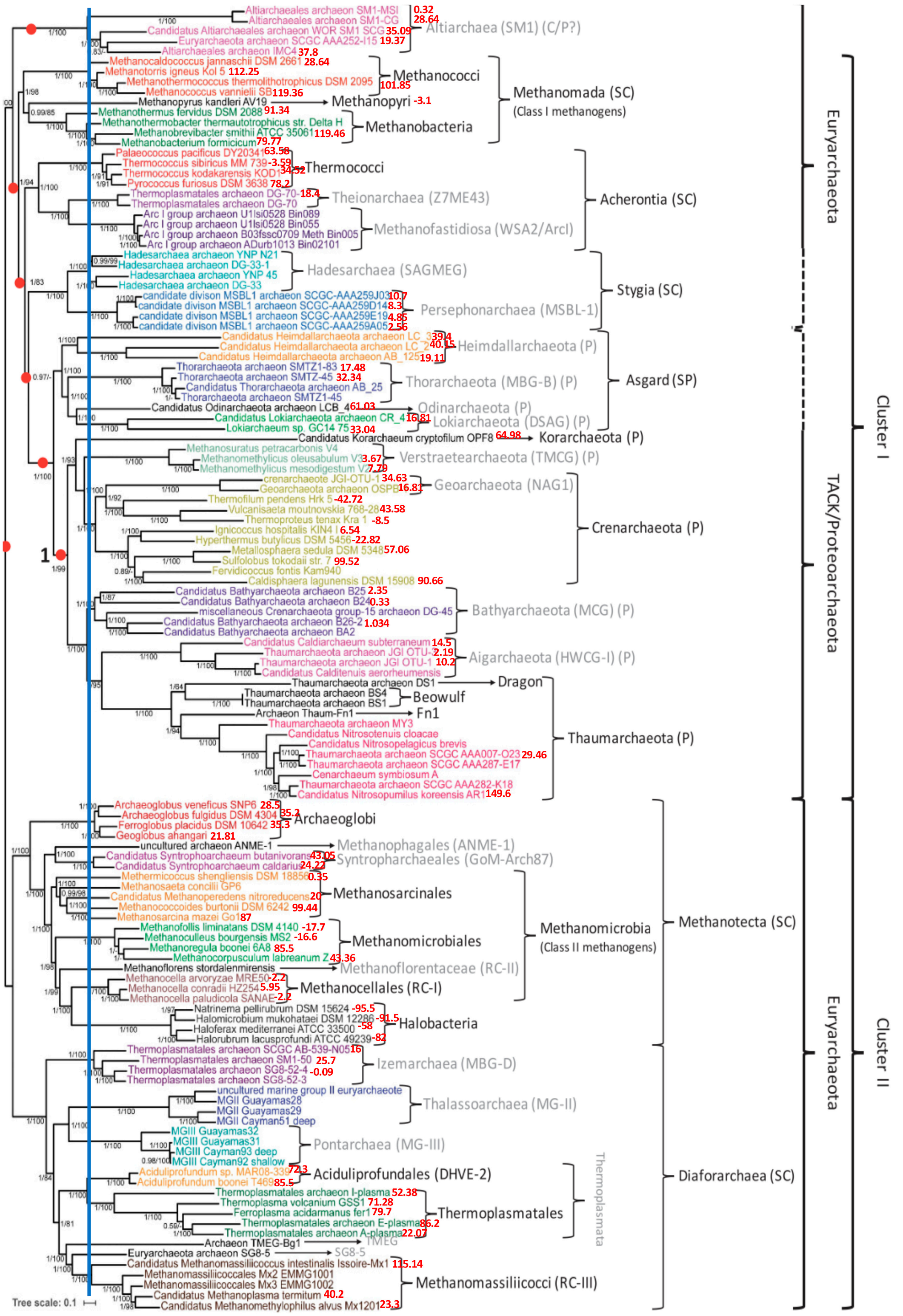

Figure 7. Phylogeny of 85 species of archaea obtained in [25] from 41 genes representing 8710 amino acid positions. The tree scale bar corresponds to the average number of substitutions per site, and node fractions refer to posterior probabilities calculated using the IQTree algorithm. The 41 genes consist of 36 genes from the Phylosift marker gene list, 2 RNA polymerase subunits A and B, and 3 universal ribosomal proteins (L7-L12, L30, S4). The taxonomic status of species families follows the following coding: C = class; P = phylum; SC = super class; SP = super phylum. The blue vertical line corresponds to the start node of both the Altiarchaea and Methanomassiliicocci sub-trees.

