Chapter 1

Phylogenetics

Eliane Barbosa Evanovich dos Santos

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.79422

Abstract

Describing the diversity of living beings has always instigated man. The classification proposed by Aristotle today seems naïve and unnatural, but it lasted from ancient Greece until the publication of the Linnaeus *Systema Naturae* in 1758. Although quite accurate, the taxonomic classification proposed by naturalist Carl Linnaeus did not consider the evolutionary relationships between living beings. This view, although prior to Charles Darwin, only gained deserved prominence after *On the Origin of Species*. Only in the twentieth century, a new area founded by Hennig, phylogenetic systematics was implemented, and with this, a series of useful methods in the construction of phylogenetic trees arose, as maximum parsimony, neighbor joining, UPGMA, maximum likelihood, and Bayesian inference. With the advancement of information technology, phylogenetic analyses have become more sophisticated and faster. The algorithms used in the analysis programs have become more complex and realistic, favoring the addition of substitution models. The application of these data and the greater facility in generating nucleotide and amino acid sequences allowed the comparison previously unimaginable, for example, between bacteria and eukaryotes. In this way, the history of the advances of phylogenetic knowledge is confused with the greater knowledge about the origin of life.

Keywords: evolution, phylogenetic systematics, phylogenetic tree, taxonomy, phylogenetic methods

1. Introduction

Different criteria of biological classification were created throughout history. Some are arbitrary and do little to reflect the evolutionary relationship between species, for example, the Aristotelian system. But not always reflecting the relations of relatives was a concern. Even the iconic classification suggested by Linnaeus was not intended to reflect this relationship (although it is very consistent with current taxonomic classification). Only with Darwin and

his successors did common ancestry gain prominence and was accepted as a fundamental tool in taxonomic analysis through Hennig. The systematic phylogenetic title of the book of the German entomologist opened the door to a new way of looking at taxonomy through kinship relations. The proposal of this new taxonomy would, therefore, be an unequivocal way of understanding the evolutionary history of the species. We now know the various phylogenetic artifacts that may mask or hinder a robust phylogenetic hypothesis. But, computational advancement and new phylogenetic approaches are emerging, reducing the effects of these artifacts. This chapter makes a narrative review of the history and current advances in phylogeny. The analysis was conducted using PubMed (https://www.ncbi.nlm.nih.gov/pubmed/), Scopus (https://www.scopus.com), and Google Scholar (https://scholar.google.com/). The first part of the review describes succinctly the work of Anaximander, Aristotle, Carl Linnaeus, Peter Simon Pallas, Charles Darwin, and Willi Hennig; the second aspect is showing the phylogenetic methods and phylogenetic analysis programs, and the third focus presents the difference between gene tree and species and shows the criteria used in building of tree of life.

2. Phylogeny

> *"...the whole system of organic bodies may be well represented by the likeness of a tree that immediately from the root divides both the simplest plants and animals, [but they remain] variously contiguous as they advance up the trunk, Animals and Vegetables; those leading, from Mollusca advancing to Pisces, with great lateral branches of Insects sent out among themselves, from here to Amphibia; and at the extreme top of the tree the Quadrupeds are supported, Aves truly thrust out as an equally great lateral branch below the Quadrupeds. At the same time this image shows the animals to be neither continuous nor neighboring, but standing like a lone tree"* [1].

Biodiversity has always instigated man to explain its origin, define it, and classify it. The precursory attempts were from the Greeks Anaximander of Miletus (610–545 BC) and Aristotle (384–322 BC). Anaximander defended the proposal that living beings originated from water and underwent transformations over time [1]. The sun would be a catalyst for these changes and would have allowed the maturation and exit of a fish-like being from the water, giving rise to more complex creatures such as man [2, 3]. Aristotle developed one of the first animal classification systems based on different pluralistic criteria, which could be based on behavior, the way of life, development, mobility, etc. [4, 5]. It was a non-hierarchical system and admitted that the same animal classified into over one group. Von Lieven and Humar [6] performed an analysis on the zoological classification performed by Aristotle. They used 157 features used by the Greek and found 58 monophyletic groups, 29 of which were consistent with the groupings created by Aristotle. Therefore, Aristotle's classification was inaccurate but not arbitrary.

The Aristotelian system was accepted for almost 2000 years ago and was definitively replaced after the publication of the 10th edition of the Linnaeus *Systema Naturae* in 1758 by the Swedish naturalist Carl Linnaeus (1707–1778). The classification presented by Linnaeus was a landmark of zoological and botanical nomenclature, standardizing the classification systems in binomial and hierarchical [7]. The taxonomy presented by him presented the taxonomic levels of kingdom (divided in Animalia, Plantae, and Protista), phyla, classes, orders, families, genera, and species. For example, one of the fish studied by Aristotle in History of Animals, the kobios (or giant goby), according to the classification of Linnaeus, came to be called *Gobius cobitis*. In the Latinized name, *Gobius* corresponds to the genus, and cobius means the specific epithet.

Linnaeus was a fixer, but admitted in his classification the similarity between man and apes, and described hybridization in plants and animals, a fact which, according to him, was contrary to the stability of divine creatures [8]. Although ancestry is a classificatory criterion by other naturalists, such as Peter Simon Pallas (1741–1811) and Carl Edward von Eichwald (1795–1876) before Charles Darwin, it only gained prominence in 1859 with the publishing of the book *On the Origin of Species* [9]. But it only merged after the apogee of the synthetic theory of evolution, a scientific theory that united the knowledge of Gregor Mendel and Charles Darwin and the principles of population genetics. The synthetic theory had a long maturation that began early in the twentieth century and gained popular visibility with the release of the book *Evolution: The Modern Synthesis* in 1942 by Julian Huxley.

The advance of evolutionary ideas brought to light the proposals of today's renowned German entomologist Willi Hennig (1913–1976). He proposed in 1950 in the book *Phylogenetic Systematics* (translated from German into English in 1966) that biological classifications based on genealogical relationships between organisms are natural and unequivocal, so it is a biological reference system [10, 11]. The taxonomy created by Linnaeus today could be called phenetic taxonomy or numerical taxonomy, while the *Phylogenetic Systematics* could also be called cladistic taxonomy. The phenetic taxonomy is based on common observable features that do not necessarily reflect the phylogenetic relationship between groups. It gained many followers in the 1950s and 1960s, with the arrival of biostatistical methods and numerical computing, and had as main representatives Peter Sneath (1923–2011) and Robert R. Sokal (1926–2012) [12].

2.1. Phylogenetic tree

The phylogenetic tree or cladogram presents the following elements: node, branch, clade, root, branch lengths, and topology. Node is the branch point in the tree; the branch represents the descendant and ancestry; a clade is the groups that include the commune and descendant; the root is the common ancestor between the clades. Topology is the branching pattern of the tree, and branch length corresponds to changes in the branches. The elements of the tree are shown in **Figure 1**.

The characteristics used in the phylogenetic trees can be classified in two ways: those shared by ancestry, that is, the homologies, and others not evolutionarily related but with an analog

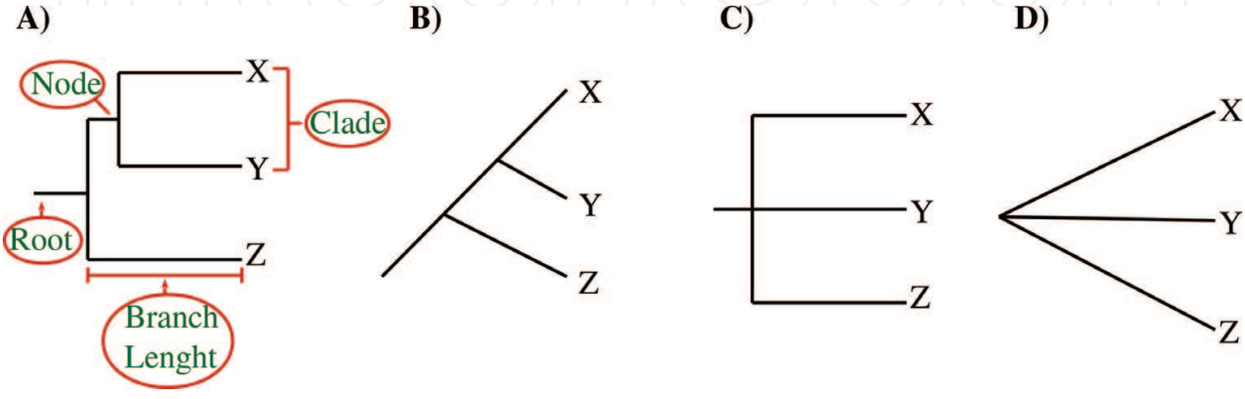

Figure 1. Phylogenetic trees showing different representations and topologies. (A) Presentation of the elements of a phylogenetic tree. (B) Another representation of the phylogenetic tree shown in A. Trees (C) and (D) are like each other and present topology that differs from trees A and B.

function called convergence or parallelism [13]. For example, homologies are present in species with recent common ancestors, as mammary glands and hairs present in mammals. The wings present in bats, birds, and insects are analogies. The comparison between analogy and homology is shown in **Figure 2**.

