Partitional Classification: A Complement to Phylogeny

Marc Salomon, Bruno Dassy

Partitional Classification: A Complement to Phylogeny

Marc Salomon and Bruno Dassy

Sorbonne Universités, UPMC Université Paris 06, Faculté de Biologie, Paris, France.

ABSTRACT: The tree of life is currently an active object of research, though next to vertical gene transmission non vertical gene transfers proved to play a significant role in the evolutionary process. To overcome this difficulty, trees of life are now constructed from genes hypothesized vital, on the assumption that these are all transmitted vertically. This view has been challenged. As a frame for this discussion, we developed a partitional taxonomical system clustering taxa at a high taxonomical rank. Our analysis (1) selects RNase P RNA sequences of bacterial, archaeal, and eucaryal genera from genetic databases, (2) submits the sequences, aligned, to *k*-medoid analysis to obtain clusters, (3) establishes the correspondence between clusters and taxa, (4) constructs from the taxa a new type of taxon, the genetic community (GC), and (5) classifies the GCs: Archaea–Eukaryotes contrastingly different from the six others, all bacterial. The GCs would be the broadest frame to carry out the phylogenies.

KEYWORDS: bioinformatics, classification, evolution, *k*-medoid analysis, cluster analysis, RNase P RNA

CITATION: Salomon and Dassy. Partitional Classification: A Complement to Phylogeny. *Evolutionary Bioinformatics* 2016:12 149–156 doi: 10.4137/EBO.S38288.

TYPE: Original Research

RECEIVED: December 24, 2015. **RESUBMITTED:** March 21, 2016. **ACCEPTED FOR PUBLICATION:** March 27, 2016.

ACADEMIC EDITOR: Jike Cui, Deputy Editor in Chief

PEER REVIEW: Two peer reviewers contributed to the peer review report. Reviewers' reports totaled 676 words, excluding any confidential comments to the academic editor.

FUNDING: Authors disclose no external funding sources.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: marc.salomon@upmc.fr

Introduction

Partitional cluster analyses (PCAs) constitute a diverse body of methods.[1,2] To our knowledge, very few taxonomic studies used PCAs, though these methods were recommended for the classification of organisms by a number of their founders.[3,4] The reason for this lies in one of the ideals of evolutionary biology, ie, to unravel the history of living beings in the form of a single phylogenetical tree, the tree of life (TOL), and simultaneously to classify them, the two activities hypothesized inseparable. In fact, each proposed TOL is a tree organizing certain taxa of the three domains of life,[5] based not necessarily on the same molecules or other characters, thus not congruent from one to another author.[5–7]

Another case against TOL is nonvertical gene transfer, namely lateral gene transfer (LGT), endosymbiosis, and chimerism. LGTs have been known since the end of the 1970s but were considered significant in the evolutionary process much later.[8] Endosymbiosis and chimerism are also invoked to explain the occurrence of main evolutionary events (eg, the emergence of Eucarya). Mitochondriae were shown to evolve from Alphaproteobacteria, chloroplasts from Cyanobacteria, and nuclei at least partially from Archaea.[9,10] In a eukaryotic cell, exchanges of material between organellar and nucleic DNA occur, a phenomenon called chimerism.

The different origins of gene acquisition launched a debate about the method to classify the Living World. Most authors have persisted to construct TOLs from genes hypothesized unaffected by LGT – the core genes –[11,12] mainly involved in transcriptional and/or translational mechanisms. Currently, strenuous efforts are made to combine the different published phylogenetic trees, taxonomical tools, and open bioinformatic systems to approach a comprehensive TOL.[13] But others criticized the phylogenetic method more deeply, arguing that LGT is still involved in a number of informational genes, and called for other representations.[14,15]

Without interfering in this discussion, we propose to construct a taxonomy based on degree of identity (DI) rather than degree of relationship. We defined the DI between two taxa as the overall distance calculated on evolutionary traits stemming both from gene vertical transmission and nonvertical transfers. The DIs were computed on the aligned DNA sequences coding for the RNA of RNase P – a universal ribozyme involved in the maturation of the tRNAs by cleaving its 5′ extremity. RNase P is an endonuclease generally comprising one RNA and a variable number of protein subunits – 1 in bacteria, 4–5 in archaea, and 8–10 in eukaryotes.[16,17] Except in the plants studied[18] and the mitochondrion of man[19] where the RNA is absent, the latter is generally the catalytic part and is widespread in a large number of taxa across the three domains.

RNase P RNA contains highly conserved regions, ie, the catalytic domain forming loops or hairpins, and highly variable regions linking them, hence the relevance of the choice of this molecule for classification. Compared with 16S–18S rRNA, RNase P RNAs are smaller sequences leading to comparable results with far less machine time. A higher rate of nucleotide

variation explains some discrepancies between the phylogenies performed with one or the other molecule.[7,20]

Methods

The material. Our initial material consisted of 564 DNA sequences coding for complete RNase P RNAs, carried by 564 different taxa (genera) and pooled together from three genetic databases, ie, Rfam, Noncode and GeneBank. The sequences obtained from Rfam originated from several built-in files where they were already displayed aligned, but this alignment was useless to us since it was performed within each file; the lengths of the sequences were different from one to another file. This length difference was even increased with the addition of the unaligned sequences coming from Noncode and GenBank. Besides, this raw material was heterogeneous concerning the presence and absence of gaps, since they were an admixture of aligned and unaligned sequences. The 564 sequences were then sorted in such a way that the *n*th sequence corresponded to the *n*th item of Dataset3.text – the file of the carriers of the sequences (cf. below). The sequences were gathered into file Dataset0.txt, whose sequences were thereafter ridden of their contingent gaps and multiply aligned (with MUSCLE[21] and Algorithms S1–S3, Figs. S1–S3 – algorithms, pieces of text, tables and figures referred to with 'S+a number' in supplementary file SupplementaryMaterial.pdf) this file and datasets 0–9 are referred to in the Supplementary Material section. The sequences resulting from these modifications were of equal length (2059 characters) and constituted file Dataset1.text. They were then numerized by an appropriate codification (Algorithm S4) and changed into numeric vectors of equal length composed of 8236 numerals, either 0 or 9 (Fig. S4). These vectors composed file Dataset2.text and were the objects on which our PCA was applied. We will now proceed to the analyses (see below).

