

## Chapter 5

# Computational molecular phylogeny: concepts and applications

Krishna Kumar Ojha<sup>1</sup>, Swapnil Mishra<sup>2</sup> and Vijay Kumar Singh<sup>1</sup>

<sup>1</sup>Department of Bioinformatics, Central University of South Bihar, India, <sup>2</sup>Center for Bioinformatics, University of Allahabad, India

### 5.1 Introduction

The biotic diversity of the earth is tremendous. Scientists have estimated that there are around 8.7 million species of plants and animals in existence (Margulis & Chapman, 2009). However, only around 1.2 million species have been identified and described so far, in which the majority are insects (Mora, Tittensor, & Adl, 2011). This means that millions of other organisms remain a complete mystery. Life originated in the ocean around 3.6 billion years ago in the form of single cells and as time passes it propagated in various directions, resulting in a tremendous diversity of habit and habitat of organisms visible today (Oparin, 1957). Notwithstanding the differences in appearance and morphological characters among the organisms, all are inherently linked evolutionarily in the form of a branched tree, in which the leaves and twigs represent recent species (offspring's) while the internal branches represent the old ancestors (Sellés-Martínez, 1996). This enigma of the ancestral relationship was solved by Carl woes in 1977 in his remarkable work of deciphering the tree of life, based on the molecular evolution of r-RNA of various domains (Woese, 2000). All organisms are morphologically or genetically linked together. Back tracing of the evolutionary history of an organism shows that organisms are connected through shared ancestors to lineages of other organisms (Field, Olsen, & Lane, 1988). Systematics is the branch of biology that constructs trees and reveals the sequence of events that is responsible for the diversity of life (McKelvey, 1982). Carolus Linnaeus, the famous philosopher of the 17th century, has tried to classify plants and animals based on similarities and differences among them (Dupré, 2006). Charles Darwin and his contemporaries biologists of the 19th century have attempted to rebuild the evolutionary history of organisms by building phylogenetic trees (Sidow & Bowman, 1991).

### 5.2 Convergent and divergent evolution

The phylogenetic relationship between organisms is represented by evolutionary distances (Tajima & Nei, 1984). The phylogenetics relationship is basic of taxonomy. It is the domain of science that is responsible for characterizing, naming, and classifying organisms. Taxonomists study and compare the characteristics or traits of different organisms and place them in taxa or groups based on the similarity and differences among them (Sokal, 1966). These taxa may further be clustered depending on biological similarities. The degree of similarity itself does not guarantee biological relatedness. The overall similarity may be ambiguous in deciding the evolutionary connectivity among organisms (Fitch, 2000). There are two routes followed by organisms to get similar characteristics. The first one is if two species descend from common ancestors they will share similar characteristics that are called as homologous characters (Simpson, 1961). For example, the forelimb of humans, bats, dogs, and flippers of whales are homologous because the anatomy of all of them is similar and they have inherited this feature from their common ancestor.

Contrarily, when different species live in similar ecological niches their body parts may differentiate similar manner with the same function to better suit their environment and this model of development is known as convergent evolution (Hubbs, 1944). The traits that are evolved and developed in organisms due to convergent evolution are known as analogous traits or characters (Adams & Nguyen, 2002).

Two species are evolutionarily related if they share homologous similarity (Stern, 2013). Even the highest homology between two organisms does not guarantee that they are closely related to each other (Gould, 1986). Homologous characters maybe recently derived or an ancient conserved feature. Only shared recent homologies can prove that two animals are closely related and have followed the same evolutionary path (Hall, 2012). This phenomenon is known as synapomorphism and such characters are called synapomorphic traits. To make this statement more clear let us take an example of the digitigrade conditions of forelimbs of the ancient vertebrates, which enables them to live on land. Many living terrestrial vertebrates also have five digits, and this suggests the inheritance of this characteristic from their common ancestor.

### 5.3 Concept of cladistics and systematics

William Hennig, a German biologist (1950s), proposed that systematics should infer the known evolutionary history of lineages, and he called this approach as phylogenetic systematics. Thus systematics is more advance as compared to taxonomy, elucidating new methods and theories to classify species (Adams & Nguyen, 2002). Hence, phylogenetic systematics helps us to understand the evolutionary relationships among many different organisms (Adams & Nguyen, 2002). The pioneering work of William Hennig establishes the importance of shared derived characters that can dig-out information about phylogeny (Hennig, 1999). This approach of grouping organisms based on their share-derived characters is called cladistics or phylogenetic systematics and the relationships among the organism are shown in the form of a branching hierarchical tree called a cladogram (Gould, 1986). The information about the closely related organism as well as those distantly related can be traced out from a tree representation.