In a phylogenetic tree, the homologous characters are wanted. The convergences (also called homoplasy) may compromise the phylogenetic inference, although often present. After determining a homology, the next step to use it in a phylogeny is to determine its character state, to establish whether it derived from ancestral, that is, it is an apomorphy or plesiomorphy. One way to determine apomorphism and plesiomorphism is through character polarization by comparison with an external group. The out-group is a related taxon (i.e., a taxonomic unit) that one hopes to analyze [13]. For example, if the aim is to analyze the class Mammalia, it is interesting to have an out-group of another class of Amniota. If the target taxon is the primate, use as an out-group of another taxon from the superorder Euarchontoglires may be ideal. Once the out-group has defined, polarization can be made by comparing common traits to determine the apomorphies and plesiomorphies. The shared traits among the members of the target group are apomorphies, while those shared with the external group are plesiomorphies. A tree without an out-group is an unrooted tree, a tree in which the phylogenetic relationship between the branches is unclear (**Figure 3**). The tree of life, for example, is an unroot, since is not known the last universal common ancestor (LUCA).

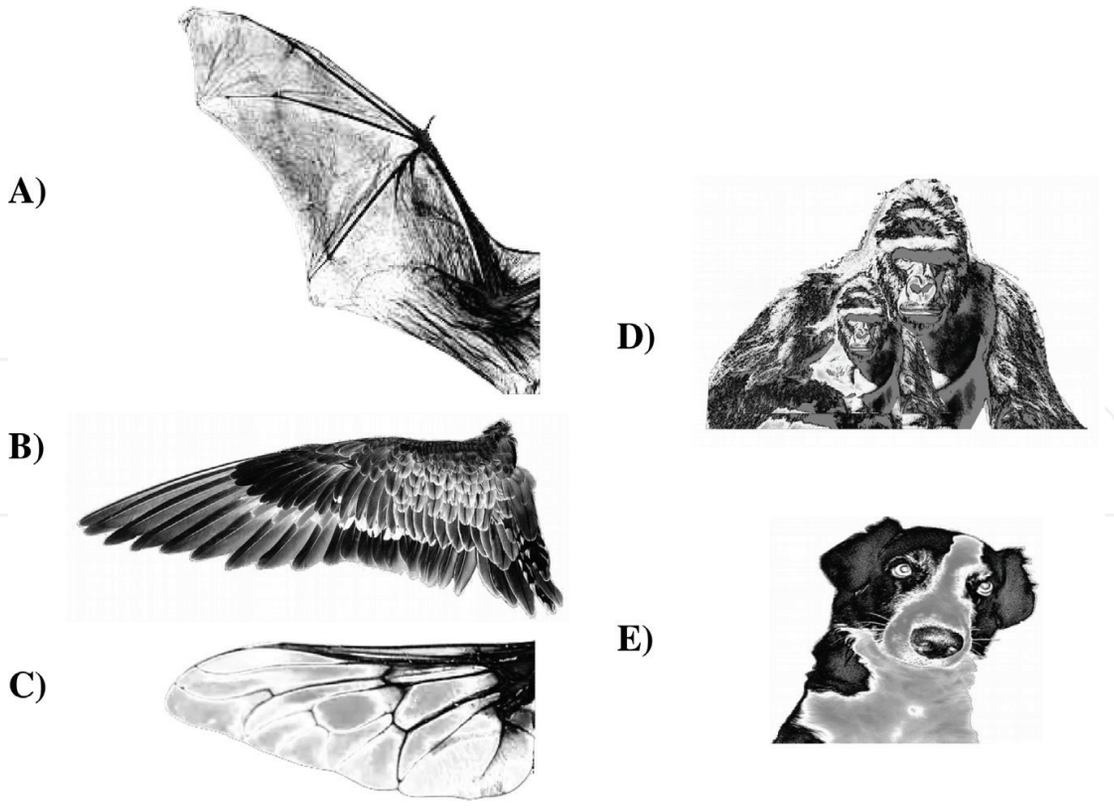

Figure 2. Difference between analogy and homology. In (A) is shown the wing of a bat, in (B) the wing of a bird, and (C) the wing of an insect. The three wings did not arise by common ancestry but by convergence or parallelism. The mammals (D) (gorilla) and (E) (dog) present homologies as mammary glands and hairs because they have a recent common ancestry.

http://dx.doi.org/10.5772/intechopen.79422

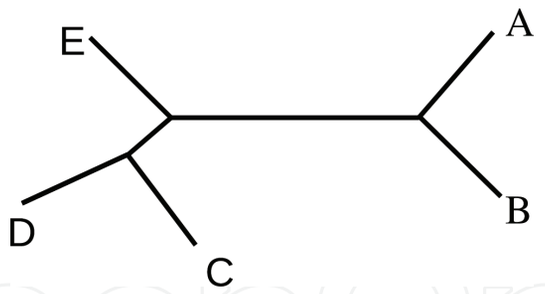

Figure 3. Representation of an unrooted tree. A, B, C, D, and E correspond to each of the taxa.

The apomorphies shared with the monophyletic group (a clade with all the ancestors and descendants) present a more recent common ancestor. Those apomorphies shared by two or more groups in a group are called synapomorphies [10, 13]. For example, having five digits is a synapomorphic trait of the modern tetrapods (the earliest tetrapods Acanthostega, Ichthyostega, and Tulerpeton presented more digits than the present species). Another type of apomorphism is the autapomorphies, specific characteristics of a group or taxon [13]. The plesiomorphy can be a symplesiomorphy that corresponds to when the ancestral characteristic is shared between certain clades [10, 13]. The different character states used in phylogeny are shown in **Figure 4**.