The analyses. Our analyses developed into the following three steps: (1) a *k*-medoid analysis revealing a number of clusters among which the sampled sequences were distributed, (2) the study of the overlap between the clusters and operational taxonomic units (OTUs), and (3) a hierarchical clustering of the clusters assimilated to the OTUs, from which we derived a typology of cluster families, very strongly overlapping reunions of OTUs, ie, the genetic communities (GCs).

The genera and their taxonomic position. The taxonomic position (TP) of a given genus was defined as a sequence of nesting taxa in decreasing ranks, ie, domain, kingdom (for eukaryotes only), phylum, class, and order, each containing the genus. This information is easily available in taxonomic databases and in the literature. File Dataset3.txt contains 564 genera and their TP (the rows). Each genus corresponding to the *n*th row of Dataset3.txt is the carrier of the sequence corresponding to the *n*th row of Dataset2.txt (for more detail, see SupplementaryMaterial.pdf, section S1).

k-medoid analysis on our data. We carried out a *k*-medoid analysis on file Dataset2.txt with the following parameters: (1) d = Manhattan distance, (2) n = number of sequences, (3) k_0 = number of clusters optimizing the partition, (4) M = method = either clustering large applications (CLARA) or partitioning around medoids (PAM) (we performed both analyses), and (5) in case, we applied CLARA, N = number of samples to be drawn for CLARA = 100 (subsection S3.3). Number k_0 was obtained with Mardia's cluster variation method, ie, $k_0 = \text{int}\left(\sqrt{\frac{n}{2}}\right)$.[22]

The analyses (Algorithms S5 launching CLARA and S7 launching PAM) resulted in (1) the construction of the clusters around their medoids, (2) the assignment of the genera to each of the clusters, and (3) the computation of the cluster means. The k_0 clusters formed our cluster partition $\{C_i\}_{i\in\{1,2,\ldots,k_0\}}$. This analysis constructed an optimal partition of clusters gathering the most similar genera.

Contingency table crossing clusters and taxa. The genera were distributed among the k taxa T_i and k_0 clusters C_j, crossed to form a contingency table (CT) – with n_{ij} representing the number of genera within T_i and C_j.

- Per taxon T_i, $C_{\max}$ is the cluster containing the largest number of genera; n_i. the number of genera; and δ_i, the degree of membership to a cluster (DMTC) defined as the percentage of the taxon within $C_{\max}$ $\left(\delta_i = \frac{\max\limits_{1\le j\le k_0}(n_{ij})}{n_i}\times 100\right)$.
- Per cluster C_j, $T_{\max}$ is the taxon containing the largest number of genera; n_{ij} the number of genera; τ_j, the taxonomic specificity (TS) defined as the percentage of the genera of the cluster within $T_{\max}$ within C_j $\left(\tau_j = \frac{\max\limits_{1\le j\le k}(n_{ij})}{n_{.j}}\times 100\right)$.
- $n = \sum_{i=1}^{k} n_{i.} = \sum_{j=1}^{k_0} n_{.j}$.

Such a CT is illustrated in Table 1.

Definite and indefinite taxa. Each taxon of a TP is a definite taxon, ie, corresponding to an acknowledged taxonomical category. We considered these taxa as mathematical sets of genera; the reunions of the most similar taxa of a TP not corresponding to an officially defined taxonomical category were the indefinite taxa.

Functional biological units and OTUs. A functional biological unit (FBU) is a definite or indefinite taxon with a given set of known evolutionary characters and useful for the construction of the OTUs.[23–26] An OTU is an FBU that has the requirement appropriate to a given study – in our case a strong overlap with the clusters. The idea is to verify whether the clusters strongly match the OTUs, so that a typology of the clusters can be assimilated to a partitional taxonomy of the OTUs.

Table 1. The OTUs crossed with the 17 clusters.

	CLUSTERS																		
OTUs	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	C_{17}	$n_{i.}$	δ_i
A	**38**	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	95
At	0	1	0	0	1	**33**	5	2	1	0	0	0	0	0	0	0	0	43	77
Ba1	0	**11**	1	0	2	0	0	0	1	0	0	0	0	0	0	0	0	15	73
FL	0	0	1	0	0	0	0	0	0	0	**10**	0	0	0	0	0	0	11	91
Cy	0	0	0	0	0	0	0	0	0	1	0	**18**	0	0	0	0	0	19	95
Co1	5	3	**6**	0	3	0	0	2	0	2	0	0	2	0	0	0	0	23	26
Ng	0	0	0	0	0	0	0	0	0	**7**	0	0	0	0	0	0	0	7	100
Al1	0	**4**	1	3	1	0	0	1	0	0	0	0	0	1	0	0	0	11	36
Al2	0	5	0	**26**	1	0	0	0	0	0	0	0	0	1	0	0	0	33	79
Al3	0	0	0	5	1	0	0	0	0	0	0	0	0	**18**	0	0	0	24	75
Bu	0	0	7	0	1	0	0	0	0	0	0	0	0	0	1	**9**	0	18	50
Ga1	0	4	0	1	**19**	0	0	0	5	0	0	0	0	0	14	0	0	43	44
Ga2	0	**17**	0	0	7	0	0	0	6	0	0	0	0	0	0	0	0	30	57
E1	9	0	6	0	1	0	0	0	0	0	0	0	**37**	0	0	0	4	57	65
E2	3	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	**22**	27	82
E3	0	**2**	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	5	40
$n_{.j}$	55	48	23	35	37	33	5	5	13	10	10	19	41	20	16	10	26		
τ_j	69	35	30	74	51	100	100	40	46	70	100	95	90	90	88	90	85		

Note: The boldfaced numbers correspond to the intersection of each OTU with its C_{max} and represent the number of genera within the OTU and C_{max} in question.