On the other hand, certain species, such as Halobacteria, have negative values of their AL-proximity. This may be due to the presence of viral infections in these species, which have caused the negative distance of their genome to AL, proposed as the initial sequence [27].

3.2. Classification of Archaea by Maxwell

The classifier Maxwell is capable of processing the entire genomes of the 85 species of archaea from [25] in a few seconds. Only a small group of the clusters in Figure 8 corresponds to the phylum Altarchaeota, the rest being given in the Supplementary Material Table S4.

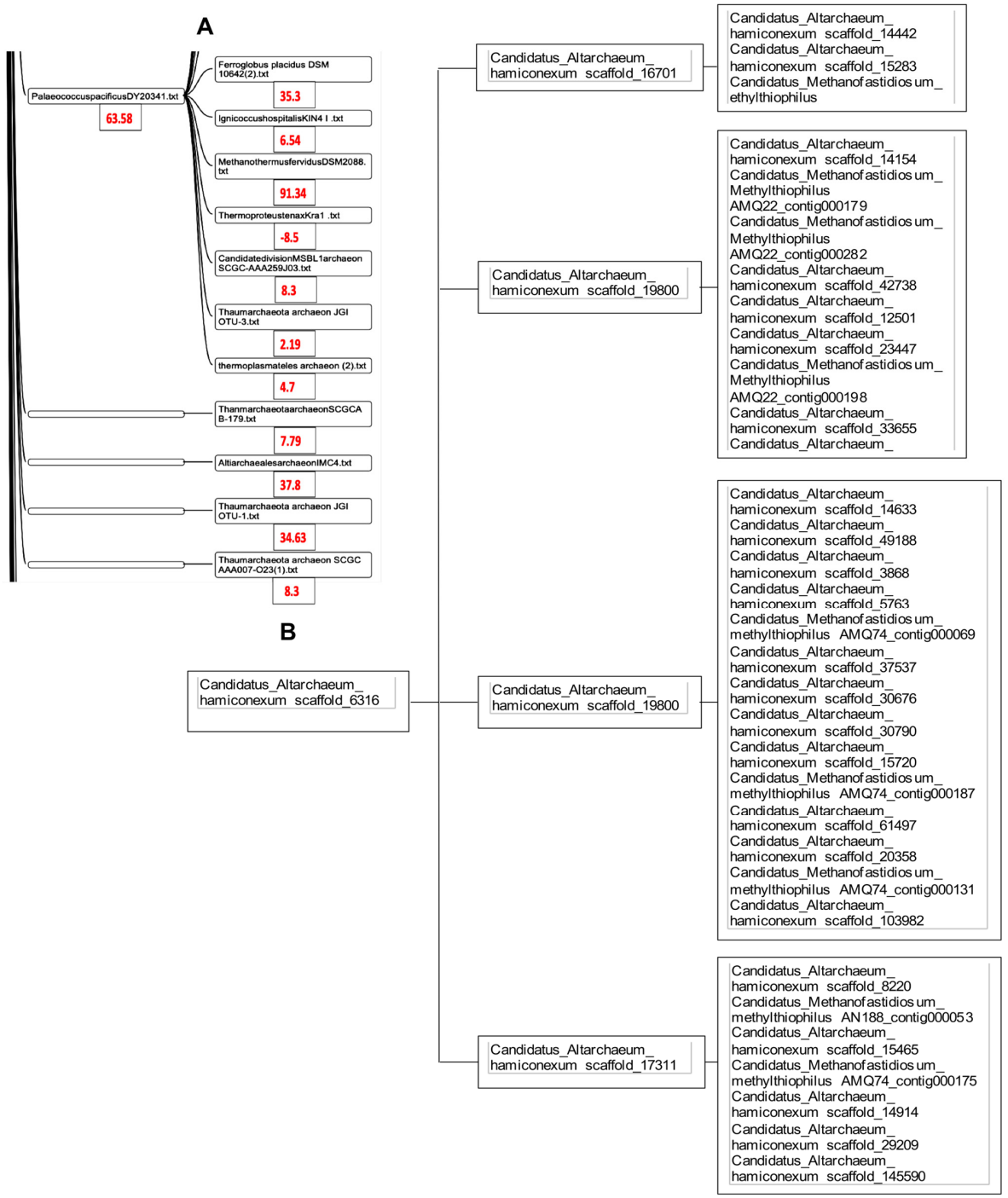

Figure 8. (**A**) Phylogenetic tree obtained using Maxwell© from complete genomes of the indicated species. (**B**) Part of this tree concerning the phylum Altarchaeota.