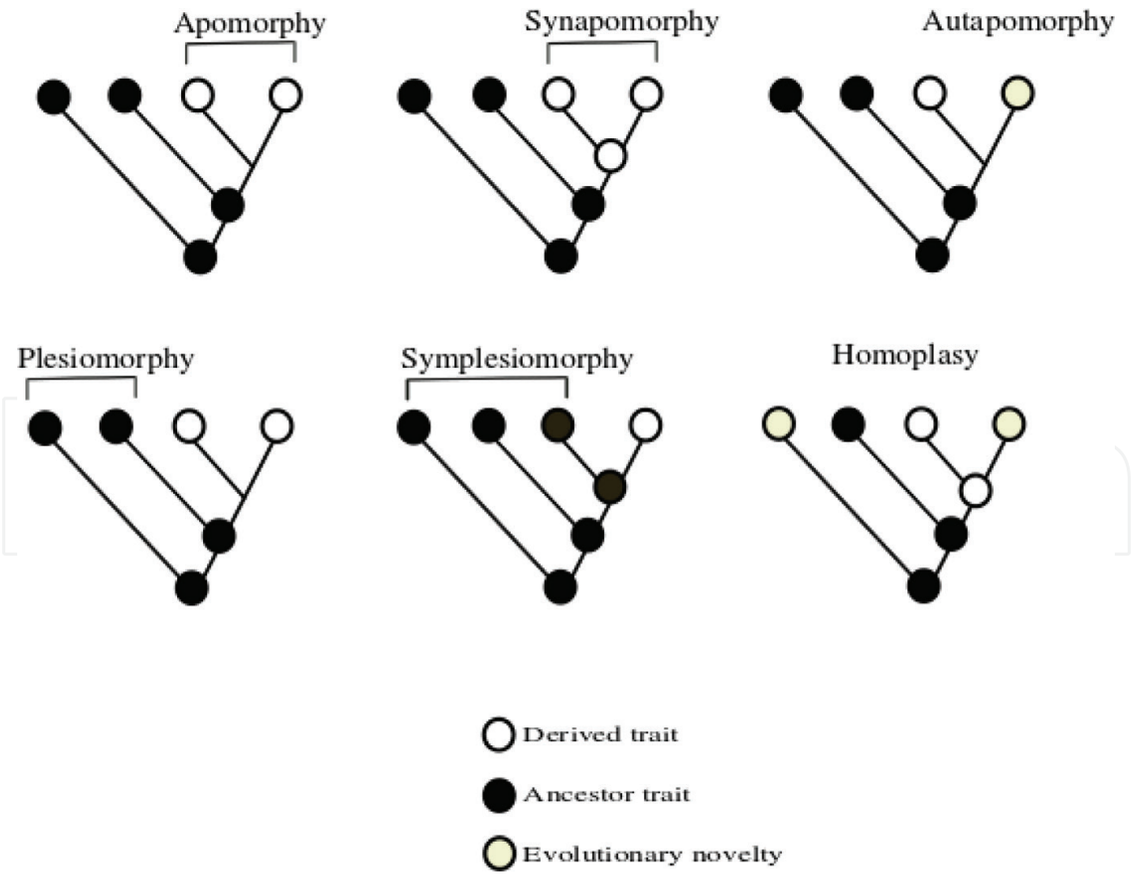

Figure 4. Types of characters used in phylogenetic trees. The different derived characters (apomorphy, synapomorphy, and autapomorphy) are shown in trees (A), (B), and (C). Meanwhile, the derived characters (plesiomorphy and symplesiomorphy) are shown in trees (D) and (E). Tree (F) presents the homoplasy, a convergence [13, 14].

2.2. Monophyletic, paraphyletic, and polyphyletic groups

Hennig assumed that phylogenetic relationships exist at different hierarchical levels, and the main role of phylogenetic systematics is to define the different degrees of kinship that can be in a phylogenetic tree [10]. One group that can be arranged in this tree is called a monophyletic group, which is defined by the author as "a group of species that contains all descendants of a single ancestral species" [10]. Within this context, species are reproductive communities isolated from others. The paraphyletic group exhibits some of its members in other groups, not monophyletic, as an example is Reptilia. It has Chondrichthyes (class formed by cartilaginous fishes) and Actinopterygii (superclass formed by ray-finned fishes). The second group presents a more recent common ancestry with Sarcopterygii (another superclass) but retained characteristics similar to those found in Chondrichthyes, such as gills. That is why Chondrichthyes and Actinopterygii often placed in the same group, but they are not. According to Hennig, the paraphyletic groups for being artificial should be abolished [10, 14]. The polyphyletic group also does not make up a natural group, and although they share common characteristics (by homoplasy), they do not have an immediate common ancestor. Both paraphyletic and polyphyletic groups produce uncertain phylogenies that are caused by a large number of homoplasies that can exceeds the amount of synapomorphies. Some fossil groups, due to the scarcity of data, may appear as paraphyletic or polyphyletic. Current groups, however, present a more robust classification because of the more sophisticated phylogenetic methods that use DNA or genome as the source of the data matrices. Hennig was right in wanting to abolish paraphyletic and polyphyletic groups, but it is not a trivial task. These groups are present even in more modern analyses based on data got by DNA sequencing. Some of these clusters result from a complex evolutionary history resulting from the exchange of genetic material between little related taxa. The horizontal transfer between bacteria can generate this evolutionary pattern of little clade containing possible polyphyletic and paraphyletic groups. But this matter will be treated more later. **Figure 5** shows the graphic representation of monophyletic, paraphyletic, and polyphyletic groups.

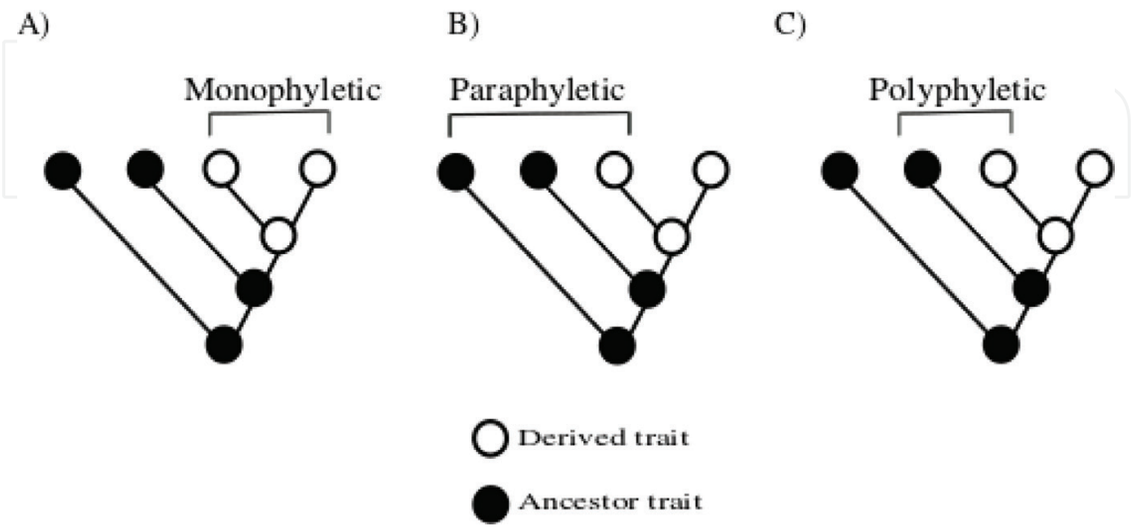

Figure 5. The phylogenetic tree shows the monophyletic, paraphyletic, and polyphyletic group. The monophyletic clade has only members with the same recent common ancestor (A), while the paraphyletic group has members in other groups (B) and the paraphyletic group (C).

2.3. Phylogenetic inference methods

2.3.1. Parsimony methods

The maximum parsimony method was one of the first methods use for construction of phylogenetic trees. This method obeys Occam's razor—a principle created by William of Ockham (1285–1347). According to this idea, the simplest hypothesis would be the best since nature tends to the economy. To analyze the best phylogenetic hypothesis, it is necessary to assemble a data matrix based on derived ancestral characters got by the comparison between the taxa with the out-group. By convention, the ancestral state that is present in the out-group is represented by 0 (zero), and the derived state is by 1 (one). Besides this binary matrix, the algorithm also performs analyses with matrices based on alignments of DNA and amino acid sequences. In the matrix only, some characters parsed. **Figure 6** shows an array constructed from a short nucleotide sequence of four hypothetical taxa (X, Y, Z, and W). In it there are 10 characters, only those of sites 2, 5, and 6 are informative (at least 2 taxa have the same nucleotide), and site 10 appears homoplasic.

The inference by maximum parsimony is inconsistent when there is a high rate of mutation in certain branches. And also presents a great problem is to consider all the sites with an equal chance of change; however, this does not correspond to biological reality [15]. Nucleotides and amino acids present different chances of change, and this should be considered when assembling a phylogeny.

During DNA replication, the DNA polymerase enzyme can incorporate nucleotide mismatch. If this failure is not repaired, the nucleotide sequence will show mutations. Transitions are mutations of the purine for a purine (e.g., A → G or G → A) or a pyrimidine for a pyrimidine (e.g., C → T or T → C), while the transversion is the shift from purine for a pyrimidine or vice versa (e.g., A → C or T, G → C or T, C → A or G, T → A or G). The transversions require more complex change, so they are less common than the transitions. Under the position of the mutation in the codon (first or second, mainly), it may cause an amino acid change and may give in the protein structure.

