frontiers | Frontiers in Applied Mathematics and Statistics

TYPE Original Research
PUBLISHED 10 December 2025
DOI 10.3389/fams.2025.1634300

OPEN ACCESS

EDITED BY
Alberto Bersani,
Sapienza University of Rome, Italy

REVIEWED BY
Nossaiba Baba,
University of Hassan II Casablanca, Morocco
Zakaria Yaagoub,
University of Hassan II Casablanca, Morocco

*CORRESPONDENCE
Jacques Demongeot
✉ Jacques.Demongeot@univ-grenoble-alpes.fr

RECEIVED 24 May 2025
REVISED 18 November 2025
ACCEPTED 24 November 2025
PUBLISHED 10 December 2025
CORRECTED 15 December 2025

CITATION
Gardes J, Tchatchueng-Mbougua JB, Maldivi C, Jelassi M, ben Khalfallah H and Demongeot J (2025) Maxwell® an AAAA classifier well-suited to biomedical data clustering.
Front. Appl. Math. Stat. 11:1634300.
doi: 10.3389/fams.2025.1634300

Maxwell® an AAAA classifier well-suited to biomedical data clustering

Joël Gardes[1], Jules Brice Tchatchueng-Mbougua[2], Christophe Maldivi[1], Mariem Jelassi[3], Houssem ben Khalfallah[4] and Jacques Demongeot[4]*

[1]Orange Laboratorys, Meylan, France, [2]Centre Pasteur du Cameroun, Yaoundé, Cameroon, [3]ENSI - Ecole Nationale des Sciences de l'Informatique, Campus Universitaire de la Manouba, La Manouba, Tunisia, [4]University of Grenoble Alpes, AGEIS EA 7407, Faculty of Medicine, La Tronche, France

Introduction: A new classifier called Maxwell®, Adiabatic, Agnostic and Almost Autonomous, is presented and used to classify species according to their early occurrence in evolution.
Methods: After a precise description of all the steps of the clustering process, two examples of application are given: first, the classification of simulated genomic data, whose simulation mode is processed by an algorithm allowing the successive application of known operators having acted during the evolution of species. The clustering thus obtained makes it possible to identify correctly the genomes of species having evolved in the same ecosystem. Then, mitochondrial genomes of mammals and giant viruses associated with their bacterial or fungal targets they infect, are classified according to the same criteria.
Results: The results show a good adequacy of the obtained classifications to the evolutionary reality and a high consistency with the known knowledge on the evolution of the oldest species.
Discussion: The Maxwell® classifier presents a unique set of properties, adiabatic, agnostic and almost autonomous, making it particularly suitable for biomedical applications.

KEYWORDS

classification, Maxwell® classifier, evolution, co-evolution cluster, mitochondrial genome, giant virus genome

1 Introduction

To be relevant, modeling in biology and medicine requires access to a large amount of data, often available in public databases such as NCBI [1]. Recently, the Covid-19 pandemic has shown that these data can feed into relevant and effective models, which can be used to explain a posteriori or predict *a priori* complex phenomena such as the transition between endemic and epidemic phases [2] or the role of quarantine and vaccination in preventing epidemic peaks [3–5]. The crucial problem posed by access to these biomedical data, particularly genomic data, is that it is constantly increasing and requires processing using descriptive statistical techniques such as classification before being incorporated into models. The aim of this article is to propose a new classification tool called Maxwell® and to verify its relevance on genomic data. This new tool is a classifier perfectly suited to AI approaches in biology and medicine because of its reversibility. Its methodology is based on a lossless compression tool, the Burrows-Wheeler transform. It can classify any digital object (image, signal, document) by retaining the intermediate results of each data processing step and allowing for reverse processing, a so-called Adiabatic quality, useful in the event of a possible medico-legal trial following the computer-assisted medical decision. Since it does not require any *a priori* knowledge, it is said to be Agnostic. Belonging to the family of unsupervised classifiers, but requiring meta-knowledge to refine the last

clustering step, it is said to be Almost Autonomous. Maxwell® can therefore be considered an AAAA classifier. In **Section 2**, we present the Maxwell®'s successive operating stages, namely the lossless compression of the digital objects to be classified, then the calculation of the distances between these compressed objects allowing the construction of their clusters, and finally their identification using semantic metadata, followed by a refinement of the classification by playing on the classification thresholds. In **Section 3**, we propose an example of application aiming to classify species according to their antiquity in evolution. Then, in **Section 4**, we present a discussion on the place of Maxwell® among the known classification tools and finally in **Section 5** perspectives and conclusion.

2 Materials and methods: Maxwell®'s operating principles

2.1 Lossless compression

The first step in Maxwell®'s operation is the lossless compression following an algorithm due to Burrows and Wheeler allowing to calculate a distance between compressed digital objects to classify [6–11]. This algorithmic approach has been already partly published [12, 13]. Consider two similar sequences of letters, X=BAIGNADE and Y=BADINAGE, with three mutations, I:D, G:I, and D:G, i.e., I changed to D, G to I, and D to G (Figure 1).

After considering all the circular permutations of the words BAIGNADE and BADINAGE, we reorder these permutations by using the alphabetic order, then we note the rank at which appears the initial word (here three for both BAIGNADE and BADINAGE) and retain the ordered last letters of the permutations, that is NBEADIAG for BAIGNADE and BNEAGADI for BADINAGE. The concatenation of the rank with this last sequence gives the Burrows–Wheeler transform BWT, e.g., BWT(BAIGNADE) = BWT(X) = 3NBEADIAG. The last step of compression is to calculate the length of the Run-length encoding (RLE) of BWT(X): here RLE(BWT(X)) = 131N1B1E1A1D1I1A1G, where consecutive occurrences of the same symbol are stored as a single occurrence of that symbol preceded by the count of its consecutive occurrences rather than as the original run: C_X = Length(RLE($BWT_{(X)}$)) = 18 Octets. In the same way, C_Y = Length(RLE(BWT(Y))) = 18 Octets. Then, we do the same for the concatenated word BAIGNADEBADINAGE, whose RLE length of its Burrows-Wheeler transform BWT(XY) is C_{XY} = 24 Octets (Figure 2).

2.2 Normalized compression distance matrix

The concept of Normalized Compression Distance (NCD) comes from a universal approach for comparing arbitrary objects. Derived from Kolmogorov Complexity (KC), NCD offers a domain-independent alternative to specific methods (sequence alignment, image analysis, text comparison, etc.) and can be considered as a computable variant of KC. The generality of its potential applications makes it a conceptually powerful tool, but it also poses practical challenges. KC calculation is unapproachable, but forms the basis of the theory. Brillouin defined Information as a form of negentropy, measurable and costly to create and to erase [6]: erasing a bit of information has a minimal energy cost of $2k_BTln2$, where k_B is the Boltzmann constant and T the temperature [7]; then, Bennett introduced the notion of logical depth taking into account the "computation time" [8]. The Kolmogorov complexity K(X) of an object X is defined by the length of the shortest program generating X and allows to define a Normalized Information Distance (NID) between two objects X and Y:

$$NID(X,Y) = \max(K(X \mid Y), K(Y \mid X))/\max, \quad (1)$$

where K(X|Y) is the conditional complexity [9, 10]. NID is approached by Vitányi's NCD [11] calculated as follows:

$$NCD(X,Y) = d(X,Y) = [C_{XY} \min(C_X, C_Y)]/\max(C_X, C_Y), \quad (2)$$

where C_X is the length of the Run-Length Encoding (RLE) of the Burrows-Wheeler Transform of X, BWT(X).

From the calculation of all the distances between several objects, we can extract the distance matrix D, whose general term D_{XY} is equal to d(X,Y). In the example above, the distance $d(X,Y)=[C_{XY} - \min(C_X,C_Y)]/\max(C_X,C_Y)=1/3$ and the matrix D is given by:

$$D = \begin{pmatrix} d(X,X) & d(X,Y) \\ d(Y,X) & d(Y,Y) \end{pmatrix} = \begin{pmatrix} 0 & 1/3 \\ 1/3 & 0 \end{pmatrix}. \quad (3)$$

2.3 From distance matrix to clusters

After getting the distance matrix D, the goal of the next step is to search for homogeneous and isotropic regions in the distance matrix, following the successive calculations (Figure 3) through 10 consecutive steps:

1) Construction of a triangulation based on the current element, the digital object X, its first neighbor Y and the first neighbor Z of Y,
2) Calculation of the triangle T=(X,Y,Z) area and evaluation of isotropy index S(T) using equations of Figure 3. Formula for calculating a triangle area was discovered by Heron of Alexandria in 1st century AD [14],
3) Calculation of a mean μ and standard deviation σ from the triangle area A histograms then obtaining by eliminating "large triangles" whose area A is over a threshold S(A) based on the number of standard deviations retained. For example, we can exclude elements corresponding to vertices of triangles whose area is more than a threshold $S(A) = \mu + 2\sigma$ (Figure 3A). In the same way, we reject vertices whose triangle is too far from equilaterality. The isotropy index $Q = \frac{3\sqrt{3}A}{a^2}$ equals 1 if the corresponding triangle is equilateral and we reject "distorded triangles" whose Q value is more than a fixed threshold S(Q) in order to ensure the distance homogeneity inside subgraphs, future cluster candidates (Figure 3B),