4. Discussion

We added to the tree in Figures 7 and 8 the values (in red) of the AL-proximity of the entire genomes of the species concerned. We see that often, these values are in agreement with the ancient character of species in the tree. This is the case, for example, for the super class Methanomada (at the top of Figure 7), considered as one of the most ancient classes of archaea, whose species have a high value of AL-proximity. On the other hand, we have already remarked that certain species, such as Halobacteria or Methanomicrobiales, have negative values of AL-proximity. This may be due to the frequent viral infections in these species, distancing their genomes from the initial RNA ring AL [27,28]. For example, among the numerous viruses infecting these extremely halophilic archaea, there are many morphotypes (round, caudo, pleomorphic and spindle-shaped viruses) with a genome far from AL. This is, for example, the case for the Halogeometricum pleomorphic virus, whose AL-proximity is negative (-3.7σ) and which was thus able to contribute, via insertion of all or part of its genome, to the negativity of the Halobacteria genomes.

Regarding the phylogenetic tree obtained using Maxwell©, we see in Figure 8 that this classifier has brought together the different genomes of Altarchaeum hamiconexum (from the phylum Altarchaeota) with the genome of Methanofastidiosum methylthiophilus (from the class Methanomassiliicocci) well, which are on opposite sides of the phylogenetic tree in Figure 7. The reason is simple: these two classes are at the same distance from the tree root, because the same blue vertical line in Figure 7 corresponds to the start nodes of both their classes, Altiarchaeota and Methanomassiliicocci. There are numerous other phylogenetic trees for the archaea domain [29–32]. They have all the same general architecture analog to those in Figure 7. The coherence with the Maxwell© architecture is not complete, but Figure 9 shows both an adequation of the AL-proximity order with the classification resulting from the phylogenetic tree given in [30] and a partial coherence with the Maxwell© tree obtained by using only the raw sequence of the nucleotides of the complete genome of the concerned species from [25].