Then, the maximum parsimony is unrealistic. For example, in **Figure 5A**, the sites 2 and 5 of the taxon Z have two transitions in 2C → T, 5 T → C base positions in relation to X and

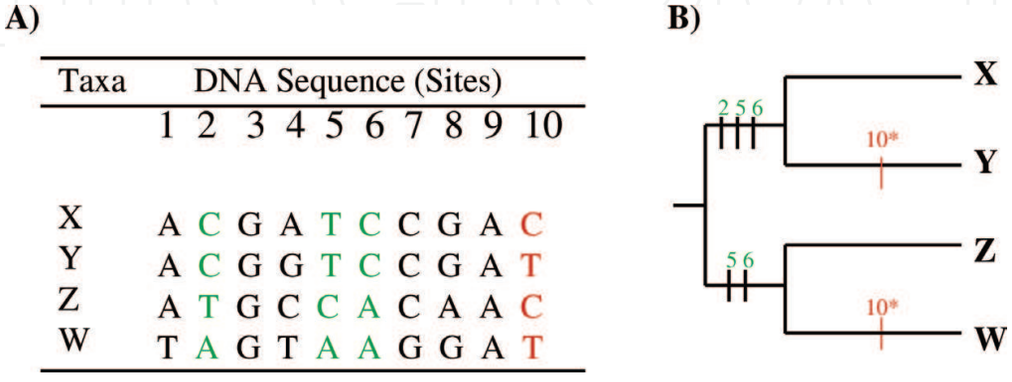

Figure 6. Simplified scheme of an array of nucleotide data characters (A) and phylogenetic tree assembled from data obtained from it (B). The sites highlighted in green in the matrix and tree correspond to the informational sites, and those in red is a supposed homoplasy site.

Y taxa. This may appear a clustering between the taxons X, Y, and Z. However, this clade is not formed due to the limitations of the method. The same problem can be observed in phylogenies built up from amino acid data. The amino acid exchanges with same physicochemical properties are not recognized by inference. Phylogenies based on maximum parsimony likewise present a major problem, the long-branch attraction (LBA), a phylogenetic artifact with high mutation rates forming erroneous groupings [16]. But, this does not mean that the maximum parsimony method will be abolished. It is even useful in analyses of conserved sequences, morphological and fossil data. In addition, the method has been optimized in different softwares [17–19].

2.3.2. Distance methods

The method infers the average number of changes per site between two rates. The total distance will be the division of the number of changes by the length of the sequence. In a sequence of 100 nucleotides, if the number of different bases between two sequences is 2, then the distance between them will be D = 0.02. The correction of this value is by the formula.

$$\mathbf{Jukes-Cantor: dxy = -(3/4)\ ln(1-4/3D).} \quad (1)$$

Dxy is the value of the correct distance between homologous sequences x and y, ln is the natural log (used to correct overlap of substitutions), and D is the observed distance between x and y. 3/4 and 4/3 reflect the nucleotides and have an equal chance of change. This formula is applied when the nucleotides have equal chances of change. Other more complex evolutionary models can also be assumed, such as the general time-reversible model (GTR), which assigns different probabilities for each type of change. Neighbor joining (NJ) method is the most used method, being fast. It uses the principle of parsimony or minimal evolution to find the best tree, based on the shortest length of the branches, with less evolutionary changes [20]. Although using the principle of parsimony, phylogenetic inference from NJ is more accurate, and together with unweighted pair group method with arithmetic mean (UPGMA), it is used in genomic analyses [21, 22].

2.3.3. Maximum likelihood (ML)

The maximum likelihood method was implemented by Anthony W. F. Edwards (1935-) and Luigi Luca Cavalli-Sforza (1922–2018) in the mid-1960s [23]. It is used to infer unknown parameters of a probability model in phylogeny analyses of different types of phylogenies and is able to estimate the length of the branches with a heuristic algorithm which is the phylogenetic tree that most likely to be generated from a given DNA sequence [24]. It can be defined by.

$$\mathbf{L = P\ (D \mid \theta).} \quad (2)$$

D corresponds to the probability of the dataset in a hypothesis θ. These hypotheses may be different parameters. The likelihood of each calculated nucleotide site and the total likelihood of the sequence are obtained from these data [24]. The probability of base substitution

occurring at time t is simplified by Pij (t). i and j correspond to the states of the sites. The probability of i changing to state j at time t is represented by Pij (t). The states correspond to bases A, C, G, or T or S = {1, 2, 3, 4}. Mutations in the bases are called random variables in a stochastic process. PMF or probability mass function of a random variable X is given by the formula.

$$\mathbf{pX(x) = P(X = x).} \tag{3}$$

This formula is applied when mutations have equal possibilities of occurring in the DNA sequence. Then, 0.25 is the probability for each of the four nucleotides, and can be represented as: pX (1) = 0.25, pX (2) = 0.25, pX (3) = 0.25, and pX (4) = 0.25.

Because the current and future states are independent, it presents a process with Markov property, and if the variables pass from another state after a certain time t, then the substitution process can be considered continuous-time Markov process and may be represented by.

$$\mathbf{Pij = P(X(t + s) = j \mid X(s) = i).} \tag{4}$$

The rows representing i (current state) and j (future state) are shown in columns. Each Pij item of the matrix is the probability of the process Markov at a time t. The ergodic Markov process (aperiodic and positive recurrent) and time reversibility properties are also assumed during the likelihood analyses. It is, therefore, the final inference of the product of different events and parameters, which makes it exhaustive and demands a great computational time, but it creates a more realistic phylogenetic scenario, as it also allows to test the phylogenetic hypothesis within complex substitution models (e.g., Hasegawa, Kishino, and Yano (HKY), and general time-reversible (GTR), established as the specific program as ModelTest or jModelTest [25, 26]. These models allow us to evaluate how nucleotide sequences evolve and which model best describes them. The best evolutionary hypothesis is tested by likelihood ratio tests (LRTs). It is an important step of phylogenetic inference because although ML is less sensitive to LBA, it is not a free method of this artifact when the assumed evolutionary model is wrong [16, 27–30].

2.3.4. Bayesian inference

As the maximum likelihood, the Bayesian inference is also a probabilistic method. The method was developed by Thomas Bayes theorem (1701–1761) and consists of describing the probability of events based on a priori knowledge about the event. The theorem is described by the equation

$$\mathbf{P(A \mid B)} = \frac{\mathbf{P(B \mid A)\, P(A)}}{\mathrm{P(B)}}. \tag{5}$$

A and B are the events, and P (B) ≠ 0.

P (A) and P (B) are the a priori probabilities of events A and B;

P (A | B) is the a posteriori probability of A conditioned to B;

and P (B | A) is the a posteriori probability of B conditioned to A.

The method was applied to phylogeny only from 1990, but the initial idea was generated in 1967 by Cavalli-Sforza and Edwards who used it in the estimation of gene frequencies in human populations [31]. The improvement of the initial idea allowed its optimization and application to nucleotide sequences and also added other mathematical processes to phylogenetic parameters. Birth-death process is used as a model of speciation and extinction of a priori distributions of phylogeny and length of branches [32]. The model of nucleotide substitution is estimated by the continuous-time Markov process [32], while the substitution models and model parameter of the branches are inferred by maximum likelihood [33]. The distribution of a posteriori is obtained through the Bayes theorem and performed through some known data a priori (D) and unknown parameters θ, applied to the equation below [34]:

$$f(\theta \mid D) = \frac{1}{z} f(\theta) f(\theta \mid D). \quad (6)$$

f (θ|D) is called likelihood and $z = \int f(\theta) f(\theta \mid D)$, normalizing constant.

The inference of the a posteriori distribution of phylogenies is performed with Markov chain Monte Carlo (MCMC) under the algorithm Metropolis-Hastings algorithm. The highest posterior probability is used to choose the best estimate [32, 33].