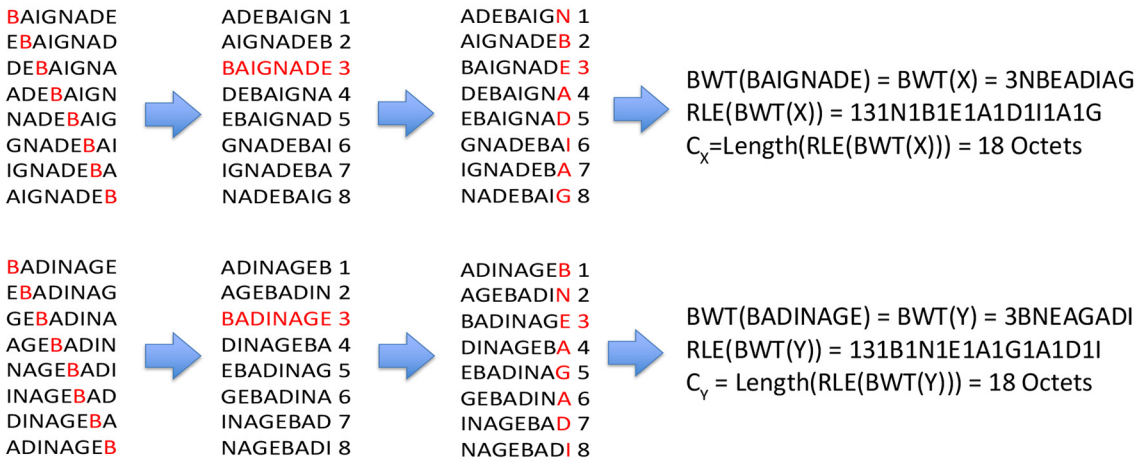

FIGURE 1
The first step of the lossless Burrows-Wheeler compression transform of the words BAIGNADE and BADINAGE.

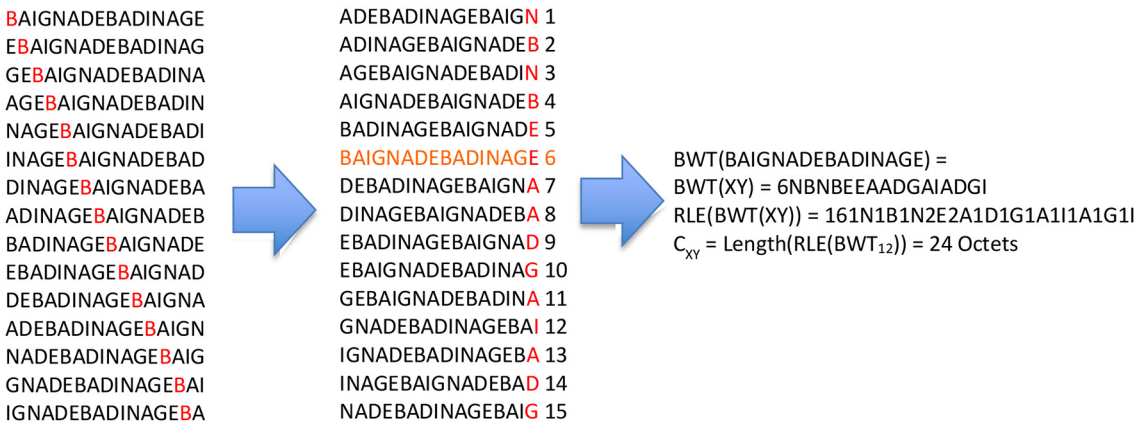

FIGURE 2
Second step of lossless Burrows-Wheeler compression transform of the concatenated word BAIGNADEBADINAGE.

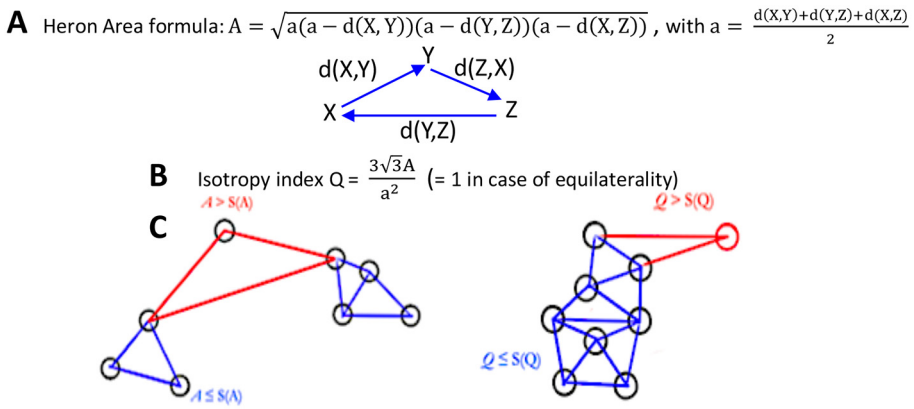

FIGURE 3
(A) Calculation of Heron area formula; **(B)** calculation of Isotropy index (distance to equilaterality); **(C)** thresholding.

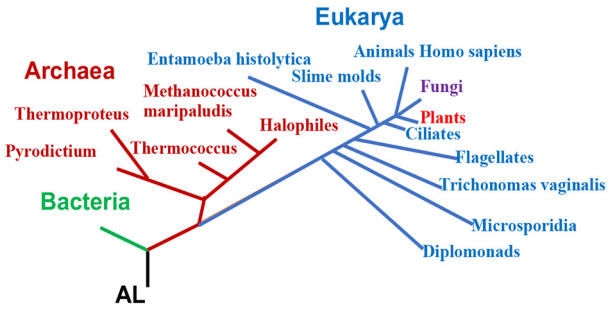

FIGURE 4
Phylogeny of the three domains of life—Archaea, Bacteria, and Eukarya—with indication of some species.

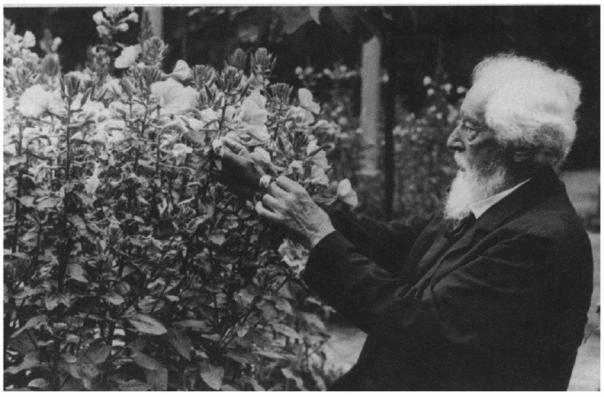

FIGURE 5
de Vries in his garden with *Oenethera Chilena Grandiflora*.

4) Edge processing of the obtained subgraphs using area and equilaterality thresholds S(A) and S(Q) to remove edges that are "useless" to the subgraph's topology and identify the "best representative" vertex as the most connected or the closest to the cluster geometric barycenter depending on the preferred criterion, i.e., realizing a local optimum (max connected or min distanced),
5) Identification of the subgraphs with multiple best representatives, by using a Voronoï tessellation algorithm (from GraphViz) for detecting internal boundaries within these subgraphs,
6) Decision test on new triangles made from internal Voronoï boundaries, at the end of which thresholds of new mean and standard deviation of area and isotropy index no longer detect new internal boundaries,
7) Storing as "singleton clusters" all elements rejected by the previous statistical calculations,
8) Recall of initial process (1) performed on the singleton population to detect new clusters until stabilization of "singleton clusters" and their affectation to the closest not singleton cluster,
9) Identification of final clusters by using metadata,
10) Final validation by experts of the field with a possible cluster concatenation under semantic arguments.

3 Results: example of application with the evolutionary random genetic operators with a constant rate in an evolutive ecosystem

3.1 The genetic operators of the evolution

Maxwell® classifier is applied to genetic data from a sample of species in the three main domains appearing progressively during the evolution and evolving parallelly until present time: Archaea, Bacteria, and Eukarya (Figure 4).

The first modification of the genomes described on *Oenothera Chilena Grandiflora* as a mutation by de Vries [15] was in reality a translocation (Figure 5). This genetic change belonged to the main genetic operators, i.e., transformations of nucleotide sequences involved in the evolution of genomes, which are the following: point mutation, insertion (e.g., after reverse transcription), deletion, inversion, translocation, transposition, duplication, symmetrization, palindrome generation, permutation, and recombination (crossing-over).