We started our partitional classification analysis with κ initial FBUs (IFBUs) and constructed the OTUs in two step: (1) from the IFBUs to $\kappa'(<\kappa)$ larger FBUs (LFBUs) and (2) from the LFBUs to $\kappa''(<\kappa')$ OTUs. The IFBUs, LFBUs, and OTUs were crossed with the k_0 clusters to build up CTs.

Taxonomic interpretation of the clusters. Two independent analyses applied on the LFBUs were carried out to interpret the clusters taxonomically: (1) a statistical one based on an overlap index (OI) and (2) a correspondence analysis (CA).[22,27]

Statistical method based on overlap index. Three overlap indices between any taxon T_i and cluster C_j were proposed, and the best one among them selected: (1) $\omega_{ij} = \frac{2n_{ij}}{n_{i.} + n_{.j}}$ (Dice index), (2) $\omega_{ij} = \frac{n_{ij}}{n_{i.} + n_{.j} - n_{ij}}$ (Jaccard index), and (3) $\omega_{ij} = \frac{n_{ij}}{\sqrt{n_{i.}} \times \sqrt{n_{.j}}}$ (cosine index).[28]

Dice, Jaccard, and cosine OIs were calculated between κ' LFBUs and k_0 clusters. Of each LFBU T_i, the maximal OI (MOI) defined as $\omega_{i.} = \max_{1 \le j \le k_0} (\omega_{ij})$ was computed. This number describes the overlap between LFBU T_i and its C_{max} and reflects, if above a threshold ω_{inf} determined statistically, a specific association between C_{max} – necessarily one of the C_js – and LFBU T_i, the last being a revealed OTU. We selected the best OI (with the strongest MOI) for partitional classification.

Correspondence analysis. CA was carried out with Algorithm S8 from a CT crossing κ' LFBUs (Dataset7.txt) with the k_0 clusters.

A hierarchical cluster analysis to infer the partitional classification. Algorithm S9 performed a hierarchical cluster analysis (HCA)[1] on the means of the taxon specific clusters and the mean of cluster C2 obtained with Algorithm S6 with (1) Manhattan distance as the dissimilarity index and (2) Ward as the agglomerative method. These means were identified to the OTUs. We considered as taxon-specific, clusters having 7+ members and a TS ≥ 50. The numerous cluster C_2 was also processed despite its low TS since it showed an interesting bimodal distribution. If the analysis showed that these clusters could be assimilated to OTUs, the inferred cluster typology would be equivalent to a taxonomic system of the OTUs based on the DIs. In this system, we gather the most similar clusters into cluster families (CFs) assimilated to the reunions of the OTUs showing the highest DIs. Such reunions of taxa were called GCs.

Abbreviated taxon names. A = Archaea, Ab = Acidithiobacillales, Ac = Actinopterygii, Ae = Aves = Ae1 ∪ Ae2, Ae1 = *Taenopygia*, Ae2 = *Gallus*, AE = Archaea or Eucarya = A ∪ E, Af = Afrosoricida, Ai = Ascidiaceae, Al2 = α-Proteobacteria 2 = Rz ∪ Ro ∪ Sh, Al3 = α-Proteobacteria 3 = Rh ∪ Mg, Am = Aeromonadales, An = Alteromonadales, Ar = Arthropoda, AT = Actinobacteria, Av = Alveolata, Ay = Artiodactyla, Ba1 = Bacteroidetes 1 = BT ∪ CT ∪ SB, BT = Bacteroidia, Bu = Burkholderiales, Ch = Chiroptera, Ci = Cnidaria,

Cm=Chromatiales, Cn=Carnivora, Co1=Clostridia 1=Cs∪Se, Cp = Cephalochordata, Cs = Clostridiales, CT = Cytophagia, Cy = Cyanobacteria, Dd = Didelphimorpha, Dp = Diprodontia, E1 = Eucarya 1 = Ac∪Av∪Ex∪Ho∪Ae1∪Ai∪Ar∪Cp∪Fn∪Ma1∪Ml∪Ne∪Pl∪Pt, E2 = Eucarya 2 = Ma2∪Ae2, E3 = GL ∪ Hr ∪ Ec ∪ Ci, Ec = Echinodermata, En = Enterobacteriales, Ex = Excavata, FL = Flavobacteria, Fn = Fungi, Ga1 = γ-Proteobacteria 1 = Ab ∪ Am ∪ An ∪ En ∪ Ps ∪ Vi ∪ Xa, Ga2 = γ-Proteobacteria 2 = Cd ∪ Cm ∪ Gais ∪ Lg ∪ Mc ∪ Oc ∪ Pd ∪ Tt, Gais = γ-Proteobacteria incertia sedis, GL = Glaucophyta, Ho = Choanomonada, Hr = Chromalveolata, Hy = Hyracoidia, La = Lagomorpha, Lg = Legionellales, Ma1 = Mammalia 1 = Af ∪ Ay ∪ Dp ∪ La ∪ Rd ∪ Sc ∪ Ty, Ma2 = Mammalia 2 = Cn ∪ Ch ∪ Dd ∪ Hy ∪ Pe ∪ Mo, Mc = Methylococcales, Mg = Magnetococcales, Ml = Mollusca, Mo = Monotremata, Ne = Nematoda, Oc = Oceanospirillales, Pd = Pseudomonadales, Pe = Perissodactyla, Pl = Placozoa, Ps = Pasteurellales, Pt = 'Platyhel-mynthes, Rd = Rodentia, Rh = Rhodobacterales, Ri = Rickettsiales, Ro = Rhodospirillales, Rz = Rhizobiales, SB = Sphingobacteria, Sc = Scandentia, Se = Selenomonadales, Sh = Sphingomonadales, Tt = Thiotrichales, Ty = Tylopodes, Vi = Vibrionales, Xa = Xanthomonadales.