One of the problems of the method is to choose the optimal size to run the MCMC string to generate good later probabilities. If the value of the string is too low, the tendency is for the data to be large deviations and not realistic. In contrast, a long time can generate a very high computational time. One way to reverse this problem is to check the stationary phase (when the values a posteriori are stable) using different string sizes through programs such as R and Tracer; plotting the data will allow evaluating the consistency of the data [34]. In addition, some authors [35–37] point out that when a large database is analyzed by Bayesian inference the tree tends to present arbitrary polytomies (unresolved branches with more than two clades appearing at the same time) with auto values of posterior probability, but this problem is easily solved by modification in the Metropolis-Hastings algorithm so that a less-resolved topology is assumed [38]. With the use of the method, it is possible to analyze DNA data, amino acids, as well as morphological data [34].

2.4. Data resampling approaches

Only the construction of a phylogeny does not support its reliability. The confidence of a given phylogenetic hypothesis is assured by support values that can be obtained by different statistical approaches. Some of these approaches are bootstrap, jackknife, Bremer support, and posterior probability. Below, I will argue each of them.

2.4.1. Bootstrap and jackknife

The most popular estimate to test the robustness of a phylogenetic hypothesis is a nonparametric method applied to a phylogenetic analysis by Joseph Felsenstein (1942-) in 1985 [39]. The method comprises a resampling with replenishment of the database. From it, pseudo-alignments (with the same length) are generated from where pseudo-trees will be created.

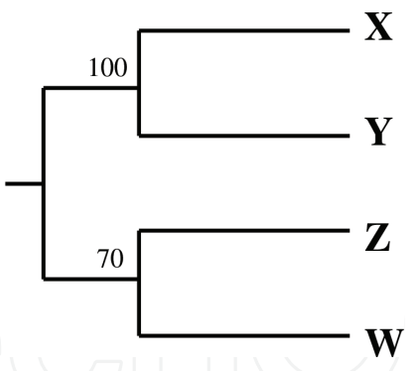

Figure 7. Phylogeny exemplifies how the bootstrap values are exposed on the nodes of the consensus tree. The groups X and Y were together during 100% of the pseudo-trees, whereas Z and W formed a clade in only 70% of them.

The number of replicates will imply the number of pseudo-alignments and generated pseudo-trees [39, 40]. For example, if the number of replicas chosen is 100, then at the end of the analysis we will have 100 pseudo-trees, which can be represented in a single phylogeny and a consensus tree. **Figure 7** shows an example of phylogeny using bootstrap.

The values on the nodes are the bootstrap which is the number of times a given clade has been repeated in pseudo-trees. Controversial groups usually present inconsistencies with low bootstrap values (i.e., below 70%), while consistent groupings present high bootstrap values (close to 100%). The method is used in some phylogenetic inferences as maximum parsimony, neighbor joining, maximum evolution, UPGMA, and maximum likelihood.

Hedges [41] suggests that 2000 replicates increase the accuracy of phylogeny because the p-value bootstrap reaches about ±1% at a 95% confidence limit. Therefore, more than 2000 replicates have little effect and increase the computational time of the analyses.

Jackknife is similar to the bootstrap, but it is an unsampled resampling with subsets of data smaller than the original. In this way, it is possible to know if the exclusion of certain characteristics will have an effect on the topology. It can also be used in analyses of maximum parsimony. For both bootstrap and jackknife, the increase of replicas reduces the standard deviation. For Müller [42] a number of replicates greater than 3458 are unnecessary since it no longer reduces the standard deviation in both confidence estimates.

2.4.2. Bremer support

Like previous methods, this method also causes disorders that may reveal data fragility and homoplasies. The support of Bremer also called decay index, however, allows verifying the number of extra steps needed to break a branch in relation to the fewer parsimony trees [43]. It is an important test to test the stability of phylogenies based on parsimony but was originally used in distance analysis [44]. It corresponds to the ratio of the consistency index to the number of steps in a given tree. For example, if the most parsimonious phylogeny has 88 steps, the consistency index will be equal to 1. If another phylogeny based on the same database presents 100 steps, then its consistency index will be 88/100, that is, 0.88 [45]. The method is a good alternative to test the monophyly of taxa whose data were generated by morphological data.

2.4.3. Posterior probability

The support value used in Bayesian inference is the posterior probability. It is a way of check the probability of a particular phylogeny, where the probability of the tree is given by P(T), given the data D, or P (T|D). A tree is characterized by the topology τ and associated with the length of β branches. Thus the value of posterior probability is given by [46].

$$\mathbf{P}(\boldsymbol{\tau} \mid \mathbf{D}). \tag{7}$$

The relationship between posterior probability value and bootstrap was not established; however, what is possible to observe is that the same phylogenetic hypothesis presents higher values of posterior probability value than bootstrap [46].

2.5. Phylogenetics software

With the development of different phylogenetic methods and technological advancement, various programs or packages were built. These programs allow the analysis of thousands of data that would be impossible to work manually. Generally, each of the programs for phylogenetic analysis uses different formats of input files. The formats can be of different types, fasta, meg, nexus, phylip, clustal, and MFS format. These formats are generated during the alignment of sequences that can be performed in the programs Clustal X, Clustal W [47], Bioedit [48], and Aliview [49]. Once the alignment is properly formatted, you can then run the analyses in the desired program. In this session, I will present some programs of phylogenetic analysis and general characteristics of them.

2.5.1. FastTree

The software is an open source and can be installed on different platforms (Mac, Linux/Unix, and Windows). It has the purpose of doing ML analyses of thousands of DNA, RNA, and protein data much faster than other programs (about 100–1000 times faster). For DNA analysis you can use the Jukes-Cantor and GTR replacement models, which is a limitation. For protein data, it uses the Jones-Taylor-Thornton 1992 (JTT) [50], Whelan and Goldman 2001 (WAG) [51], and Le and Gascuel 2008 (LG) [52] models. One of the great advantages of the program is to use a category of each site (or CAT model) approach, and it reduces the computational time during the analyses, mainly of amino acids [53–55]. The program uses a specific type of support value, called local-bootstrap support values that can vary throughout the search, but the traditional bootstrap can be obtained by using the SEQBOOT program (belonging to the phylogeny inference package) that resamples the data. The program written in Perl CompareToBootstrap.pl. can be used to compare the tree generated by FastTree and this, with resampling of the data. The program uses the multiple sequence alignment (MSA), fasta, and interleaved phylip format formats.

2.5.2. Molecular evolutionary genetics analysis (MEGA)

It presents a very friendly graphical interface, besides being free [56, 57]. It also works on Mac, Linux/Unix, and Windows. It has some advantages, such as the ability to perform sequence

alignment in the program itself through MUSCLE or Clustal. The program also has the option of looking for the appropriate replacement model for the data (however it is little used for this) and the possibility of constructing the distance matrix. Phylogenies can be based on ML, NJ, minimal evolution, and UPGMA. Bootstrap values can be added to trees or the tree consensus easily by choosing the number of replicas. It allows the analysis of DNA, RNA, and protein and also the distance of the matrix. The tree created can be viewed and edited in the program itself, which increases its practicality than other phylogenetic analysis programs. ML analyses have Jukes-Cartor models, Kimura 2-parameters, Tamura 3-parameters, Hasegawa-Kishino-Yano, and GTR for nucleotide data and 13 models for amino acid sequences (Poisson, equal entry, Dayhoff, JTT, JTT + F, WAG, WAG + F, LG, F + LG, mtREV, mtREV + F, cpREV, cpREV). The version for the Microsoft Windows operating system can execute strings of different extensions (.an, .nexus, .phylip, .gcg, fasta, .pir, .nbrf, .msf, .ig, and .xml) which must be converted into extension. meg, usually found by the program.

2.5.3. MrBayes

It is a most commonly used Bayesian analysis programs [58, 59]. It is also free and serves all major operating systems, but it needs to be compiled in the Unix/Linux version. It allows the analysis of DNA, RNA, and protein restriction sites morphological data and also from a mixed file containing the mixture of these data. The input file is the nexus format. This file, in addition to the nucleotide sequence, protein, etc., should also have additional information such as the specific evolution model and other useful parameters to perform the analysis. The choice of each of the parameters of the input file should be placed with great care, preferably following the steps in the program manual, as errors may interfere with the final result of the analysis.