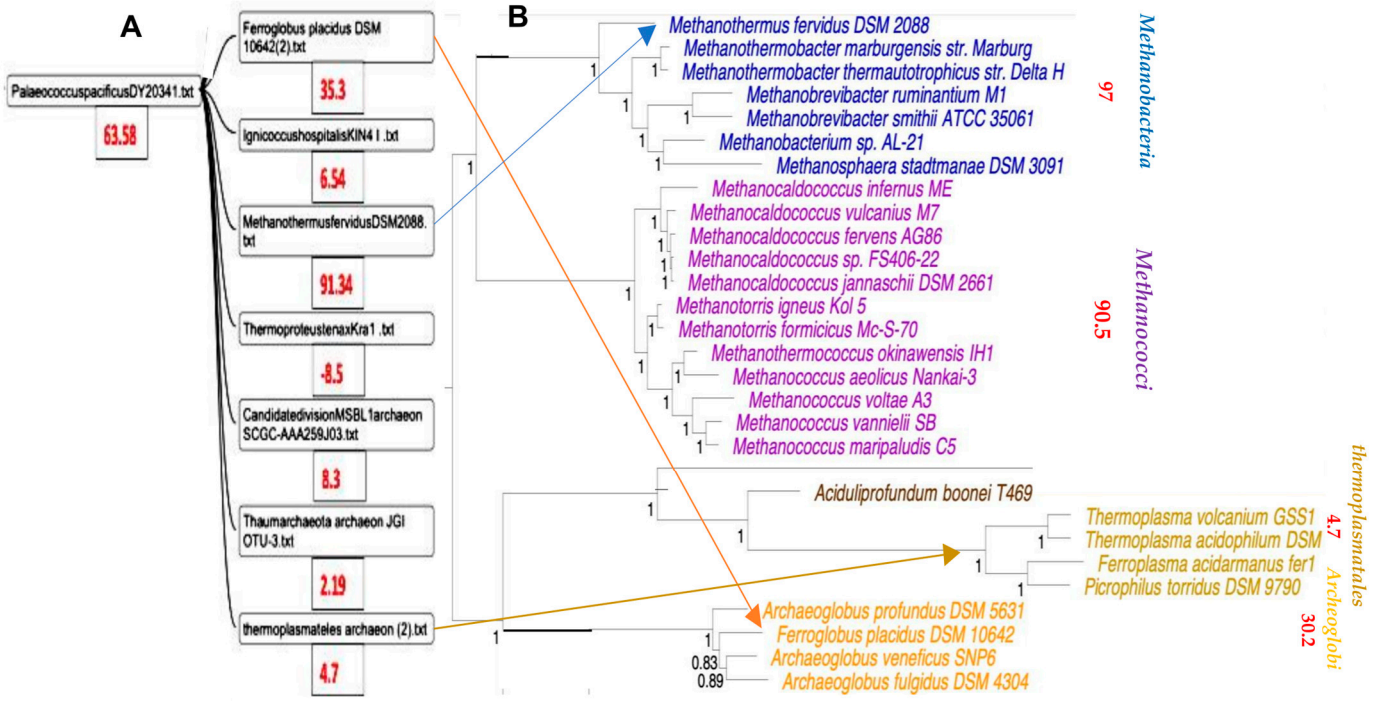

Figure 9. (**A**) Phylogenetic tree by Maxwell©. (**B**) Part of the archaea phylogenetic tree from [30] with mean AL-proximities (in red).

Numerous such examples of rapprochement in the same cluster can be found in Supplementary Material Table S4, and we will continue to multiply the species, by looking for the traces of AL in even more species of archaea, bacteria and eukaryotes, in order to strengthen the hypothesis of the "emergence of an RNA world, defined by RNA molecules with catalytic and replicative properties" [32] already published by Manfred Eigen [13].

5. Methodology

5.1. The Burrows–Wheeler Transform

The Burrows–Wheeler transform (BWT) is a lossless compression algorithm that rearranges strings into runs of similar characters in a reversible way [33,34]. When combined with a run-length encoding (RLE) algorithm, it yields a function that can be used in the computation of "Normalized Compression Distance" (NCD) or Vitányi distance, enabling the detection of similarities between information sequences such as repeated motifs, common deletion or insertions and other evolutionary patterns. BWT can be considered as a simplified compression algorithm with regard to more sophisticated ones commonly used in information theory applications, which can be lossless [35] or lossy [36]. Its main advantage over them is mathematically retrieving the symmetry of NCD. Such a property is essential to compare the genomic sequences of multiple species having coevolved under the action of the same operators. In the context of evolution, this indeed allows us to consider all of the eleven different genomic established operators, namely crossing over, mutation, translocation, insertion, deletion, transposition, inversion, repetition, symmetrization, palindromization and circular permutation. When these operators have been used with the same frequency during evolution, the Burrows–Wheeler transform is useful to compress the sequences of nucleotides coming from the same origin and having a similar evolutionary history.

5.2. Principles of the Burrows–Wheeler Transform (BWT)

The BWT algorithm transforms a string S of n characters by considering its n circular permutations (cyclic shifts), listing them lexicographically and retaining only the last character of each permutation. A final string F = BWT(S) is formed from these characters, where the ith character of F is the last character of the ith permutation. In addition to F, the BWT algorithm computes the rank s of the original string S in the list of permutations, and it is possible to integrally reconstruct S from F plus rank s. The run-length encoding of the string F, denoted RLE(F), is eventually obtained by replacing any sub-sequence of F repeating the same character, like TTT, by the number of repetitions followed by the repeated character, here 3T. Hence, the format of the final expression of the RLE(BWT(S)) is made first of the rank s, then of the characters of BWT(S) with an ultimate simplification: if a character C is met n(C) consecutive times, the final sequence is, after s, made of the succession of the characters C of BWT(S) preceded by their number n(C) of repetitions.

5.3. "Normalized Compression Distance" (NCD) or Vitányi Distance

The Vitányi distance is a measure used to quantify the similarity between two sequences, x and y, based on their compressed representation [37,38]. Information compression is used here as a proxy for similarity. The intuition that stands behind this approach is that similar sequences can be compressed with the same efficiency, in contrast to dissimilar ones. In the context of Maxwell© functioning, this consists of calculating for x, y and the concatenated word xy the length of the RLE version of their Burrows–Wheeler transform (BWT), that is, respectively, if Len[string] denotes the string length, the values of the coefficients Cx = Len[RLE(BWT(x)], Cy = Len[RLE(BWT(y)] and Cxy = Len[RLE(BWT(xy)], and then calculating the ratio (Figure 10):