Results

Relevant taxa. *Results from the k-medoid analysis.* Our data showed that the optimal number of clusters, obtained with Mardia's cluster variation method, was $k_0 = 17$. Our *k*-medoid analysis, carried out with method CLARA, resulted in (1) the assignment of each of the 564 genera to one of the 17 clusters (Dataset4.txt) and (2) the computation of the mean vectors of the 17 clusters (Dataset5.txt). We performed the same analysis with method PAM and obtained an almost identical assignment of the genera to the 17 clusters (Dataset6.txt) except for four among the 564 genera (*Sebaldella*, *Liberibacter*, *Novosphingobium*, and *Nautilia*). We decided to proceed to the analyses with CLARA (see Discussion section).

The three successive CTs. The 564 genera were distributed into three successive CTs – taxa crossed with the same $k_0 = 17$ clusters:

- A CT involving $\kappa = 100$ IFBUs (Table S1).
- A CT on $\kappa' = 33$ LFBUs (Table S2). Each of these taxa is the reunion of IFBUs included in the same taxon of the immediate higher rank (TIHR) as displayed in Dataset3.txt and shares the same C_{max}. For example, LFBU *Archaea* (*A*) is the reunion of IFBUs *Crenarchaeota* (*Cr*), *Euryarchaeota* (*Er*), *Korarchaeota* (*Kr*), and *Thaumarchaeota* (*TH*); the genera of the member IFBUs of *A* overwhelmingly belong to cluster $C_{max} = C_1$.
- A CT involving $\kappa'' = 16$ OTUs (Table 1). The OTUs are heuristically defined as (1) LFBUs having a DMTC ≥ 50 and represented by seven or more genera and (2) LFBUs belonging to the same TIHR as other member OTUs, ie, Al1, Ga1, and E3 (69.3% of the sampled genera).

Correspondence between cluster groups and taxa. *With the statistical method based on the OIs.* Tables S3–S5 display the overlap between the LFBUs and the clusters – assessed respectively with Dice, Jaccard, and cosine indices (MOIs ω_i in right margin of the tables). From these tables, we calculated (1) $\bar{o}$ and $\hat{\sigma}_O$, respectively, mean and standard deviation of random variable *O* taking on values ω_i and (2) threshold $\omega_{inf} = \bar{\omega} - 1.65 \times \frac{\hat{\sigma}_O}{\kappa}$ (kappa being the number of the taxa involved) after normality of the ω_is was verified (Table 2). Each LFBU with a $\omega_i \geq \omega_{inf}$ was considered as significantly superimposed to its C_{max} cluster, which we called its corresponding cluster. (We called these LFBUs candidate OTUs.)

Tables S6–S8 present the three OIs between the candidate OTUs and the clusters. The cosine index was the best OI, with the highest mean MOI and largest number of candidate OTUs above ω_{inf} (cf. Table 2) and, thus, chosen as our OI for the rest of the study.

From the CA. We applied the CA to Dataset7.txt and obtained file Dataset8.txt, the listing of the analysis, from which we plotted CA diagrams (Fig. 1). The results of the analysis, ie, the relationships between cluster and OTU as revealed by the CA, are reported in Table 3.

Both methods give the same association between cluster and taxon (Table 3). Remarkably, (1) the associated clusters unveiled by CA are the C_{max}s of the descriptive method and (2) the taxa revealed as overlapping the clusters were all OTUs as determined in the previous subsection. A solid underpinning between clusters and OTU is thus highlighted. The clusters are identified to OTUs.

The DIs revealed by HCAs. The HCA on the cluster means (Dataset5.txt), restricted to the taxon-specific clusters, calculated the distances between them, organized these distances in a distance matrix (Dataset9.txt), and drew from the latter our dendrogram (Fig. 2). We associated the CFs with their corresponding GCs (cf. Table 4).

The dendrogram plot highlighted a typology with seven cluster families (CFs): $\mathcal{A}_1 = \{C_1, C_{13}, C_{17}\}$; $\mathcal{A}_2 = \{C_2, C_5, C_{10}\}$; $\mathcal{A}_3 = \{C_4, C_6, C_{14}\}$; $\mathcal{A}_4 = \{C_{11}\}$; $\mathcal{A}_5 = \{C_{12}\}$; $\mathcal{A}_6 = \{C_{15}\}$; and $\mathcal{A}_7 = \{C_{16}\}$.

Table 2. Comparison of statistic descriptors of variable *O* for the three OIs (analysis on the LFBUs).

OI TYPE	TABLE	$\bar{o}$ (MOI)	$\hat{\sigma}_O$	JB	ω_{inf}	NUMBER OF TAXA WITH $\omega_i > \omega_{inf}$
Dice index	S3	0.54	0.28	ns	0.46	11
Jaccard index	S4	0.42	0.26	ns	0.34	10
Cosine index	S5	0.56	0.26	ns	0.49	11

Abbreviations: JB, Jarque–Bera normality test statistic[49]; $\bar{o}$, MOIs; ns, nonsignificant.