2.5.4. Phylogenetic analysis using PAUP (PAUP)*

PAUP is one of the most popular software for maximum parsimony, but can also be used in the phylogenetic reconstruction of neighbor-joining. The original version was paid for. Some changes are happening in PAUP version 4.0 of the software; there are options to run on Mac OS X, Windows, and Linux. An open-source command-line version (need the Fortran runtime) is under construction, as is a graphical user interface (GUI) for Windows [60, 61]. According to the developers of the program, the trial versions are still free, but those with a graphical interface will expire, except for the command-line version. The default input file is nexus (.nex). The default input file is nexus (.nex), and all information about the sequence must be in it, such as alignment, substitution model, if the given ones are partitioned and how, if the used sequence is coding, etc. Due to a large amount of data contained in this file, it should be built with care and attention (especially to the symbols accepted by the program), otherwise, a series of bugs will appear. The program is easy to execute, mainly in the version with a graphics interface. It works with DNA, RNA, proteins, and discrete character data (1/0).

2.5.5. Phylogeny inference package (PHYLIP)

It is a package consisting of about 30 programs in C source code. It is free and can be used on Mac, Linux/Unix, and Windows. It has programs that run analysis of parsimony, neighbor Joining, UPGMA, and likelihood. It can create phylogenies based on a distance matrix

(fitch program). It is quite versatile working with data from DNA, RNA, amino acid, gene frequency, and discrete character data (1/0). It can use bootstrap or jackknife (SEQBOOT) to determine the support of the branches and also presents a specific program (consense) for building a consensus tree [62–64]. The program does not have a graphics interface and presents few substitution models for both DNA and proteins. However, it performed a large number of phylogenetic methods.

2.5.6. PHYML

The great advantage of the software is to present a likelihood analysis for nucleotides and proteins. Unlike most programs, they only analyze nucleotides. The input file is in phylip (.phy) format. The program is free code and can be installed on all platforms; however, the installation presents particularities that must be respected. It has a list of choices which facilitates its execution and is one of the software with bigger options for models of substitution (JC69, K80, F81, F84, HKY85, etc.). The number of bootstrap replicas is not automatically generated, it must be chosen, with 100 being the default amount. With each replicate, a phylogeny is generated [65–67]. The program currently has smart model selection in PHYML (SMS) [67], another program that assists in the search for the best replacement model for nucleotides and proteins. Both PHYML and SMS have versions for online execution. Although quite versatile, developers recommend that the database has between 100 and 200 sequences and a maximum of 2000 characters. The software becomes slow and consumes a lot of memory with larger banks [67].

2.5.7. Randomized accelerated maximum likelihood (RAXML)

The program is an open source for ML analysis, an alternative to PHYML for long databases. It is derived from dnaml, one of the programs available in PHYLIP [63]. The input files are in phylip (.phy) or fasta (.fas) formats. It can perform binary, nucleotide, and protein data. It is one of the programs that have more options of substitution models for phylogenetic inferences based on data of proteins. It is available for all platforms, but the form of installation depends on the type of platform and also the configuration of the processor. The AVX version can run on more modern processors (e.g., the Intel i7 series or AMD Bulldozer systems) and runs faster than the SSE3 version. In addition, Mac and Linux will have different compilation forms that must meet the instructions in the manual for the correct installation. The likelihood value is more similar to the PHYML values found because they use similar methods, but it is not comparable to those obtained by other ML analysis programs. CAT model of rate heterogeneity can be used in long databases (over 50 taxa) to accelerate the phylogenetic inferences of the initial trees. Later the search of the trees is refined with the use of RΓ (refinement under Γ) search algorithm [68–70].

2.6. Visualization tree tools

Many of the software for phylogenetic construction at the end of the analysis generate a tree in non-graphical, difficult-to-interpret formats. Except for the MEGA program that automatically opens a tree after its construction, most phylogenies will need to be shown using other features, such as Archeopteryx, TreeView, and iTol.

2.6.1. Archaeopteryx

It is a free code software that allows viewing and editing of phylogeny. The program is written in Java and this allows it to be installed on all platforms. It can read phylogenies in different formats (phyloXML, newick, nexus, nhx, etc.). The options are quite versatile, being possible to edit different informations (color, root, form of phylogeny, etc.) [71–77]. It is important that the user fully exploits the program and chooses the best options to represent their phylogeny.

2.6.2. Dendroscope

The software is written in Java available for all platforms. It is an easy-to-use program, but it does not have editing options as sophisticated as Archeopteryx. It accepts the formats nexml, .dendro, .tre, and nexus and also has different options for editing phylogenies [71, 72].

2.6.3. iTol

It can be used online and has several editing possibilities. Phylogenies can be seen in a circular, normal, or non-root fashion. Branches can be colored differently for better identification of taxonomic groups. The program also allows the addition of captions, connections, heat maps, box plots, protein domains, and annotation data. The input files can be of the newick, nexus, phyloXML, jplance, QIIMe2, and NHX types [73, 74].

2.6.4. TreeView

The open-source software can interpret a large number of phylogeny formats [75, 76]. It is quite simple and easy to use, but it does not have editing options as sophisticated as Archeopteryx.

2.7. Gene tree versus species tree

Not all phylogenetic reconstructions can reflect the evolutionary history of a group; sometimes the evolutionary history of the gene is shown in the phylogenetic hypothesis. Pamilo and Nei [77] emphasized that it is important to distinguish between a gene tree and species tree. Gene tree shows the history of paralogous genes in different species. While the specie tree reflects the processes of speciation within a lineage, through the use of orthologous genes. Orthologs have similar functions among the organisms that possess them.

The genes undergo multiple duplication processes and may present multiple copies with distinct functions in the genome of the same species, as an example, the glycosyltransferase 6 gene family, which possesses the ABO gene. Some parallel copies may still lose function in some groups and become pseudogenes, as an example GGTA1 in Catarrhini (human, apes, and old world monkeys) [78]. **Figure 8** shows the difference between two types of homologous genes.

The misuse of paralogous genes while attempting to construct the phylogeny of a taxon is a recurring problem and needs the care to ensure the use of orthologous genes. The most effective procedure is to verify the similarity of the target sequence through the basic local

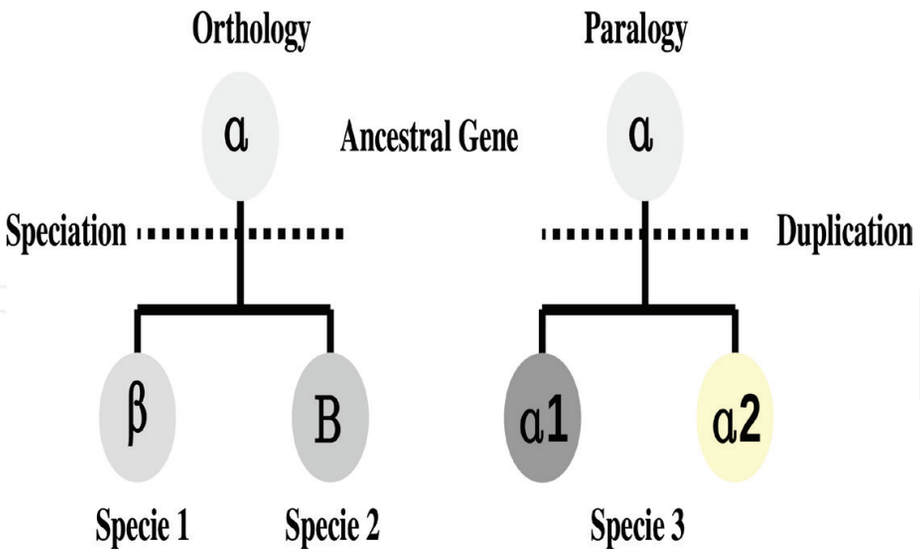

Figure 8. Difference between orthologous genes and paralogs, when the ancestral gene is being represented by α. The first phylogenetic tree shows the origin of the orthologues β and B, coming from the speciation between species 1 and 2. The second phylogeny represents the process of formation by duplication of the genes α1 and α2.

alignment search tool (BLAST) and to analyze the results with the best scores (with the lowest values of e-value). The orthologous gene will tend to exhibit more similarities than the paralogs (one of the parallel graphic copies maintains the original function, while others may have multiple mutations). Another strategy is to verify in the literature whether the target gene has copies within the genome of the species analyzed.