3.2 A proposal for the first RNA

Many authors have proposed explanations for the first stages of life on Earth, 3.5 billion years ago. Among these, after supporters of DNA as the primordial molecule [16, 17] a school gradually emerged proposing RNA as the initial molecule [18–23], then a group proposing a third way, between DNA first and RNA first, "the third way of evolution" [24] in favor of an early interaction between RNAs in equilibrium between an active linear form or ring and a memory hairpin form, the first catalyzing the polymerization of the first peptides, from nucleotides and amino acids synthesized following the hypothesis of Miller [25]. Inspired by these works, we have proposed in a series of previous articles an RNA candidate for serving as catalyzer of the first peptide syntheses [26–32]. The first steps for finding this RNA we will call in the following AL (for ALpha or Archetypal Loop) was supposing that it had two satisfy four criteria for optimizing the primordial catalysis of peptide synthesis. Then, the concept of a ring for the structure of AL has been considered as a sort of circular consensus capable of embedding all possible genetic encodings, with the four choice criteria summarized as follows:

1) AL must satisfy the principle "be as short as possible and contain at least one codon per synonymy class of the genetic code,"
2) AL codon sequence obtained with overlap after three turns of its circular form must begin with the start codon and end with the stop codon,
3) the AL must have a hairpin configuration in balance with its circular shape, and this hairpin must have a

minimum head length (3nt) and a maximum number (9) of codon pairs,

4) if multiple rings possess properties (1) to (3), they must have a single barycenter for classical inter-ring distances (circular Hamming, permutation, and editing distances), i.e., the AL ring.

By formalizing the problem in a search for a Hamiltonian path between the nodes of a circular graph representing the 20 amino acids (Figure 6), we get among 4^{22} possible solutions:

* No solution if AL contains 20 or 21 nucleotides, i.e., 20 ou 21 overlaping triplets,
* 29 520 solutions if AL contains 22 nucleotides, i.e., 22 triplets ending with an END codon,
* from where 25 with a maximal hairpin form (3 head free bases + 9×2 stem bases + 1 tail free base),
* with only one starting with AUG, repeating AUG and being barycenter of the 24 others as the ring built from sequence AUGGUACUGCCAUUCAAGAUGA. More, it the closest to the set of all known tRNAs.

In the following, we will use a proximity of a given genome to the RNA AL, denoted P_{AL} Doublet [32] by calculating the number of standard deviations between the observed and expected numbers of motifs common with AL (here pairs of closed trimers from AL), which allows to calculate the probability of getting by chance [33] these traces from AL in current genomes (e.g., the mitochondrial one obtained from NCBI [1]).

3.3 The generation of synthetic genomes having been changed by a sequence of successive of genetic operators respecting a certain proportion of each

We developed a custom genome evolution simulator to generate artificial nucleotide sequences across multiple generations, mimicking biological diversification processes due to evolution operators. The simulation is implemented in R and is centered around the function, which takes as input a reference genome and simulates its descent through *n* generations.

3.3.1. Initial input and generation process

The simulation begins with a single reference genome, represented as a character string (e.g., RNA or DNA sequence). This genome constitutes Generation 1. For each subsequent generation (i = 2, ..., *n*), a random subset of genomes from the previous generation (i – 1) is selected to act as parental genomes. The number of selected parents is capped at p, a user-defined parameter controlling the branching factor of the simulation. Each parent can produce descendant genomes, each of which being derived by applying a sequence of randomly selected evolutionary transformations. The number of transformations applied to each descendant is also randomly chosen between 1 and 11.

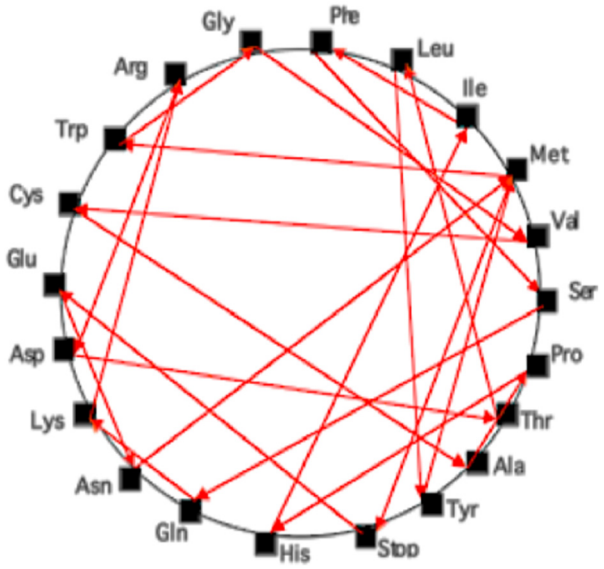

FIGURE 6
Hamiltonian path used on the amino-acids set to find the archetypal RNAs.

3.3.2 Implemented evolutionary operations

Each descendant genome is subjected to a pipeline of transformations chosen from the following set:

- Point Mutation: random substitution of individual nucleotides,
- Insertion: addition of a random subsequence at a random position,
- Deletion: removal of a segment of the genome,
- Inversion: reversal of a segment's nucleotide order,
- Translocation: movement of a segment from one position to another,
- Transposition: segment is excised and reinserted at a new location,
- Duplication: a region is copied and inserted elsewhere,
- Symmetrization: production of the reverse-complement (for RNA sequences)
- Palindrome generation: creation of palindromic sequences to simulate structural patterns,
- Permutation: random shuffling of a genomic segment,
- Recombination (Crossing-over): exchange of regions between sequences.

These operations are designed to replicate biologically plausible events observed in genome evolution and introduce substantial sequence variability across generations.

3.3.3 Output format and file naming convention

Each descendant genome is saved as a text file in the format: descendant_i_k_j, where:

- i refers to the generation number (e.g., i = 2 for Generation 2),

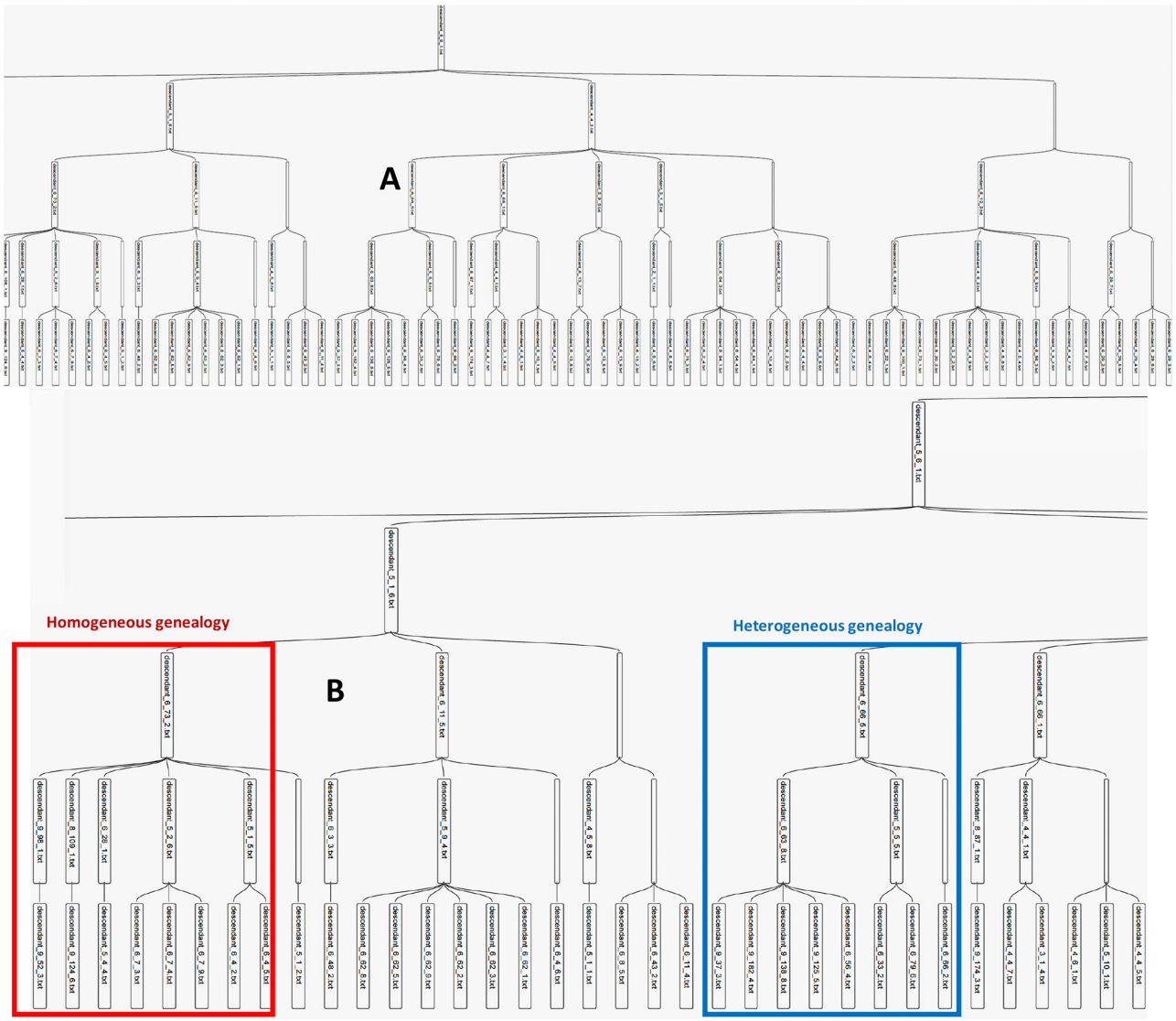

FIGURE 7
(A) Five successive generations of simulated genomes (with random action at each generation of some of the eleven evolution operators) classified by Maxwell®. **(B)** A part of the clustering showing a subtree respecting the generation order (in red) and another part mixing the generations (in blue).