$$d(x,y) = [Cxy - \min(Cx,Cy)]/\max(Cx,Cy) \quad (1)$$

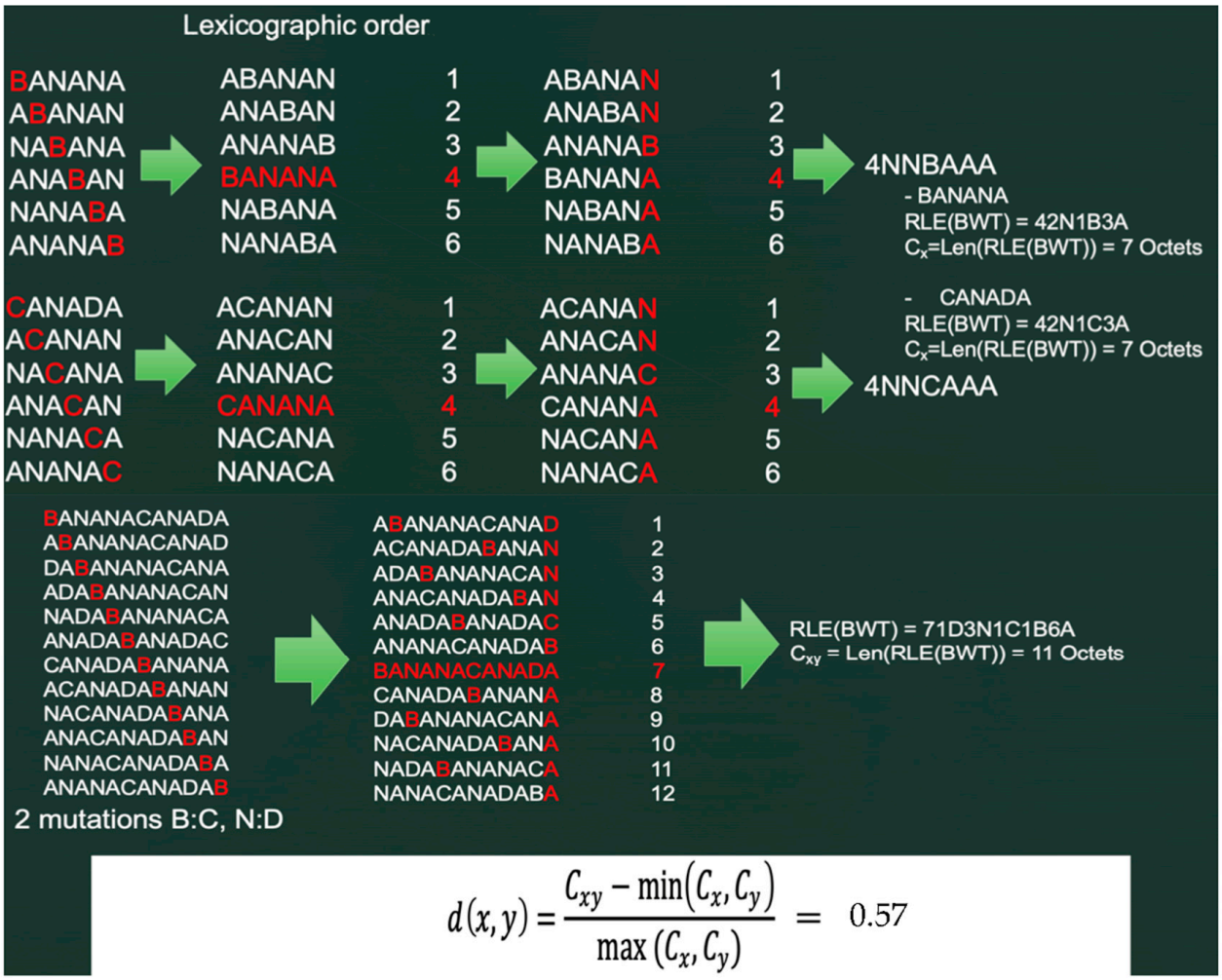

Figure 10. Burrows–Wheeler transform (BWT) of two words BANANA and CANADA, with two mutations B:C and N:D. Run-lengths (RLEs) of BWT transforms of BANANA, CANADA and concatenation BANANACANADA have, respectively, 7 and 11 Octets, and their NCD distance equals 0.57.

At this stage, the NCD between x and y is the ratio between the size of the compressed representation of the concatenated sequence Cxy minus min(Cx,Cy) and the maximum of size of the compressed representations of each sequence individually, max (Cx,Cy).

The NCD or Vitányi distance is a real mathematical distance, with d(x,x) = 0, d(x,y) = d(y,x), and satisfies the triangular inequality $d(x,z) \leq d(x,y) + d(y,z)$.

Consider now the two words BANANA and CANADA and calculate the Vitányi distance between them (Figure 10). This Vitányi distance using the Burrows–Wheeler transform and run-length compression equals 0.57 (see https://gitlab.com/Orange-OpenSource/documentare/for the calculation program, accessed on 15 October 2023).