The third type of homologous gene may compromise the validity of a phylogenetic hypothesis, the xenologous genes, which were obtained by horizontal event gene transfer (HGT) between two species. Although they may act as phylogenetic artifacts, these genes are of extremely evolutionary importance (the possibility of contamination should be considered first). They can be easily identified through the BLAST tool, which shows rather unusual results, indicating similarities between the target sequence and others of bacterial or viral origin. These genes are quite common among prokaryotes, and the best known are those related to antibiotic resistance. HGT also played a key role in the evolution of eukaryotes, mainly in the origin of this domain from several events of serial endosymbiosis, fundamental in the acquisition of nucleus and organelles [79]. Another striking example was the syncytin gene, originated from reiterated endogenous retrovirus (ERV) sequences that were fundamental in the formation of placental structures in eutherian mammals [80, 81].

2.8. Tree of life

From Charles Darwin to today, it is difficult to determine what would be the real tree of life, complete and unequivocal. What is concrete today is that life is composed of the domains (or superkingdom) Bacteria, Eukarya, and Archaea [82]. The proposal of the third group was observed by Carl Woese and George Fox based on the 16S ribosomal gene [83]. This taxonomic proposal is shown in **Figure 9**. The relationships between these three groups are controversial. Bacteria has the wall cellular with peptidoglycans, different from members of Archaea and Eukarya.

http://dx.doi.org/10.5772/intechopen.79422

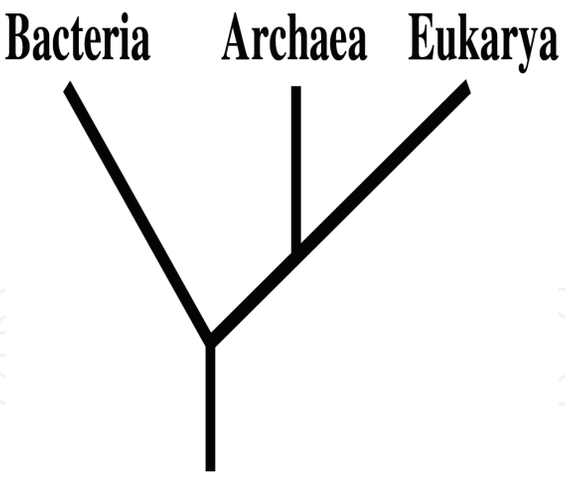

Figure 9. The three domains (or superkingdom) of life: Bacteria, Archaea and Eukarya.

These two groups, in turn, have similar replication, transcription, and translation mechanisms. However, only Eukarya has a cytoskeleton containing tubulin and actin [83, 84]. Thus, although this classification is robust, it still needs more information about a possible LUCA.

3. Conclusions

Phylogenetic systematics emphasized the investigation of phylogenetic relationships among living beings. However, this view is pre-Darwinian and had supporters over 2 millennia ago, in antique Greece. From then on, the relationships of ancestry between living beings are described as cladograms constructed from homologies and homoplasies. The computational advancement allowed new methods and phylogenetic approaches based on advanced mathematical assumptions. As a result, current free computational tools are rising, aiming at the analysis of long databases faster and higher, and with fewer phylogenetic artifacts (such as homoplasies). Each method and software, however, should be appropriate to the database (sample number and a number of characters) and computational power. But deserves attention in terms of the support value that differs from that applied in classical methods. Visualization and editing of phylogenies are further possible through various tools easy to install and run. Despite all the advantages, it is essential that the researcher knows what evolutionary history wants to produce in his phylogeny, whether it is the genes or species history. For this, it is indispensable to identify orthologous, paralogues, and xenologous genes. It is frequent for concatenated analysis of these genes to generate a puzzling and unclear phylogeny. It is important to consider that each gene can show ancestral histories of particular lineages, such as 16S gene which provided the tree of life, comprising the Bacteria, Eukarya, and Archaea domains.

Conflict of interest

The author has declared that no competing interests exist.

Author details

Eliane Barbosa Evanovich dos Santos

Address all correspondence to: lianevanovich@gmail.com

Laboratório de Genética Humana e Médica, Instituto de Ciências Biológicas—Universidade Federal do Pará, Pará, Brasil

References

[1] Pallas PS. Elenchus zoophytorum sistens generum adumbrationes generaliores et specierum cognitarum succintas descriptiones, cum selectis auctorum synonymis. The Hague (The Netherlands): Apud Petrum van Cleef, 1766

[2] Trevisanato SI. Reconstructing anaximander's biological model unveils a theory of evolution akin to darwin's, though centuries before the birth of science. Acta Medico-Historica Adriatica. 2016;**14**:63-72

[3] Couprie DL, Kočandrle R. Anaximander's 'Boundless Nature'. Peitho examina. antiqua. 2013;**1**:63-91

[4] Dunn PM. Aristotle (384-322 bc): philosopher and scientist of ancient Greece. Archives of Disease in Childhood. Fetal and Neonatal Edition. 2006;**91**:F75-F77

[5] Ragan MA. Trees and networks before and after Darwin. Biology Direct. 2009;**4**(43)

[6] von Lieven AF, Humar M. A Cladistic Analysis of Aristotle's Animal Groups in the Historia animalium. History and Philosophy of the Life Sciences. 2008;**30**:227-262

[7] Müller-Wille S, Charmantier I. Natural history and information overload: The case of Linnaeus. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences. 2012;**43**:4-15

[8] Paterlini M. There shall be order. The legacy of Linnaeus in the age of molecular biology. EMBO Reports. 2007;**8**:814-816

[9] Darwin C. On the origin of species. London: Murray; 1859

[10] Phylogenetic Systematics HW. Annual Review of Entomology. 1965;**10**:97-116

[11] Andersen NM. The impact of W. Hennig's "phylogenetic systematics" on contemporary entomology. European Journal of Entomology. 2001;**98**:133-150

[12] Jensen RJ. Phenetics: revolution, reform or natural consequence? Taxon. 2009;**58**:50-60

[13] Brooks, DR, Caira JN, Platt TR, Pritchard MH. Principles and methods of phylogenetic systematics : a cladistics workbook. 1984. Harvard Botany Libraries. University of Kansas, Lawrence, USA

[14] Dupuis C. Willi Hennig's impact on taxonomic thought. Annual Review of Ecology and Systematics. 1984;**15**:1-25

[15] Mount DW. Maximum Parsimony Method for Phylogenetic Prediction. CSH Protocols. 2008;**3**

[16] Bergsten J. A review of long-branch attraction. Cladistics. 2005;**21**:163-193

[17] Gregor I, Steinbrück L, McHardy AC. PTree: pattern-based, stochastic search for maximum parsimony phylogenies. Peer-reviewed Journal. 2013;**25**:e89

[18] Hoang DT, Vinh LS, Flouri T, Stamatakis A, von Haeseler A, Minh BQ. MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. BMC Evolutionary Biology. 2018 Feb 2;**18**(1):11

[19] Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution. 2018;**35**:518-522

[20] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution. 1987;**4**:406-425

[21] Nascimento M, Sousa A, Ramirez M, Francisco AP, Carriço JA, Vaz C. PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. Bioinformatics. 2017;**33**:128-129

[22] Hedge J, Wilson DJ. Bacterial Phylogenetic Reconstruction from Whole Genomes Is Robust to Recombination but Demographic Inference Is Not. MBio Journal. 2014;**5**:e02158-e02114

[23] Cavalli-Sforza LL, Edwards AW. Phylogenetic analysis. Models and estimation procedures. American Journal of Human Genetics. 1967;**19**:233-257

[24] Cho A. Constructing Phylogenetic Trees Using Maximum Likelihood. Scripps Senior Theses. 2012;**46**

[25] Posada D, Crandall KAMODELTEST. testing the model of DNA substitution. Bioinformatics. 1998;**14**:817-818

[26] Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models. new heuristics and parallel computing. Nature Methods. 2012;**772**