- k refers to the index of the parent genome within Generation i−1,
- j refers to the index of the descendant genome in Generation i derived from this parent k.

This naming scheme ensures traceability and allows reconstruction of the full ancestral path of any genome, which is critical when evaluating hierarchical clustering or tree-based reconstruction methods such as Maxwell®'s classification algorithm (Figures 7A, B).

3.3.4 Interpretation of the Maxwell® classification results

The hierarchical classification obtained using the Maxwell® algorithm (Figures 7A, B) was evaluated against the known genealogy of simulated genomes. In cases of homogeneous classification, the dendrogram exhibited a clear structure in which genomes derived from the same ancestor (identified by shared parent and generation indices in their filenames, e.g., descendant_i_k_j) were grouped within the same branches. This concordance suggests that the classifier successfully captured the evolutionary relationships embedded in the data, even in the presence of complex mutational events such as insertions, deletions, and transpositions. Conversely, in heterogeneous classifications, descendant genomes originating from the same parent were dispersed across multiple branches, and sequences from distant generations were occasionally grouped together (Figure 8). This pattern indicates potential challenges for the classifier in preserving genealogical coherence when evolutionary noise increases or when sequence divergence becomes too pronounced. Such discrepancies underscore the limits of structural

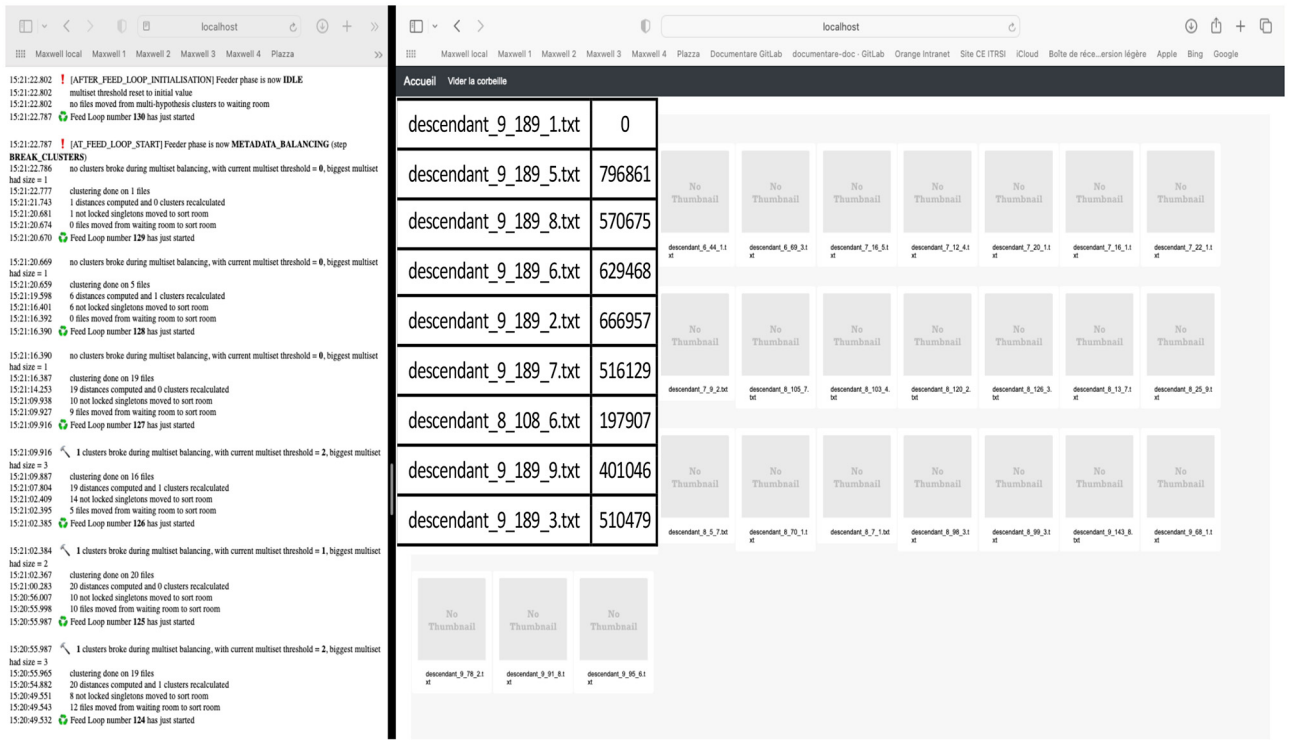

FIGURE 8
Examples showing the management of clusters with some not respecting the surface nor equilaterality criteria **(left)**, a homogeneous cluster made of members of the same descendance **(middle)** and singletons from clusters broken expecting their rearrangement in new clusters respecting the both criteria **(right)**.

similarity measures in reconstructing deep or highly perturbed evolutionary histories. Overall, these results demonstrate the ability of the Maxwell® classifier to retrieve hierarchical structure under moderate evolutionary variation, while also highlighting the importance of controlling for transformation intensity in simulated datasets when evaluating classification robustness.

3.4 Classification of mitochondrial genomes

The mitochondrial genomes of 10 mammalian species were obtained from the NCBI site and their classification allowed to obtain from interspecies distance matrix four classes of size 2 (hominidae, whales, seals, and murines) and two singletons (horse and cat) of which we will only comment on two intra-class and two inter-class proximities (Figure 9), comparing the results of Maxwell® to those of the NCD classical classifier [10].

The primate cluster {human, chimpanzee} is obtained in both classifiers with a distance intraclass smaller for Maxwell® (0.553) than for NCD (0.655). It is the same for the whale cluster {blueWhale, finWhale} (0.49 vs. 0.61). The mean interclass distance between both primate and whale cluster is smaller for Maxwell® (0.90166) than for NCD® (0.9226). Hence with Maxwell®, the internal homogeneity was favored over external heterogeneity, contrary to what is observed for NCD.

3.5 Mitochondrial genome classification and proximity to AL

Comparing the primate classification to equid one (Figures 10A, B), several points of convergence are observed:

- The consistency with previous studies on the evolution of mammalian families: chimpanzees are close to humans and more distant from other simians (gorilla, orangutan, gibbon and baboon) [35],

- Perissodactyla species (Figure 10C) are in the same cluster, with rhinoceroses appearing older than equids, i.e., having a greater proximity (in red in Figure 10B, see Supplementary material S1) to archetypal RNA AL [36, 37].

4 Discussion

The choice of the compressor strongly influences the metric structure of the NCD. The BWT followed by RLE with introducing a locking pattern (or sentinel) outside the alphabet preserves the reversibility during concatenations, and this reversible transformation tends to satisfy both symmetry and triangle inequality axioms of the distance C: $C_{XY} \approx C_{YX}$ and $C_{XZ} \leq C_{XY} + C_{YZ}$. This brings the NCD closer to a Hilbert-type distance and reinforces its Euclidean character. For DNA sequences where the alphabet is {A,C,G,T}, the letter Z can be used. For example, with X=ACGTTAAAA and Y=AATGCT, a naive

NCD®

	graySeal	rat	mouse	finWhale	harborSeal	human	horse	cat	blueWhale	chimpanzee
graySeal	0	0,919705	0,924487	0,908946	0,38332	0,928216	0,889022	0,878782	0,90257	0,925278
rat	0,913329	0	0,847857	0,918233	0,911846	0,927023	0,913696	0,916699	0,916817	0,926073
mouse	0,919107	0,845826	0	0,928828	0,917987	0,931597	0,926512	0,912967	0,915612	0,934022
finWhale	0,89739	0,919432	0,922631	0	0,900357	0,922649	0,897241	0,908055	0,606357	0,926669
harborSeal	0,385103	0,91462	0,916204	0,907884	0	0,924128	0,882726	0,872102	0,893819	0,919176
human	0,930205	0,934381	0,936966	0,92762	0,924921	0	0,923842	0,931827	0,923842	0,659181
horse	0,880056	0,906488	0,9169	0,890444	0,875792	0,919467	0	0,877014	0,889467	0,920707
cat	0,866012	0,914931	0,916896	0,911788	0,869352	0,926719	0,887623	0	0,903929	0,922004
blueWhale	0,893007	0,921037	0,919028	0,614154	0,891838	0,923245	0,901882	0,90668	0	0,92349
chimpanzee	0,919714	0,925477	0,927266	0,921701	0,91462	0,65163	0,921502	0,923772	0,917329	0