5.4. Steps of the Maxwell© Algorithm of Clustering (MAC)

The principle of the Maxwell© used to classify words belonging to the set $\{x_i\}_{i=1,n}$ is to constitute clusters from the distance matrix $D_{ij} = d(x_i,x_j)$. The process is a dynamic and tessellation-based variant of the tree-based clustering approach proposed originally in [38]. In the Maxwell© version, each triplet of words (x,y,z) constitutes a triangle in the graph associated with D, and the area A of this triangle is calculated using the classical Héron formula:

$$A = [p(p - a)(p - b)(p - c)]1/2, \quad (2)$$

with a = d(x,y), b = d(y,z), c = d(z,x) and p = (a + b + c)/2.

The learning procedure can hence follow two strategies:

(1) *One-shot learning*. This procedure is static. It includes a unique forward pass whose role is to calculate the similarity clusters.

(2) *Active learning*. This procedure is dynamic. It includes one or several cycles of forward–backward passes. A backward pass consists of adjusting the validity of the clusters calculated previously in the forward pass based on a feedback loop. Feedback can be carried out either through automated decision, based on a validity calculation, or through the intervention of a human expert decision. At the end of a forward–backward cycle, invalidated data and/or clusters are rejected as singleton to possibly be then incorporated into a new forward–backward learning cycle, which can be summarized as follows:

(1) *Forward pass*

- The forward pass is the core of the original algorithm of Maxwell©. It takes as input the distance matrix D_{ij} and the triangulation standard deviation parameter σ to then execute the following steps:
- Triangulation property calculation: calculation of 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 observed number of standard deviations σ from the area mean value retained;
- First-of-first (1-1) neighborhood pruning: examining sub-graphs whose "useless" (respectively, "best") representative edges are identified as attached to the least (respectively, the most) connected nodes and remove (respectively, keep) them as central nodes;
- Local minima detection: processing subgraphs with several local minima, i.e., nodes whose neighborhood does not contain another node that is closer to the subgraph than the node itself, by using Voronoï networking (as in [5]) for detecting internal boundaries. In practice, this step can be implemented using the Graphiz open API [39] by testing at the end for sub-graphs of which mean and standard deviation depend on local minima, until Graphviz no longer detects any internal boundary;
- Singleton formation: storage of all the elements rejected by this statistical calculation in the form of "singleton clusters";
- Final recall: re-clustering the population of singletons to detect new clusters.

(2) *Backward pass*

In the case of active learning, human expert intervention is permitted to validate singletons after their calculation as to invalidate other cluster elements as a feedback decision. A new recall of the forward pass can be then processed after such a feedback decision.

5.5. Toward Auto-Correction in Maxwell© with Multisets Used for Best Representative Nodes

The auto-correction method was recently added to the Maxwell© platform as an active learning process and has enriched rules of rejects and recalls in an automatized way. When determining the statistical values of a cluster (barycenter node coordinates and mean distance of each element to this node), Maxwell© computes which node is the best representative object of this cluster. In the case of inability to choose a unique node, it builds a "multiset node" containing all the best representative nodes and computes the distances between each element and this multiset (definition of a multiset node is described well in [38,40,41]).

This method allows us to include variability in the representativity of a cluster. For example, one can consider a set of genomes representing different individuals from the same species. Users can determine a threshold for the multiset size, meaning a limit of its variability. This threshold parameter also allows us to detect an excessive cluster growth. In the case of exceeding this threshold, the considered cluster is deleted, and all nodes of its content are distributed again in their nearest neighbor cluster in a new cycle of active learning. We are now experimenting with the use of a similar strategy for labelled object

processing (not applicable here), i.e., in order to minimize clusters having several labelled objects related to multi-hypothesis semantic values, a multi-hypothesis cluster with too many nodes can be deleted, and its content will be reprocessed in a new clustering cycle.

6. Conclusions

Without any contextual data, the classifier Maxwell© is capable of clustering long, full-genome sequences of archaea (for example, 1.78 M bp for Thermofilum pendens) by exploiting only their internal repetitions of motifs (such as pentamers common with an ancestral RNA ring) that appeared during evolution through the effect of the classical genetic operators responsible for genetic variability: mutations, translocations, insertions, deletions, transpositions, inversions, repetitions, symmetrizations, palindromizations, permutations and crossings over. Species that have co-evolved in the same environment have tended to be subject to the action of the same operators, of which their genomes keep track, and Maxwell is able to find the corresponding motifs and cluster genomes in relation to their frequency of occurrence. The example of archaea must be reinforced by experiments with larger genomes from other areas of evolution (bacteria and eukaryotes). In Supplementary Material Figures S2 and S3, we give some examples of such classifications with complete nuclear or mitochondrial genomes of several species (viral, bacterial and mammiferous).