[27] Gaut BS, Lewis PO. Success of maximum-likelihood phylogeny inference in the 4-taxon case. Molecular Biology and Evolution. 1995;**12**:152-162

[28] Huelsenbeck JP. Performance of phylogenetic methods insimulation. Systematic Biology. 1995;**44**:17-48

[29] Chang JT. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across char-acters. Mathematical Biosciences. 1996;**134**:189-215

[30] Lockhart PJ, Larkum AWD, Steel MA, Waddell PJ, Penny D. Evolution of chlorophyll and bacteriochlorophyll: theproblem of invariant sites in sequence analysis. PNAS. 1996;**93**:1930-1934

[31] Cavalli-Sforza LL, Edwards AWF. Phylogenetic analysis: models and estimation procedures. Evolution. 1967;**21**:550-570

[32] Yang Z, Rannala B. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Molecular Biology and Evolution. 1997;**14**:717-724

[33] Rannala B, Yang Z. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. Journal of Molecular Evolution. 1996;**43**:304-311

[34] Nascimento F, Reis M. Yang Z. A biologist's guide to Bayesian phylogenetic analysis. Nature Ecology & Evolution. 2017;**1**:1446-1454

[35] Suzuki Y, Glazko GV, Nei M. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. PNAS. 2002;**99**:16138-16143

[36] Alfaro ME, Zoller S. Lutzoni f. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Molecular Biology and Evolution. 2003;**20**:255-266

[37] Douady CJ. Delsuc f, Boucher Y, Doolittle WF, Douzery EJP. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Molecular Biology and Evolution. 2003;**20**:248-254

[38] Lewis PO, Holder MT, Polytomies HKE. Bayesian phylogenetic inference. Systematic Biology. 2005;**54**:241-253

[39] Felsenstein J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution. 1985;**39**:783-791

[40] Efron B, Halloran E, Holmes S. Bootstrap confidence levels for phylogenetic trees. PNAS. 1996;**93**:7085-7090

[41] Hedges SB. The number of replications needed for accurate estimation of the bootstrap P value in phylogenetic studies. Molecular Biology and Evolution. 1992;**9**:366-369

[42] Müller KF. The efficiency of different search strategies in estimating parsimony jackknife, bootstrap. and Bremer support. BMC Evolutionary Biology. 2005;**5**:58

[43] Bremer K. Branch support and tree stability. Cladistics. 1994;**10**:295-304

[44] Farris JS, Kluge AG, Mickevich MF. Immunological Distance and the Phylogenetic Relationships of the Rana boylii Species Group. Systematic Zoology. 1982;**31**:479-491

[45] Bremer K. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. Evolution. 1988;**42**:795-803

[46] Cummings MP, Handley SA, Myers DS, Reed DL, Rokas A, Winka K. Comparing bootstrap and posterior probability values in the four-taxon case. Systematic Biology. 2003;**52**:477-487

[47] Larkin MA1, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;**23**:2947-2948

[48] Hall TA. Bioedit: a user-friendly biological sequence aligniment editor and analysis program for windows 95/98/NT. Nucleic acids symposium series. 1999:95-98

[49] Larsson AAV. a fast and lightweight alignment viewer and editor for large data sets. Bioinformatics. 2014;**30**:3276-3278

[50] Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Computer Applications in the Biosciences. 1992;**8**:275-282

[51] Whelan S. Goldman N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Molecular Biology and Evolution. 2001;**18**:691-699

[52] Le SQ, Gascuel O. An improved general amino acid replacement matrix. Molecular Biology and Evolution. 2008;**25**:1307-1320

[53] Le SQ, Lartillot N, Gascuel O. Phylogenetic mixture models for proteins. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences. 2008;**363**: 3965-3976

[54] Price MN, Dehal PS, Arkin APFT. Computing Large Minimum-Evolution Trees with Profiles instead of a Distance Matrix. Molecular Biology and Evolution. 2009;**26**:1641-1650. DOI: 10.1093/molbev/msp077

[55] FastTree [Internet]. 2018. Available from: http://www.microbesonline.org/fasttree/ [Accessed: 2018-06-06]

[56] Kumar S, Stecher G, Tamura K. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. Molecular Biology and Evolution. 2016;**33**:1870-1874

[57] MEGA [Internet]. 2018. Available from: https://www.megasoftware.net [Accessed: 2018-06-06]

[58] Ronquist F, Huelsenbeck JPMRBAYES. 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003;**19**:1572-1574

[59] MrBayes [Internet]. 2018. Available from: http://mrbayes.sourceforge.net/manual.php [Accessed: 2018-06-06]

[60] Maddison DR, Swofford DL, Maddison WPNEXUS. an extensible file format for systematic information. Systematic biology. 2003;**46**:590-621

[61] PAUP* [Internet]. 2018. Available from: https://paup.phylosolutions.com [Accessed: 2018-06-06]

[62] Felsenstein J. PHYLIP (Phylogeny Inference Package) version 3.7a. Distributed by the Author. Department of Genome Sciences, University of Washington, Seattle. 2009

[63] PHYLIP [Internet]. 2018. Available from: http://evolution.genetics.washington.edu/phylip/getme-new1.html [Accessed: Jun 6, 2018]

[64] Felsenstein J. PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics. 1989;**5**: 164-166

[65] Guindon S, Gascuel O. PhyML: A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology. 2003;**52**:696-704

[66] PHYML [Internet]. 2018. Available from: http://www.atgc-montpellier.fr/phyml/ [Accessed: 2018-06-06]

[67] Lefort V, Longueville JE, Gascuel OSMS. Smart Model Selection in PhyML. Molecular Biology and Evolution. 2017;**34**:2422-2424

[68] Stamatakis A. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics. 2014;**30**:1312-1313

[69] RAXML [Internet]. 2018. Available from: https://sco.h-its.org/exelixis/web/software/raxml/index.html [Accessed: 2018-06-06]

[70] RAXML [Internet]. 2018. Available from: https://sco.h-its.org/exelixis/web/software/raxml/index.html [Accessed: Jun 6, 2018]

[71] Huson DH, Scornavacca C. Dendroscope 3: An interactive tool for rooted phylogenetic-trees and networks. Systematic Biology. 2012;**61**:1061-1067

[72] Dendroscope 3 [Internet]. 2018. Available from: http://dendroscope.org [Accessed: Jun 6, 2018]

[73] Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. Science 2006;**311**:1283-1287

[74] iTol [Internet]. 2018. Available from: https://sco.h-its.org/exelixis/web/software/raxml/index. html [Accessed: Jun 6, 2018]

[75] Page RDMTREEVIEW. An application to display phylogenetic trees on personal computers. Computer Applications in the Biosciences. 1996;**12**:357-358

[76] Treeview [Internet]. 2018. Available from: http://taxonomy.zoology.gla.ac.uk/rod/treeview.html [Accessed: Jun 6, 2018]

[77] Pamilo P, Nei M. Relationships between Gene Trees and Species Trees. Molecular Biology and Evolution. 1988;**5**:568-583

[78] Galili U. Significance of the Evolutionary α1,3-galactosyltransferase (GGTA1) gene inactivation in preventing extinction of apes and old world monkeys. Journal of Molecular Evolution. 2015;**80**:1-9

[79] Keeling PJ, Palmer JD. Horizontal gene transfer in eukaryotic evolution. Nature Reviews Genetics. 2008;**9**:605-618

[80] Cornelis G, Heidmann O, Degrelle SA, Vernochet C, Lavialle C, Letzelter C, et al. Captured retroviral envelope syncytin gene associated with the unique placental structure of higher ruminants. PNAS. 2013;**110**:E828-E837

[81] Lavialle C, Cornelis G, Dupressoir A, Esnault C, Heidmann O, Vernochet C. Heidmann T Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. Philosophical Transactions. 2013;**368**:20120507

[82] Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. PNAS. 1990;**87**:4576-4579

[83] Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. PNAS. 1977;**74**:5088-5090

[84] Doolittle WF. Phylogenetic classification and the universal tree. Science. 1999;**284**: 2124-2129

IntechOpen

IntechOpen