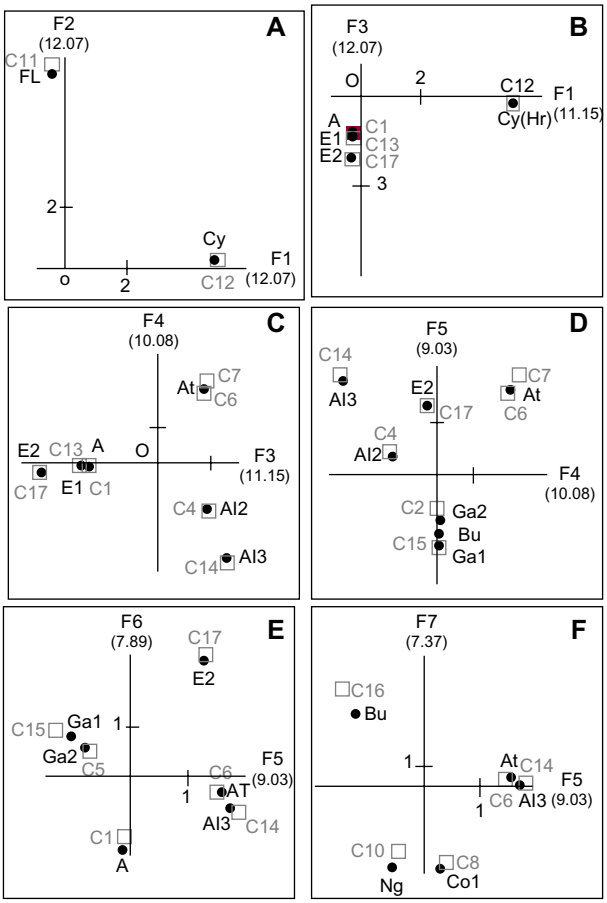

Figure 1. Plot diagrams inferred by CA. Inertia rates in brackets next to the factorial axes (FAs). Squares with C_n in gray are clusters. Dots with abbreviated names in black are taxa. Factorial planes generated by two factorial axes: (**A**) F1 and F2; (**B**) F1 and F3; (**C**) F3 and F4; (**D**) F4 and F5; (**E**) F5 and F6; (**F**) F5 and F7.

We identified each CF with a *potential GC* (PGC) defined as the reunion of the OTUs corresponding to the clusters composing the CF. For example, to C_1 corresponds Archaea, to C_4 Eucarya 1, and to C_{17} Eucarya 2. Hence, to $\mathcal{A}_1$, we could identify the PGC obtained by reuniting these three taxa. We considered the typology displayed in Table 4 to be good because $\widetilde{OI} > \omega_{inf}$ (calculated from the data of Table 4). The PGCs with $MOI \geq \omega_{inf}$, boldfaced, were defined as GCs. Hence, the GCs are (1) the Archaea and Eukaryotes altogether (AE), (2) the Burkholderiales (Bu), (3) Bacteroidetes 1 (Ba1), (4) the Cyanobacteria (Cy), (5) the γ-Proteobacteria (Ga), (6) the α-Proteobacteria (Al), and (7) the Actinobacteria (AT). The genera processed numbered 333, accounting thus for 59% of sample *S*.

Discussion

Justification of our methological principle. LGT and endosymbiosis may have played a key role in the emergence of new groups in certain circumstances (such as, after massive extinctions or radical changes in their environment). These events could have introduced novelties in organisms,

Table 3. Comparative results between the OIA and CA.

OIA			CA	
OTUs	C_{max}	MOIs	ASSOCIATED CLUSTERS	FACTORIAL PLANES
FL	C_{11}	0.95	C_{11}	F1 × F2
Cy	C_{12}	0.95	C_{12}	F1 × F2
E2	C_{17}	0.83	C_{17}	F1 × F3
Al3	C_{14}	0.82	C_{14}	F3 × F4
At	C_6 (C_7)	0.78	C_6, C_7	F3 × F4
E1	C_{13}	0.78	C_{13}	F1 × F3
A	C_1	0.76	C_1	F1 × F3
Co1	C_8	0.73	C_8	F5 × F7
Al2	C_4	0.72	C_4	F3 × F4
Bu	C_{16} (C_3)	0.67	C_{16}	F4 × F5
Ga2	C_{15} (C_9)	0.53	C_{15}	F4 × F5
Ng	C_{10}	0.45	C_{10}	F5 × F7
Ga1	C_2 (C_9)	0.34	C_2	F4 × F5
Ba1	C_2	0.32	C_2	F1 × F12

Notes: OTUs sorted in decreasing order of MOI. Clusters in parentheses, in OIA, cluster with the second largest number of genera for a given taxon.
Abbreviations: OIA, overlap index analysis; CA, correspondence analysis.

shared thereafter by their descendants via classical vertical gene transmission if these gene acquisitions conferred to the bearers increased selective advantages.[10,29] Hence, entire historical communities could have emerged this way, introducing evolutionary discontinuities, possibly the GCs. We propose that phylogenies could be unraveled within the GCs.

The construction of the TOL implicitly accepts the hypothesis of the constancy of the molecular clock – at least

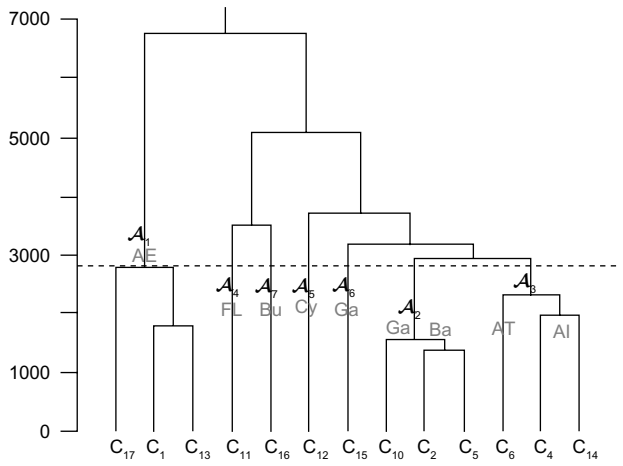

Figure 2. HCA dendrogram. Distance = DI = Manhattan; aggregation method = Ward. Cut at distance ca. 3200. $\mathcal{A}_i$ = cluster families: $\mathcal{A}_1 = \{C_1, C_{13}, C_{17}\}$; $\mathcal{A}_2 = \{C_2, C_5, C_{10}\}$; $\mathcal{A}_3 = \{C_4, C_6, C_{14}\}$; $\mathcal{A}_4 = \{C_{11}\}$; $\mathcal{A}_5 = \{C_{12}\}$; $\mathcal{A}_6 = \{C_{15}\}$; and $\mathcal{A}_7 = \{C_{16}\}$.

Table 4. Cluster families inferred by the HCA of the TSCs.