The Maxwell MAC technique requires no training and can be used as a primer step for classical classification tools (like deep learning methods) requiring a training data set or an initial unbiased reference clustering. We will experiment in the future with using Maxwell for labeled object processing (in the case of a preorganization of the knowledge) in order to minimize clusters with several labeled objects related to multi-hypothesis semantic values. In genetics, this labeling can come from previous knowledge on genomes of other classes of the same phylum than the genomes to be classified, or it can result from a previous classification of which we have only kept the barycenter of the classes to accelerate the grouping process of labeled objects. Maxwell's applications are therefore potentially very numerous in genetics, but more generally in all biomedical fields providing numerous data that are difficult to interpret without an extensive prior syntactic and/or semantic grouping work.

Supplementary Materials: The supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms242216278/s1.

Author Contributions: J.G. and J.D. are responsible for the investigation and the writing of the first draft, and C.M., D.B. and T.A. participated in the elaboration of the Maxwell software (Version September 2023) and the final rewriting. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article and Supplementary Materials.

Acknowledgments: The authors thank the NCBI data center, which offers an inexhaustible source of reliable genetic data.

Conflicts of Interest: Joël Gardes, Christophe Maldivi and Denis Boisset were employed by the company Orange Labs. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Steinhaus, H. Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci.* **1957**, *4*, 801–804.
2. MacQueen, J.B. Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 21 June–18 July 1965 and 27 December 1965–7 January 1966; University of California Press: Berkeley, CA, USA, 1967; Volume 1, pp. 281–297.