Maxwell®

	blueWhale.t	cat.txt	chimpanzee	finWhale.tx	graySeal.txt	harborSeal.t	horse.txt	human.txt	mouse.txt	rat.txt
blueWhale.tx	0	0,867045	0,910633	0,488832	0,857191	0,848628	0,855595	0,906014	0,898921	0,907915
cat.txt	0,867045	0	0,902411	0,868379	0,807907	0,801234	0,838268	0,906081	0,89482	0,888648
chimpanzee.	0,910633	0,902411	0	0,908174	0,899543	0,90401	0,894777	0,552993	0,925555	0,906902
finWhale.txt	0,488832	0,868379	0,908174	0	0,872099	0,859688	0,849002	0,901434	0,915159	0,901741
graySeal.txt	0,857191	0,807907	0,899543	0,872099	0	0,284255	0,839319	0,916991	0,900051	0,888447
harborSeal.t	0,848628	0,801234	0,90401	0,859688	0,284255	0	0,830984	0,914057	0,898269	0,891262
horse.txt	0,855595	0,838268	0,894777	0,849002	0,839319	0,830984	0	0,894732	0,899778	0,885363
human.txt	0,906014	0,906081	0,552993	0,901434	0,916991	0,914057	0,894732	0	0,937569	0,914581
mouse.txt	0,898921	0,89482	0,925555	0,915159	0,900051	0,898269	0,899778	0,937569	0	0,793703
rat.txt	0,907915	0,888648	0,906902	0,901741	0,888447	0,891262	0,885363	0,914581	0,793703	0

FIGURE 9
Matrices of distances between different species calculated by two different classifiers, NCD on the top [10] and Maxwell® on the bottom, with indication (in red) of the human to chimpanzee distance.

FIGURE 10
(A) Taxonomy of hominoides; **(B)** Three clusters obtained by Maxwell® confirming the antiquity of the species as known in classical taxonomic studies. The proximity to AL (P_{AL} Doublet) of mitochondrial genome is indicated in red; **(C)** Taxonomy of Perissodactyla.

concatenation can produce ambiguities during decompression. ZACGTTAAAAZAATGCT explicitly marks the boundary, ensuring consistent cross-compression and a more stable distance, improving the Euclidean quality of the distance matrices. The NCD, based on compression, offers a conceptual and practical framework for comparing heterogeneous objects. This universality

FIGURE 11
(A) Cluster of giant viruses from Pandoravirus family with their target Acanthamoeba (ribosomal genome); **(B)** Cluster Legionella (ribosomal genome); ((**C**) Cluster Mimivirus with its virophages Zamilon and Sputnik, and their targets Acanthamoeba (mitochondrial genome) and Woesearchaeota (whole genome); **(D)** Giant viruses classification based on the RNA-dependent RNA polymerase using FastTree algorithm [34]. AL proximity P_{AL} Doublet is indicated (in red).

in the application to any type of data and any type of domain brings adavantages such as: (i) no need for alignment, annotation, or expert features and (ii) consistent results on a small corpus. On a small corpus of biological sequences, NCD approach finds expected structures without requiring specific knowledge. Its limitations—cost and dependence on the compressor—call for further work, but its interdisciplinary potential is evident. The above results in genomics show that the information carried by nucleotide sequences alone allows genomes to be organized in a way that respects known knowledge of the evolution of species. This information concerns the way in which species have evolved while being subject to the operators of evolution. If they have co-evolved while belonging to the same ecosystem where they maintained relationships of the parasitism, saprophytism or simply commensalism type, it is not surprising to find them in the same cluster. In Figure 11 for example, the well-characterized giant viruses have well-identified clusters, linked for example to Pandoravirus and Mimivirus. The trace of the contamination of giant viruses by the virophages Zamilon and Sputnik and that of the contamination of the amoeba Acanthamoeba by a virus of the Mimivirus class (trace in its mitochondrial DNA) and by a virus of the Pandoravirus class (trace in its tyrosine-tRNA ligase) is justified by their evolution in the same marine ecosystem. Legionella bacteria are isolated by Maxwell® clustering classification, but having traces of ancient contamination by Acanthamoeba, it joins the Pandoravirus cluster, if we relax the clustering constraints. This clustering based only on nucleotide sequences shows that they could have a common origin, possibly due to the fact that first peptides are supposed to be formed with catalysis by primordial RNAs [19, 38–45], analogous to experimental synthesis of dipeptides on RNA template [46–50]. These peptides left relics in current proteins of ancient organisms like Entamoebae (PAL Doublet proximity to AL in red, see Supplementary material S1), reinforcing the hypothesis of existence of weak bonds between RNA and amino acids, in connection with a progressive appearance of the current genetic code [46–50].

Concerning the metrology, Maxwell® relies on high-performance computing standards, guaranteeing a good scalability:

- Performance automatically adapts to the software environment (memory, CPU cores number),
- NCD formula (2) demonstrates its algorithmic simplicity. The relative slowness in calculation is common to all lossless compressors and is the reasons for the weak lack of gains with the GPU,
- Clustering is handled by graph theory and performance is that of the graph management library, here Graphviz®, which is a benchmark,
- The use of a NoSQL database (MongoDB) for the data storing frees from memory management and input/output issues, delegating these to PostgreSQL.

Concerning the usage, Maxwell® can operate in incremental mode thanks to an epoch mechanism, similar to neural network learning systems, where each epoch alternates between a

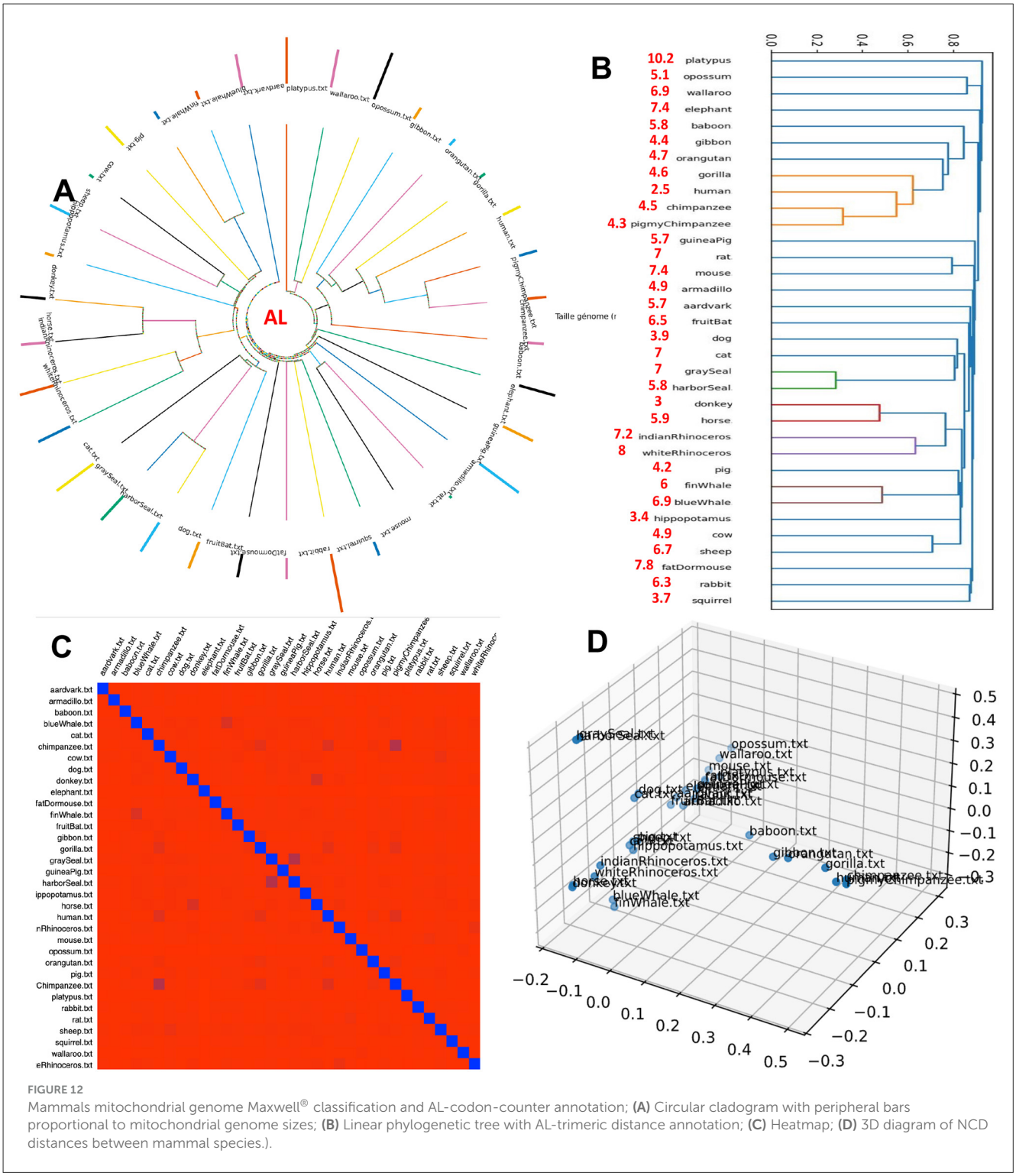

FIGURE 12
Mammals mitochondrial genome Maxwell® classification and AL-codon-counter annotation; **(A)** Circular cladogram with peripheral bars proportional to mitochondrial genome sizes; **(B)** Linear phylogenetic tree with AL-trimeric distance annotation; **(C)** Heatmap; **(D)** 3D diagram of NCD distances between mammal species.).

classification phase and a clustering phase for unclassified elements. This property means that it is not necessary to repeat the entire learning process when a new dataset is added. This ensures compliance with a reference base while allowing it to evolve. The result is a system capable of merging datasets.