	CLUSTER FAMILIES $\mathcal{A}_j$								
T_i	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_3$	$\mathcal{A}_4$	$\mathcal{A}_5$	$\mathcal{A}_6$	$\mathcal{A}_7$	$n_{i.}$	$\omega_{i.}$
AE	115 (0.99)	2 (0.02)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	117	0.99
Ga	0 (0)	47 (0.68)	1 (0.01)	0 (0)	0 (0)	14 (0.46)	0 (0)	62	0.68
Ba1	0 (0)	13 (0.41)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	13	0.41
AI	0 (0)	12 (0.17)	54 (0.71)	0 (0)	0 (0)	0 (0)	0 (0)	66	0.71
AT	0 (0)	2 (0.04)	33 (0.59)	0 (0)	0 (0)	0 (0)	0 (0)	35	0.59
FL	0 (0)	0 (0)	0 (0)	10 (1)	0 (0)	0 (0)	0 (0)	10	1
Cy	0 (0)	1 (0.03)	0 (0)	0 (0)	18 (0.97)	0 (0)	0 (0)	19	0.97
Bu	0 (0)	1 (0.03)	0 (0)	0 (0)	0 (0)	1 (0.08)	9 (0.90)	11	0.90
$n_{.j}$	116	12	22	19	68	57	61	355	

Notes: T_i, PGCs. At the intersection of T_i and A_j: n_{ij} = number of genera in taxon T_i and cluster family A_j; $n_{i.}$ number of genera in taxon in T_i, and $n_{.j}$ number of genera in taxon in A_j. In brackets, OI between PGCs and CFs. $\omega_{i.}$ = MOI of T_i. Mean of MOI = $\widetilde{\text{OI}}$ = 0.81. From calculation, ω_{inf} = 0.69.

stochastically – throughout the geological eras, within the organisms classified. However, it has been shown that in the remote past, radiation rates coupled with atmosphere composition varied, entailing a variation of the rate of molecular evolution between the taxa.[30–32]

TOLs based on core genes might trace back the phylogenies of only parts of organisms, if these are phylogenetically too distant. The aim of a sound taxonomical system being the objective comparison of whole organisms, we suggest to carry out phylogenetical taxonomy only on restricted groups where one can take nearly for granted that the overwhelming part of the genetical material has been acquired by vertical transmission, like for instance in the Metazoa or γ-Proteobacteria. Thus, we propose to apply partitional clustering mainly to higher ranked taxa and phylogenetical analyses principally on lower ranked taxa, when the molecular clock can be reasonably calibrated and the genes shown to be transmitted vertically.

One might object against partitional clustering that the latter is equivalent to rootless tree analyses, as used in previous studies.[33–35] In our opinion, the two approaches are distinct, and the main differences between them are as follows: (1) In a rootless phylogeny, one poses a hypothesis on the relationships between taxa of a given group, which would constitute a community of related taxa exclusively sharing a set of characters between themselves, hypothesized to be relevant for the group and supposed to be possessed by a common (unknown) ancestor. These characters are termed *polarized*. A rootless tree, like any tree, is a hypothetico-deductive construction. (2) On the contrary, partitional clustering is not based upon an a priori hypothesis. The global DI between the taxa is revealed by structures underpinning the data. This approach is inductive.

Our analysis revives the old-standing debate between the tenants of the deductive methods and those of the inductive methods in systematics and evolutionary biology.[36,37] Deductive methods have been favored for the last three decades, and inductive methods on the contrary hardly evoked. However, though the deductive methods have been extremely useful and fruitful in the explanation of many evolutionary phenomena, inductive methods can also deliver very interesting information.[38,39]

The choice of the k-medoid analysis. We chose to apply *k*-medoid analysis because contrary to *k*-mean and *k*-median analyses, it does not rely on means or medians, not appropriate to our data (binary numerals). In addition, *k*-medoid analyses are less influenced by outliers, and they are more robust than *k*-mean or *k*-median analyses, ie, their results depending less on the initial conditions (the choice of the first centroids).[40]

There are two methods for *k*-medoid analysis on a given sample, ie, PAM and CLARA.[41] PAM handles all the objects and is appropriate for relatively small samples. CLARA on the other hand selects, from a large sample, a series of randomly drawn subsamples. The seeds are selected in each subsample by means of a program similar to PAM; thereafter, the objects of the entire sample are assigned to each of these seeds by means of a chosen (either Euclidian or Mahattan) index distance. CLARA is best suited for large files since the complexity of this algorithm rises arithmetically and not exponentially like in *k*-mean and *k*-median analyses. This property makes it possible to process large samples of long sequences in a reasonable time period and in portable computers. We compared the two methods and found that among the 564 genera analyzed, only 4 genera were not assigned to the same cluster. Hence, for us, the methods are comparable, and we can use either method, perhaps with a preference for CLARA to minimize the complexity of the algorithm.

The GCs. Our analysis revealed taxa, ie, the GCs, overlapping the cluster families very significantly and gathering the most similar organisms, ie, the genera whose DI between themselves are smallest. This may be explained by the fact that mathematical clustering does not assemble the genera randomly. Organisms are hypercomplex systems highly constrained phenotypically, hence also genetically. This mere fact probably imposed on them a relatively small number of solutions for their structuration, reflected by the strong genetic resemblance of the organisms within a small number of sets.

Figure 2 shows that the nonbacterial organisms are genetically less differentiated than the bacterial ones, the largest between-cluster distance (LBCD) of $\mathcal{A}_1$ being about 2775 and of the reunion of the other cluster families ca 5080. Cluster family $\mathcal{A}_1$ is remarkable in the sense that Eucarya 2 (one of the avian and about half of the mammalian orders) is contained in

cluster C_{17}, which is more distant from cluster C_3 (comprising almost the rest of the Eukaryotes), than C_1 (the cluster gathering almost all the Archaea). One of the possible explanations lies in the acquisition of extra protein subunits partially involved in catalytic activity in the eucaryal genera, a situation that would have correlatively entailed a weaker involvement of the RNA subunit in that activity, and consequently a structural simplification of the latter. This could explain some structural convergence between very distinct groups of nonbacterial genera in CF $\mathcal{A}_1$.[16] A huge gap exists between the Eukaryotes and Archaea on one hand and the Bacteria on the other hand. The LBCD between these two groups is 6780. Thus, GC Archaea-Eucarya forms a consistent group, in opposition with the remaining GCs forming another and as consistent group of GCs, all bacterial. Remarkably, within this group, the GCs Burkholderiales ($\mathcal{A}_2$) and Cyanobacteria ($\mathcal{A}_4$) are more distant from their neighboring bacterial CFs than the nonbacterial clusters between themselves. Reversely, two composite GCs, the one containing most of the γ-Proteobacteria and Bacteroidetes 1 ($\mathcal{A}_2$) on one hand, and the other composed of the α-proteobacteria and the Actinobacteria ($\mathcal{A}_3$) on the other hand, are less diversified than the nonbacterial GC (respective LBCDs $\simeq$ 1590 and 2320).