3. Diday, E. Une nouvelle méthode en classification automatique et reconnaissance des formes la méthode des nuées dynamiques. *Rev. Stat. Appl.* **1971**, *19*, 19–33.
4. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
5. Gualtieri, J.A.; Cromp, R.F. Support vector machines for hyperspectral remote sensing classification. *Proc. SPIE* **1998**, *3584*, 221–232.
6. Hopfield, J.J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA* **1982**, *79*, 2554–2558. [CrossRef]
7. Mattes, J.; Demongeot, J. Dynamic confinement, classification and imaging. In *Studies in Classification, Data Analysis, and Knowledge Organization*; Springer: Berlin/Heidelberg, Germany, 1999; Volume 14, pp. 205–214.
8. Demongeot, J.; Sené, S. The singular power of the environment on nonlinear Hopfield networks. In *CMSB'11, ACM Proceedings*; ACM: New York, NY, USA, 2011; pp. 55–64.
9. Hinton, G. A Practical Guide to Training Restricted Boltzmann Machines. In *Neural Networks: Tricks of the Trade*; Lecture Notes in Computer Science Series; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7700, pp. 599–619.
10. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
11. Cox, A.J.; Bauer, M.J.; Jakobi, T.; Rosone, G. Large-scale compression of genomic sequence databases with the Burrows-Wheeler transform. *Bioinformatics* **2012**, *28*, 1415–1419. [CrossRef] [PubMed]
12. tRNAviz. Available online: http://trna.ucsc.edu/tRNAviz/ (accessed on 23 May 2023).
13. Eigen, M. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **1971**, *58*, 465–523. [CrossRef]
14. Demongeot, J.; Moreira, A. A circular RNA at the origin of life. *J. Theor. Biol.* **2007**, *249*, 314–324. [CrossRef]
15. Rigden, D.J.; Fernández, X.M. The 2021 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Res.* **2021**, *49*, D1–D9. [CrossRef]
16. Lee, B.D.; Neri, U.; Oh, C.J.; Simmonds, P.; Koonin, E.V. ViroidDB: A database of viroids and viroid-like circular RNAs. *Nucleic Acids Res.* **2022**, *50*, D432–D438. [CrossRef] [PubMed]
17. Seligmann, H.; Raoult, D. Stem-Loop RNA Hairpins in Giant Viruses: Invading rRNA-Like Repeats and a Template Free RNA. *Front. Microbiol.* **2018**, *9*, 101. [CrossRef] [PubMed]
18. Stockert, J.C. Prebiotic RNA Engineering in a Clay Matrix and the Origin of Life: Mechanistic and Molecular Modeling Rationale for Explaining the Helicity, Antiparallelism and Prebiotic Replication of Nucleic Acids. *BME Horiz.* **2023**. to appear.
19. Demongeot, J.; Seligmann, H. Spontaneous evolution of circular codes in theoretical minimal RNA rings. *Gene* **2019**, *705*, 95–102. [CrossRef] [PubMed]
20. Demongeot, J.; Gardes, J.; Maldivi, C.; Boisset, D.; Boufama, K.; Touzouti, I. Genomic phylogeny by Maxwell®, a new classifier based on Burrows-Wheeler transform. *Computation* **2023**, *11*, 158. [CrossRef]
21. Demongeot, J.; Thellier, M. Primitive oligomeric RNAs at the origins of life on Earth. *Int. J. Mol. Sci.* **2023**, *24*, 2274. [CrossRef] [PubMed]
22. Novozhilov, A.S.; Koonin, E.V. Exceptional error minimization in putative primordial genetic codes. *Biol. Direct.* **2009**, *4*, 44. [CrossRef]
23. Trifonov, E.N. Consensus temporal order of amino acids and evolution of the triplet code. *Gene* **2000**, *261*, 139–151. [CrossRef] [PubMed]
24. Harish, A.; Caetano-Anollés, G. Ribosomal History Reveals Origins of Modern Protein Synthesis. *PLoS ONE* **2012**, *7*, e32776. [CrossRef]
25. 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] [PubMed]
26. NCBI. Available online: https://www.ncbi.nlm.nih.gov/refseq/ (accessed on 23 June 2023).
27. Luk, A.W.; Williams, T.J.; Erdmann, S.; Papke, R.T.; Cavicchioli, R. Viruses of haloarchaea. *Life* **2014**, *4*, 681–715. [CrossRef]
28. Ngo, V.Q.H.; Enault, F.; Midoux, C.; Mariadassou, M.; Chapleur, O.; Mazéas, L.; Loux, V.; Bouchez, T.; Krupovic, M.; Bize, A. Diversity of novel archaeal viruses infecting methanogens discovered through coupling of stable isotope probing and metagenomics. *Env. Microbiol.* **2022**, *24*, 4853–4868. [CrossRef] [PubMed]
29. Matte-Tailliez, O.; Brochier, C.; Forterre, P.; Philippe, H. Archaeal phylogeny based on ribosomal proteins. *Mol. Biol. Evol.* **2002**, *19*, 631–639. [CrossRef] [PubMed]
30. Petitjean, C.; Deschamps, P.; López-García, P.; Moreira, D.; Brochier-Armanet, C. Extending the conserved phylogenetic core of archaea disentangles the evolution of the third domain of life. *Mol. Biol. Evol.* **2015**, *32*, 1242–1254. [CrossRef]
31. Tahon, G.; Geesink, P.; Ettema, T.J.G. Expanding Archaeal Diversity and Phylogeny: Past, Present, and Future. *Annu. Rev. Microbiol.* **2021**, *75*, 359–381. [CrossRef] [PubMed]
32. Demetrius, L. Directionality Theory and the Origin of Life. *arXiv* **2023**, arXiv:2304.14873.
33. 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.
34. Burrows, M.; Wheeler, D.J. A block-sorting lossless data compression algorithm. *Digit. SRC Res. Rep.* **1994**, *124*, 1–24.
35. Royer, L.; Reimann, M.; Andreopoulos, B.; Schroeder, M. Unraveling Protein Networks with Power Graph Analysis. *PLoS Comput. Biol.* **2008**, *4*, e1000108. [CrossRef]
36. Agustsson, E.; Mentzer, F.; Tschannen, M.; Cavigelli, L.; Timofte, R.; Benini, L.; Gool, L.V. Soft-to-Hard Vector Quantization for End-to-End Learning Compressible Representations. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1141–1151.

37. Cilibrasi, R.; Vitányi, P.M.B. Clustering by compression. *IEEE Trans. Inf. Theory* **2005**, *51*, 1523–1545. [CrossRef]
38. Cohen, A.R.; Vitányi, P.M.B. Normalized Compression Distance of Multisets with Applications. *IEEE Trans. PAMI* **2015**, *37*, 1602–1614. [CrossRef] [PubMed]
39. Graphviz. Available online: https://graphviz.org/ (accessed on 23 May 2023).
40. Vardasbi, A.; Faili, H.; Asadpour, M. On the Reselection of Seed Nodes in Independent Cascade Based Influence Maximization. *Int. J. Inf. Commun. Technol. Res.* **2018**, *10*, 11–21.
41. Pastor-Satorras, R.; Castellano, C.; Van Mieghem, P.; Vespignani, A. Epidemic processes in complex networks. *Rev. Mod. Phys.* **2015**, *87*, 925–979. [CrossRef]