One of the main advantages of Maxwell is the absence of semantic data, because Maxwell uses only information in octets allowing decision tests based on deliberately simple statistical values (means and standard deviations) due to their algorithmic simplicity. However, once the cluster calculations have been performed, we can evaluate the robustness of the relationships between groups of measured data and create a posteriori semantic classes using metadata to detect semantically ambiguous clusters that will be distributed across several semantic classes.

Another important advantage is the reversibility of the classification process, that is, the ability to return to all stages of this process, which may be necessary to find a detachable error in a medico-legal process of proving the origin of bad diagnostic or therapeutic advice.

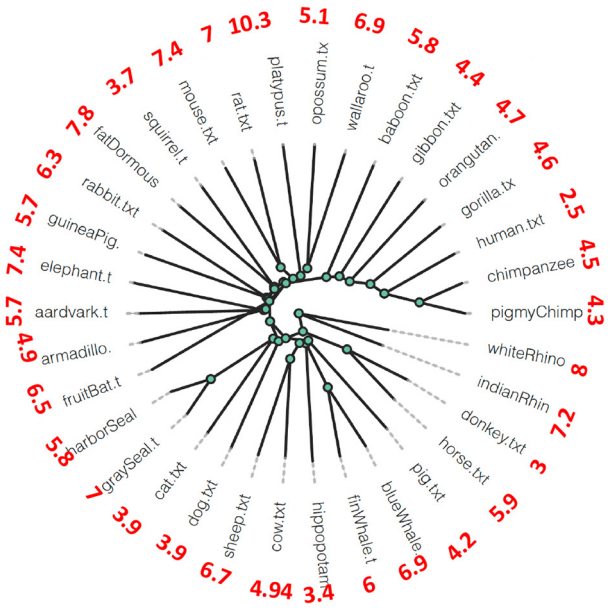

FIGURE 13
Maxwell's phylogenetic tree of mammal species based on their mitochondrial genome. The bifurcation points between species are indicated in green. AL proximity P_{AL} Doublet is indicated (in red).

5 Perspectives and conclusion

We have presented a new classification tool called Maxwell®, whose main characteristics are to be adiabatic (reversible), agnostic (without the need to introduce *a priori* knowledge) and almost autonomous (given the possibility of refining a posteriori clusters that are too large, therefore often heterogeneous, or too small, such as singletons). The application of the Maxwell® classifier to the genome shows that it is particularly well-suited to detect transformations of primary nucleotide information due to evolutionary operators, therefore to classify in the same cluster species that have evolved in the same eco-system. Only a thousand synthetic genomes and forty species were studied in this article, but the results demonstrate the relevance of the classifier to bring together neighboring evolved genomes.

In this direction, three research topics can be considered, in order of increasing complexity: (i) generalizing to all species the calculation of the age of their genome and their attachment to existing clusters in the three kingdoms of life: Archaea, Bacteria, and Eukaryotes; (ii) processing data related to the interaction between genes in genetic control networks. Classifying networks requires the introduction of new distances between graphs, but the problem can be solved using Maxwell's algorithm; (iii) processing epigenetic control networks, which encompass genetic networks and their exogenous control parameters related to infectious agents, environmental factors, and biological clocks. These three domains of research are challenging but represent essential opportunities in biomedical research. In view of the current work, two areas of ongoing work can be identified:

1) restitution of the results of the data processing

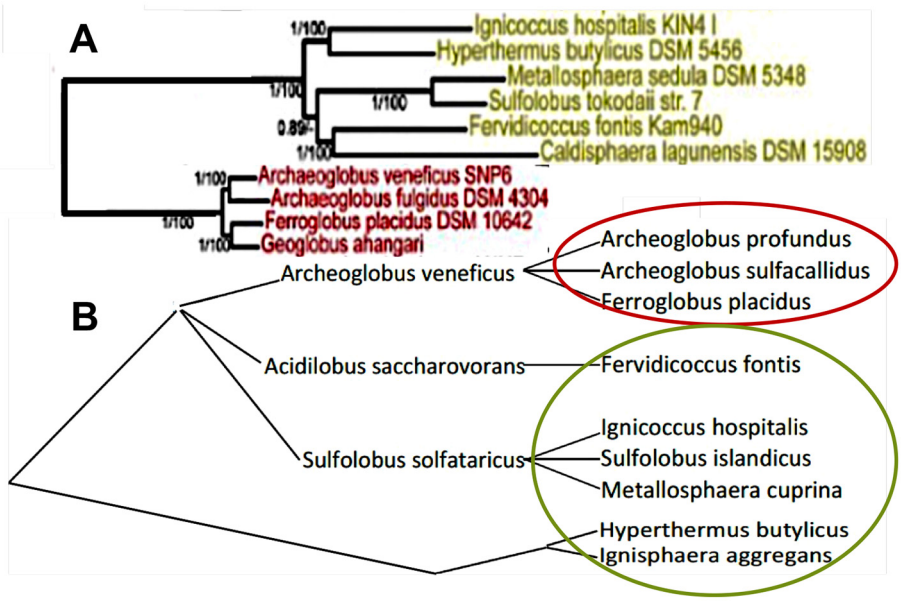

FIGURE 14
Phylogenetic tree of crenoarchaeota (in green) and archaeoglobi (in red). **(A)** Classification method based on a neighbor-joining analysis of 637 homologous sequences chosen in the genome of each archaeum [54]; **(B)** Maxwell's classification of chosen crenarchaeota and archaeoglobi.