A number of taxa of sample S are not members of any of the GCs, namely those which are scattered among the clusters with no preferential connections (and thus showing a weak DMTC), or those connected to clusters C_1, C_2, or C_7, which do not enter in the composition of the cluster families. Of the first category, one can mention Bacilli; and of the second, one can mention Clostridia 23 and Negativicutes, and δ- and ε-Proteobacteria. Such a result is not compatible with the systematics inferred from the phylogeny based on 16S/18S rRNA. However, the heterogeneity of the Firmicutes and the Proteobacteria highlighted by our analysis was also revealed in a number of phylogenetic studies on universal molecules other than 16S/18S rRNA,[20,42,43] inciting the authors to question the monophyly of these taxa.

Two biases can be encountered in classification based upon aligned sequences, namely the convergence of homologous blocks resulting from plesiomorphic sequence position,[42] and the compensatory base changes not necessarily leading to a phenotypic differentiation (in the case of noncoding RNA, no change in secondary structures).[44,45] But this remark is not only valid about our study but also to the vast majority of the current phylogenetic studies exclusively involving the primary structures.

The method is tributary to the sequences existing in the genetic databases. The material obtained have a strong influence on the optimization of k-medoid analysis, hence on k_0 – the optimal number of clusters – and consequently all the genera will not necessarily be processed. But this problem also exists in phylogenetic analysis, where a decision is always made concerning a hypothesis, necessarily concealing – in parts of a tree in construction – uncertainties or lack of knowledge.

Conclusion

The seven GCs would be the result of the plurality of the sources of genetic heritage that would render the history linking them blurred and tremendously difficult to unravel. The nonbacterial GC is distinct from all the other, bacterial, GCs taken altogether. And within the bacterial GCs, *surprisingly*, Actinobacteria have a relatively strong DI with α-proteobacteria, which again does not mean that α-Proteobacteria are more related to Actinobacteria, than they are to γ-Proteobacteria. The same holds for Burkholderiales – an order of β-Proteobacteria – which show a smaller DI with the Bacteroidetes (Flovobacteria) than with the other Proteobacteria. This shows that the dendrogram interpreting the DIs is not a phylogeny but add information to it, contributing hopefully to the construction of a taxonomy at the highest ranks, when all cellular organisms are compared, perhaps more based on partitional than purely phylogenetical reasoning. Interestingly, each GC is genetically so consistent that this does not seem fortuitous. It appeared to us very likely that vertical gene transmission did play a great role in this internal coherence. Therefore, we propose that the seven GCs be the broadest frames for phylogenetic reconstructions.

At the highest rank, ie, that of the domain, our results are strikingly compatible with the three-partite division of the Living World present in the TOL of Woese et al[5]; Archaea-Eucarya is, also with our method, the sister group of all the remaining known cellular organisms, ie, bacteria, but at the same time, our proposition introduces an uncertainty principle in the search of the phylogenetic relationships between all of the cellular organisms. We based our analysis on a universal albeit single molecule, and further studies on other molecules or parts of the genome are needed to check consistency and thus validate the method. Some of the validating approaches, with appropriate modifications, could be applied to our method, eg, benchmarking.[46,47] We could compare our method with other classificatory systems, eg, the Cluster of Orthologous Groups of proteins for prokayotic or eukaryotics organisms (COG/KOG).[48]

Acknowledgments

We thank A. Carbone, T. Dagan, A.-L. Haenni, M. Jousselin-Hosaja, and S. Kruglik for their valuable comments and constructive discussions.

Author Contributions

Conceived and designed the experiments: MS, BD. Analyzed the data: MS. Wrote the first draft of the article: MS. Contributed to the writing of the article: MS, BD. Agreed with the article results and conclusions: MS, BD. Jointly developed the structure and arguments for the article: MS, BD. Made critical revisions and approved the final revisions: MS, BD. Both authors reviewed and approved the final article.

Supplementary Material

SupplementaryMaterial.pdf. The supplementary material (algorithms, figures, tables and texts) are gathered and described in this file.

Dataset0.txt – Dataset9.txt. Datasets 0–9, as described in SupplementaryMaterial.pdf, are included as separate files.