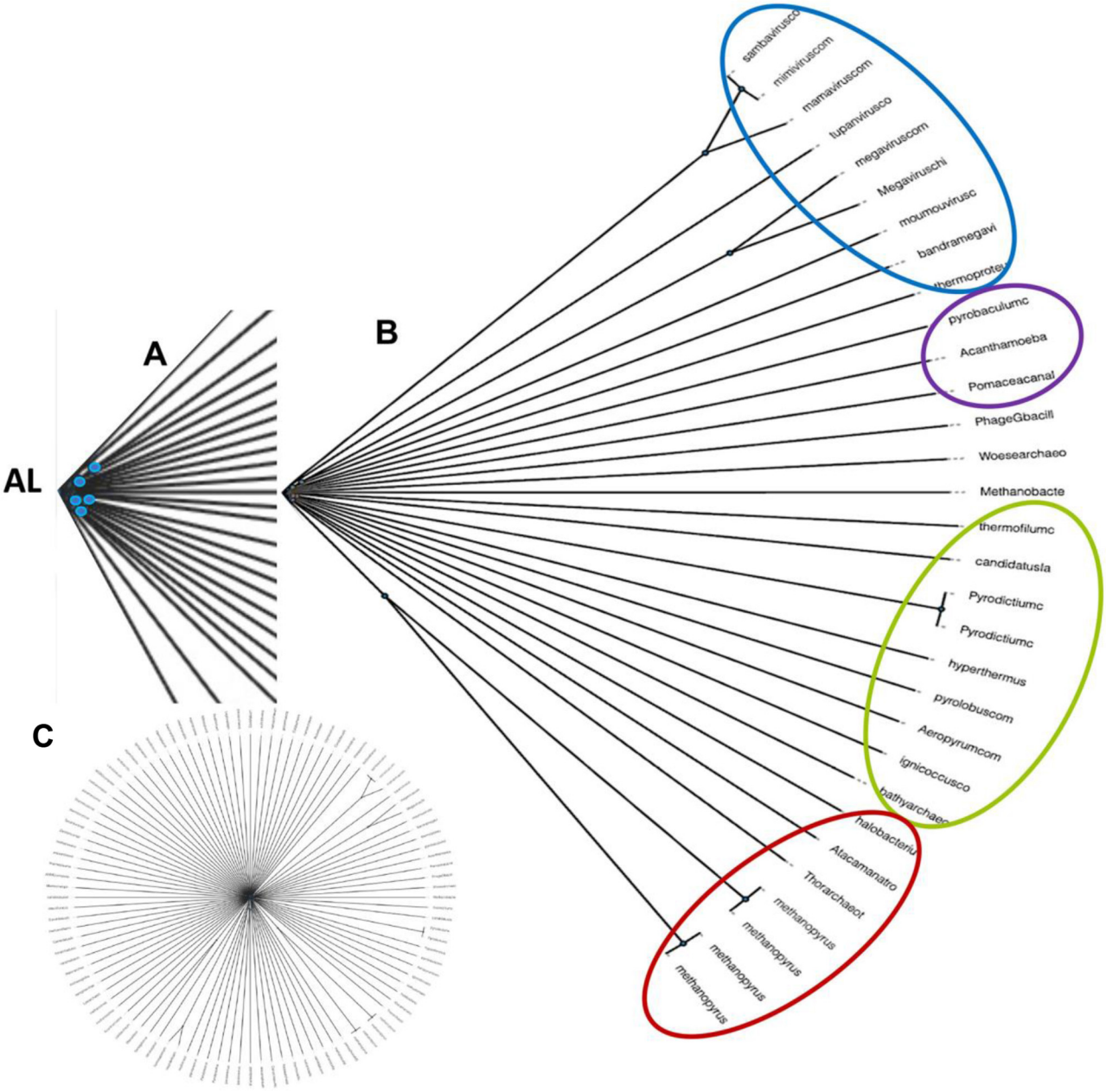

FIGURE 15
Maxwell's classification of giant viruses and their putative targets chosen in amoebae, crenarchaeota and archaeoglobi families. **(A)** thumbnail image showing the start of the phylogenetic tree with the bifurcations of the branches (in blue); **(B)** classification of giant viruses (blue circle), their targets (violet circle), crenarchaeota (in green) and euryarchaeota (in red); **(C)** whole cladogram of Maxwell's classification.

Linear phylogenetic trees are limited by their height. We have therefore started to replace them with more compact circular cladograms (Figure 12).

Work remains to be done to refine the graph at its root level. The same applies to the output in the form of a linear dendogram, whose bifurcation points have to be emphasized (as on Figure 13) and annotations more linked to the classification by showing their possible inconsistency with the clustering due to Maxwell® (Figure 12B). The restitution in the form of a heatmap (Figure 12C) or a 3D representation of the distances between species (Figure 12D) must provide additional information on the proximities between species, to be incorporated into the display of the results, to be able to use them in real time, if there is a modification of the input data.

2) generalization of the method to large data sets.

The NCBI Nucleotide repository contains the genomes of more than 160,000 species, each of which containing approximately 600 RNAs of evolutionary interest (tRNA, rRNA, miRNA, circRNA, mRNA and the corresponding proteins [51, 52]), primarily those involved in transcription, translation, and cellular energy, which are essential for cellular life. The next step will be the progressive generalization of the classifications presented here to the 160,000

species currently present in the NCBI server [1]. Expanding the classification work proposed here and in [32] to very large data sets will be a natural program for the future of our current approach.

3) Comparison with other clustering methods

In [53], we have compared 25 classification techniques of sepsis diagnosis, from classical k-means and multiple regression tools to the deep learning methods after what we have decided to use a new classifier having the same performance than the best one in sepsis diagnosis (surprisingly the multiple regression), but having the advantage to be reversible (each step is explainable), which is a necessary condition of acceptance in a medico-legal context. In Figure 14, a comparison between the Maxwell's approach and a classical classifier [54] has been done on genomic data coming from the whole genome of a sample of Archaea in case of Maxwell® and from a chosen set of genes for the alternative method. In this last example, the obtained phylogenetic tree (Figure 14A) was inferred by using a neighbor-joining analysis of 637 homologous sequences chosen in genome of Archaea.

The scale bar on Figure 14A represents 10 mutations per 100 nt of the positions of these homologous sequences. It is indicated as a fraction the percentage of 100 bootstrap resamplings supporting the topology of the neighbor joining skeleton. On Figure 14B, the Maxwell's tree has been obtained by using the whole genome of soe Aechaea chosen in two families, the crenarchaota (in yellow in Figure 14A) and the archaeaglobi (in red in Figure 14A).

In conclusion, the Maxwell® classifier presents a unique set of characteristics, making it particularly suitable for biomedical applications:

1) Its adiabatic or reversible nature makes it possible to isolate the steps involved in a potential diagnostic or therapeutic advice error. This ability to demonstrate a detachable error, excluding, for example, the physician's liability, makes it a means of meeting the medico-legal requirements of medical practice.
2) Its agnostic nature, i.e., its independence of *a priori* semantics or semiology, makes it suitable for processing data such as nucleotide or amino acid strings coming from a sequencer for example, without any information other than the obtained sequences. In Figure 15, using only the whole sequence of the archaeal or viral genomes, it is possible to obtain clusters coherent with the classical taxonomy.
3) Its autonomy, stemming from its unsupervised nature, makes it capable of automatically providing a classification, which will be interpreted using metadata and may subsequently lead to adjustments to the cluster aggregation thresholds in a second, partially non-autonomous phase. The sensitivity to the threshold distance parameter for cluster assignment is handled interactively by Maxwell®, allowing singletons to be assigned to the nearest cluster or, conversely, to split up clusters that are too large and heterogeneous.

In summary, Maxwell® is a new adiabatic, agnostic and almost autonomous classifier, whose efficiency will be systematically compared in a future work on very large genetic databases to the existing classifiers listed in [53, 55]. The final choice of a classifier, in the case of a biomedical application, is the subject of a compromise between the precision and speed of its clustering and its capacity to provide clear answers in the identification of the faulty stages of its reasoning, in the context of forensic investigations, similar to those carried out on the non-digital medical human chain of diagnosis and care, to identify and qualify the responsibilities of a possible error. Then, its major application concerns the medical diagnosis in which a prior classification of a training set of patients allows, in the generalization phase to a larger patient population, to test the accuracy of the assignment of a new patient to the class that corresponds to him. In the case of the genome, the major interest of Maxwell® lies in its capacity to find clusters using only the sequence of the entire genome or that of cellular structures such as mitochondria (for mammals) and chloroplasts (for plants). A version of Maxwell® can be found online in [55].

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JG: Conceptualization, Investigation, Methodology, Software, Supervision, Validation, Writing – original draft, Writing – review & editing. JT: Investigation, Software, Writing – original draft, Writing – review & editing, Data curation, Formal analysis. CM: Data curation, Investigation, Software, Writing – review & editing, Conceptualization, Methodology, Validation. MJ: Conceptualization, Investigation, Methodology, Software, Writing – review & editing. HK: Conceptualization, Investigation, Software, Data curation, Writing – review & editing. JD: Conceptualization, Data curation, Investigation, Software, Writing – review & editing, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The 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.

Correction note

This article has been corrected with minor changes. These changes do not impact the scientific content of the article.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fams.2025.1634300/full#supplementary-material

References

1. NCBI (2025). Available online at: https://www.ncbi.nlm.nih.gov/refseq/ (Accessed May 23, 2025).

2. Demongeot J, Magal P. Data-driven mathematical modeling approaches for COVID-19: a survey. *Phys Life Rev*. (2024) 50:166–208. doi: 10.1016/j.plrev.2024.08.004

3. Yaagoub Z, Sadki M, Allali K. Global stability of spatio-temporal with quarantine and vaccination. *J Indonesian Math Soc.* (2024) 30:321–37. doi: 10.22342/jims.30.2.1452.321-337

4. Yaagoub Z, Farah EM, Ahmad S. Three-strain epidemic model for influenza virus involving fractional derivative and treatment. *J Appl Math Comput.* (2025) 71:1247–66. doi: 10.1007/s12190-024-02284-0

5. Sadki M, Yaagoub Z, Allali K. Qualitative analysis of a fractional-order for a within-host infection dynamics with adaptive immunity using caputo derivative. *Iranian J Sci.* (2025) 49:847–69. doi: 10.1007/s40995-024-01768-9

6. Brillouin L. *Science and Information Theory*. New York, NY: Academic Press. (1956).

7. Landauer R. Irreversibility and heat generation in the computing process. *IBM J Res Dev*. (1961) 5:183–91. doi: 10.1147/rd.53.0183

8. Bennett CH. Logical depth and physical complexity. In : Herken R, editor. *A Half-Century Survey on the Universal Turing Machine*. New York, NY: Oxford University Press (1988). pp. 227–57. doi: 10.1093/oso/9780198537748.003.0008

9. Li M, Vitányi PMB. *An Introduction to Kolmogorov Complexity and its Applications*. New York, NY: Springer (1993). doi: 10.1007/978-1-4757-3860-5

10. Cilibrasi R, Vitányi P. Clustering by compression. *IEEE Trans Inf Theory*. (2005) 51:1523–45. doi: 10.1109/TIT.2005.844059

11. Burrows M, Wheeler DJ. A block-sorting lossless data compression algorithm. *Digit SRC Res Rep*. (1994) 124:10009821328.

12. 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) 264:1464–5. doi: 10.3233/SHTI190486

13. Demongeot J, Gardes J, Maldivi C, Boisset D, Boufama K, Touzouti I. Genomic phylogeny using the Maxwell® classifier based on Burrows–Wheeler Transform. *Computation*. (2023) 11:158. doi: 10.3390/computation11080158

14. Drachmann AG. Heron and ptolemaios. *Centaurus*. (1950) 1:117–31. doi: 10.1111/j.1600-0498.1950.tb00576.x

15. De Vries H. *Die Mutationstheorie*. Leipzig, Germany: Veit & Co (1901).

16. Watson JD, Crick FHC. The structure of DNA. *Cold Spring Harbor Symp Quant Biol.* (1953) 18:123–13. doi: 10.1101/SQB.1953.018.01.020

17. Monod J. *Chance and Necessity: An Essay on the Natural Philosophy of Modern Biology*. New York, NY: Knopf (1971).

18. Paecht-Horowitz M, Berger J, Katchalsky A. Prebiotic synthesis of polypeptides by heterogeneous polycondensation of amino-acid adenylates. *Nature*. (1970) 228:636–9. doi: 10.1038/228636a0

19. Eigen M. Selforganization of matter and the evolution of biological macromolecules. *Naturwiss*. (1971) 58:465–523. doi: 10.1007/BF00623322

20. Katchalsky A. Prebiotic synthesis of biopolymers on inorganic templates. *Naturwiss*. (1973) 60:215–20. doi: 10.1007/BF00625709

21. Tamura K, Schimmel P. Oligonucleotide-directed peptide synthesis in a ribosome- and ribozyme-free system. *Proc Natl Acad Sci USA*. (2001) 98:1393–7. doi: 10.1073/pnas.98.4.1393

22. Xiao H, Murakami H, Suga H. Ferré-D'Amaré AR. Structural basis of specific tRNA aminoacylation by a small *in vitro* selected ribozyme. *Nature*. (2008) 454:358–61. doi: 10.1038/nature07033

23. Deng J, Wilson TJ, Wang J, Peng X, Li M, Lin X, et al. Structure and mechanism of a methyltransferase ribozyme. *Nat Chem Biol.* (2022) 18:556–64. doi: 10.1038/s41589-022-00982-z

24. Shapiro JA. Why the third way of evolution is necessary. *Theor Biol Forum*. (2021) 114:13–26. doi: 10.19272/202111402002

25. Miller SL. A production of amino acids under possible primitive Earth conditions. *Science*. (1953) 117:528–9. doi: 10.1126/science.117.3046.528

26. Demongeot J. *Au Sujet de Quelques Modèles Stochastiques Appliqués à la Biologie*. PhD Thesis, Université Joseph Fourier, Grenoble, France (1975). Available online at: https://tel.archives-ouvertes.fr/tel-00286222 (Accessed on January 5, 2025).

27. Demongeot J. Sur la possibilité de considérer le code génétique comme un code à enchaînement. *Rev Biomaths*. (1978) 62:61–6.

28. Demongeot J, Besson J. Code génétique et codes à enchaînement I. *CR Acad Sc III*. (1983) 296:807–10.

29. Demongeot J, Besson J. Genetic code and cyclic codes II. *CR Acad Sc III*. (1996) 319:520–8.

30. Weil G, Heus K, Faraut T, Demongeot J. An archetypal basic code for the primitive genome. *Theoret Comp Sc*. (2004) 322:313–34. doi: 10.1016/j.tcs.2004.03.015

31. Demongeot J, Moreira A. A circular RNA at the origin of life. *J Theor Biol*. (2007) 249:314–24. doi: 10.1016/j.jtbi.2007.07.010

32. Demongeot J. Traces of a primitive RNA ring in current genomes. *Biology*. (2025) 14:538. doi: 10.3390/biology14050538

33. Edous M, Eidous O. A simple approximation for normal distribution function. *Math Stat*. (2018) 6:47–9. doi: 10.13189/ms.2018.060401

34. Aherfi S, Colson P, LaScola B, Raoult D. Giant viruses of amoebas: an update. *Front Microbiol*. (2016) 7:349. doi: 10.3389/fmicb.2016.00349

35. Saneda TM, Field M. *Biological Anthropology: A Brief Introduction*. Open WA, Bothell WA: Cascadia College Pressbooks (2022).

36. Chimento NR, Agnolin FL. Phylogenetic tree of Litopterna and Perissodactyla indicates a complex early history of hoofed mammals. *Sci Rep*. (2020) 10:13280. doi: 10.1038/s41598-020-70287-5

37. Carranza J, Pérez-Barbería FJ. Sexual selection and senescence: male size-dimorphic ungulates evolved relatively smaller molars than females. *Am Nat*. (2007) 170:370–80. doi: 10.1086/519852

38. Zaia DA, Zaia CT, De Santana H. Which amino acids should be used in prebiotic chemistry studies? *Orig Life Evol Biosph*. (2008) 38:469–88. doi: 10.1007/s11084-008-9150-5

39. Robinson R. Jump-starting a cellular world: investigating the origin of life, from soup to networks. *PLoS Biol*. (2005) 3:e396. doi: 10.1371/journal.pbio.0030396

40. Seligmann H, Raoult D. Unifying view of stem-loop hairpin RNA as origin of current and ancient parasitic and non-parasitic RNAs, including in giant viruses. *Curr Opin Microbiol*. (2016) 31:1–8. doi: 10.1016/j.mib.2015.11.004

41. Muller HJ. The gene as the basis of life. In: Duggar BM editor. *Proceedings of the International Congress of Plant Sciences*. Ithaca, NY 1926, Menasha: Banta WI, (1929). pp. 897–921.

42. Maturana HR, Varela FJ. *Autopoiesis and Cognition: The Realization of the Living*. Boston MA: Reidel (1980). doi: 10.1007/978-94-009-8947-4

43. Bourgine P, Stewart J. Autopoiesis and cognition. *Artif Life*. (2004) 10:327–45. doi: 10.1162/1064546041255557

44. Ono N, Ikegami T. Self-maintenance and self-reproduction in an abstract cell model. *J Theor Biol*. (2000) 206:243–53. doi: 10.1006/jtbi.2000.2121

45. Ono N, Ikegami T. Artificial chemistry: computational studies on the emergence of self-reproducing units. In: Kelemen J, Sosik S, editors. *Proceedings of the 6th European conference on Artificial Life (ECAL'01)*. Prague, Czech Republic, September 2001. Berlin, Germany: Springer (2001). pp. 186–95. doi: 10.1007/3-540-44811-X_20

46. Tamura K, Schimmel P. Chiral-selective aminoacylation of an RNA minihelix. *Science*. (2004) 305:1253. doi: 10.1126/science.1099141

47. Tamura K, Schimmel P. Chiral-selective aminoacylation of an RNA minihelix: Mechanistic features and chiral suppression. *Proc Natl Acad Sci USA*. (2006) 103:13750–2. doi: 10.1073/pnas.0606070103

48. Beringer M, Rodnina MV. Importance of tRNA interactions with 23S rRNA for peptide bond formation on the ribosome: studies with substrate analogs. *Biol Chem*. (2007) 388:687–91. doi: 10.1515/BC.2007.077

49. Koonin EV, Novozhilov AS. Origin and evolution of the genetic code: the universal enigma. *Life*. (2009) 61:99–111. doi: 10.1002/iub.146

50. Rodin AS, Szathmáry E, Rodin SN. On origin of genetic code and tRNA before translation. *Biol Direct*. (2011) 6:14. doi: 10.1186/1745-6150-6-14

51. Seligmann H. Protein sequences recapitulate genetic code evolution. *Comput Struct Biotechnol J*. (2018) 16:177–89. doi: 10.1016/j.csbj.2018.05.001

52. Fei H, Li Y, Liu Y, Wei J, Chen A, Gao C. Advancing protein evolution with inverse folding models integrating structural and evolutionary constraints. *Cell*. (2025) 188:4674–92. doi: 10.1016/j.cell.2025.06.014

53. Ben Khalfallah H, Jelassi M, Demongeot J, Bellamine Ben Saouda N. Advancements in predictive analytics: machine learning approaches to estimate length of stay and mortality in sepsis. *Computation*. (2025) 13:8. doi: 10.3390/computation13010008

54. Bintrim S, Donohue T, Handelsman J, Roberts G, Goodman R. Molecular phylogeny of Archaea from soil. *Proc Natl Acad Sci USA*. (1997) 94:277–82. doi: 10.1073/pnas.94.1.277

55. Maxwell (2025). Available online at: https://gitlab.com/Orange-OpenSource/documentare?filter=maxwell (Accessed May 23, 2025).