REFERENCES

1. Everitt BS, Landau S, Leese M. *Cluster Analysis*. London: Wiley; 2001.
2. Gan G, Chaoqun M, Wu J. *Data Clustering: Theory, Algorithms and Applications; 20 of Series on Statistics and Applied Probability*. Philadelphia, PA; Alexandria, VA: Siam Press; 2007.
3. Edwards AWF, Cavalli-Sforza LL. A method for cluster analysis. *Biometrics*. 1965;21:362–75.
4. Gower JC. A comparison of some methods of cluster analysis. *Biometrics*. 1967;23:623–37.
5. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms - proposal for the domains Archaea, Bacteria, and Eukarya. *Proc Natl Acad Sci U S A*. 1990;87:4576–9.
6. Lawson FS, Charlebois RL, Dillon JAR. Phylogenetic analysis of carbamoylphosphate synthetase genes: complex evolutionary history includes an internal duplication within a gene which can root the tree of life. *Mol Biol Evol*. 1996;13:970–7.
7. Sun FJ, Caetano-Anollés G. The ancient history of the structure of ribonuclease P and the early origins of Archaea. *BMC Bioinformatics*. 2010;11:153.
8. Smith MW, Feng DF, Doolittle RF. Evolution by acquisition – the case for horizontal gene transfers. *Trends Biochem Sci*. 1992;17:489–93.
9. Schwartz RM, Dayhoff MO. Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. *Science*. 1978;199:395–403.
10. Williams TA, Foster PG, Cox CJ, Embley TM. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*. 2013;504:231–6.
11. Brochier C, Bapteste E, Moreira D, Philippe H. Eubacterial phylogeny based on translational apparatus proteins. *Trends Genet*. 2002;18:1–5.
12. Gribaldo S, Poole AM, Daubin V, Forterre P, Brochier-Armanet C. The origin of eukaryotes and their relationship with the Archaea: are we at a phylogenomic impasse? *Nat Rev Microbiol*. 2010;8:743–52.
13. Hinchliff CE, Smith SA, Allman JF, et al. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proc Natl Acad Sci U S A*. 2015;112:12764–9.
14. Doolittle WF. Phylogenetic classification and the universal tree. *Science*. 1999;284:2124–8.
15. Lopez P, Bapteste E. Molecular phylogeny: reconstructing the forest. *C R Biol*. 2009;332:171–82.
16. Esakova O, Krasilnikov AS. Of proteins and RNA: the RNase P/MRP family. *RNA*. 2010;16:1725–47.
17. Mondragón A. Structural studies of RNase P. *Annu Rev Biophys*. 2013; 42:537–57.
18. Krehan M, Heubeck C, Menzel N, Seibel P, Schoen A. RNase MRP RNA and RNase P activity in plants are associated with a Pop1p containing complex. *Nucleic Acids Res*. 2012;40:7956–66.
19. Holzmann J, Frank P, Loeffler E, Bennett KL, Gerner C, Rossmanith W. RNase P without RNA: identification and functional reconstitution of the human mitochondrial tRNA processing enzyme. *Cell*. 2008;135:462–74.
20. Haas ES, Banta AB, Harris JK, Pace NR, Brown JW. Structure and evolution of ribonuclease P RNA in Gram-positive bacteria. *Nucleic Acids Res*. 1996;24:4775–82.
21. Gouy M, Guindon S, Gascuel O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol*. 2010;27:221–4.
22. Mardia KV, Kent JT, Bibby JM. *Multivariate Analysis*. London: Academic Press;1979.
23. Sokal RR, Sneath PHA. *Principles of Numerical Taxonomy*. San Francisco: Freeman; 1963.
24. Sneath PHA, Sokal RR. *Numerical Taxonomy. The Principles and Practice of Numerical Classification*. San Francisco: Freeman; 1973.
25. Sastre C. Paléoclimats, spéciation et taxonomie. Quelques exemples chez les Ochnacées néotropicales. *Mém Soc Biog* 3 *sér*. 1994;4:3–10.
26. Ness JH, Rollinson EJ, Whitney KD. Phylogenetic distance can predict susceptibility to attack by natural enemies. *Oikos*. 2011;120:1327–34.
27. Benzécri J-P. *L'analyse des données. T*2 *– L'analyse des correspondances*. Paris: Dunod; 1973.
28. Legendre L, Legendre P. *Ecologie numérique: Le traitement multiple des données cologiques. La structure des données écologiques*. Paris: Masson; 1984.
29. Schönknecht G, Weber APM, Lercher MJ. Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution. *Bioessays*. 2014;36:9–20.
30. Graur D, Martin W. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends Genet*. 2004;20:80–6.
31. Ho SYW, Lanfear R, Bromham L, et al. Time-dependent rates of molecular evolution. *Mol Ecol*. 2011;20:3087–101.
32. Lanfear R, Ho SYW, Love D, Bromham L. Mutation rate is linked to diversification in birds. *Proc Natl Acad Sci U S A*. 2010;95:9413–7.
33. Felsenstein J. *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates; 2004.
34. Gascuel O. *Mathematics of Evolution and Phylogeny*. Oxford: Oxford University Press; 2005.
35. Lapointe F-J, Lopez P, Boucher Y, Koenig J, Bapteste E. Clanistics: a multi-level perspective for harvesting unrooted gene trees. *Trends Microbiol*. 2010;18:341–7.
36. Lienau EK, DeSalle R. Evidence, content and corroboration and the Tree of Life. *Acta Biotheor*. 2009;57:187–99.
37. Schwartz JH. Reflections on systematics and phylogenetic reconstruction. *Acta Biotheor*. 2009;57:295–305.
38. Kell DB, Oliver SG. Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays*. 2004;26:99–105.
39. Guthery FS. Deductive and inductive methods of accumulating reliable knowledge in wildlife science. *J Wildl Manage*. 2007;71:222–5.
40. Sarada W. A review on clustering techniques and their comparison. *Int J Adv Res Comput Eng Technol*. 2013;2:2806–12.
41. Kaufman L, Rousseeuw PJ. *Finding Groups in Data. An Introduction to Cluster Analysis*. Hoboken: John Wiley & sons; 2005.
42. Ludwig W, Schleifer KH. Phylogeny of Bacteria beyond the 16S rRNA standard. *ASM News*. 1999;65:752–7.
43. Sutcliffe IC. A phylum level perspective on bacterial cell envelope architecture. *Trends Microbiol*. 2010;18:464–70.
44. Caetano-Anollés G. Evolved RNA secondary structure and the rooting of the universal tree of life. *J Mol Evol*. 2002;54:333–45.
45. Pace NR, Smith DK, Olsen GJ, James BD. Phylogenetic comparative analysis and the secondary structure of ribonuclease P RNA: a review. *Gene*. 1989;82: 65–75.
46. Löytynoja A. Alignment methods: strategies, challenges, benchmarking, and comparative overview. *Methods Mol Biol*. 2012;855:203–235.
47. Iantorno S, Gori K, Goldman N, Gil M, Dessimoz C. Who watches the watchmen? An appraisal of benchmarks for multiple sequence alignment. *Methods Mol Biol*. 2014;1079:59–73.
48. Tatusov RL, Fedorova ND, Jackson JD, et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*. 2003;4:41.
49. Jarque CM, Bera AK. A test for normality of observations and regression residuals. *Int Stat Rev*. 1987;55:163–72.