During earlier times phylogenetic relationships among organisms were generally supported by morphological data. Taxonomists used to collect and examine various characteristics and seek to establish the degree of similarity among organisms (Lee, 2004). We can define cladistics as a subordinate of systematic in which we study phylogenetic association in organisms elicited from shared, derived traits (Rieppel, 2020). There are two types of characteristics, primitive traits and derived traits. *Primitive traits* are characteristics of organisms that were present in the ancestor of the organisms that are under study. They have no role to infer about the evolutionary connection of species within a group since they are inherited from the predecessors to all of the descendent of the clades without major change (Rieppel, 2020). A derived character is one that arose in the lineage preceding a clade and distinguishes members of that clade from others. Derived characters are crucial and very advantageous because they can assist us to know why some species possess common traits or characteristics.

Cladistic and phenetic methods are used to analyze phylogenetic relationships among various groups of organisms. Phenetic methods or numerical taxonomy uses various standards to quantify the overall resemblance for grouping of species (Szalay, 1977). These methods may choose any number or type of trait, which is later changed into a numerical value for analysis (Sneath & Sokal, 1962). The characteristics of organisms groups are pairwise compared with each other and the resemblance of traits is calculated as numerical values. This step is followed by a general clustering approach, which is used to group organisms based on the resemblance. These clusters are known as phenograms and they do not necessarily infer evolutionary relationship. The shared derived traits of organisms are called synapomorphies (Wilkins & Philosophy, 2003). The two important aspects that conclusively revamp the study of molecular phylogenetics are the development of efficient computer algorithms that can generate phylogenetic trees and the availability of nonredundant molecular sequence data (nucleotide or protein) from various organisms for phylogenetic studies, which will enhance the accuracy of results (Yang, 2006). Nowadays we can use both molecular and morphological data to establish phylogenetic resemblance among organisms. Molecular methods are based on studies of either genomic sequences or sequences of the amino acid residue of a peptide chain (Field et al., 1988). The idea behind this approach is that variation in genomes is responsible for variation in traits; thus resemblance between genomes of organisms will assist us to find the taxonomic association among these species. Morphological methods use the phenotypic traits for phylogeny hence and they are conventionally known as character-based phylogeny (Lewis, 2001). Nevertheless, these two methods are associated because it is the genome of an organism that decides the phenotypic traits of that organism. In general practice today, almost all evolutionary relationships among organisms are inferred from either nucleotide sequence of DNA/RNA or amino acid sequences of proteins (Field et al., 1988). For phylogeny, we use molecular data because it is either inherited material (DNA/RNA) or translated product of genetic material. Today with the availability of high-end sequencers and next-generation sequencing methods, we can sequence whole genomes of organisms easily, quickly, at a very cheap cost (Perelman, Johnson, & Roos, 2011). Generally, molecular sequences are highly specific and information rich (Hillis & Huelsenbeck, 1992). Now we use morphological traits to infer evolutionary relationships only in the cases where genetic material is not available like very ancient fossil samples. Nevertheless, this approach is

not very reliable as compared to molecular data because of the events of multiple independent evolutionary lineages, which give rise to the development of analogous traits.

The phylogenetic tree has several nodes and branches in which closely related species share a more recent common ancestor than distantly related species. The nodes represent taxonomic units and the bridge point from where speciation can start. Branches show the relationships of descendants with these nodes. In some phylogenetic trees, branch length differs which usually indicates the rate of evolution and thus evolutionary distance among different clades. The tip of the branches represents the most recent species or actual existing species, which is called the operational taxonomic units (OTUs).

## 5.4 Phylogenetic trees' terminology

The phylogenetic tree shows the evolutionary relatedness among a group of organisms that may be living or extinct. Molecular phylogeny uses molecular sequence data that may be nucleotide or protein sequence, which is orthologous to infer the evolutionary relationship among organisms.

### 5.4.1 Phylogenetic tree

The phylogenetic tree is an acyclic directed graph that represents evolutionary relationships among taxa. The graph is a schematic entity that has some set of nodes that are connected themselves with a set of edges. A phylogenetic tree is a special graph that has no cycles. A connected graph will have a path from any node to any other node. The acyclic graph shows only one path between any two nodes. A graph with at least one cycle is known as a cyclic graph (Fig. 5.1A). A graph with no cycle in it is known as an acyclic graph (Fig. 5.1B).

### 5.4.2 Taxon

A taxon is a group of organism population that can form a unit of taxonomy. Hence, it acts as a unit of classification. All classification divisions, such as species, populations, genera, families, orders, and phyla, may act as a taxon. The plural of the taxon is taxa. The terminal leaf of a phylogenetic tree represents a taxon.

### 5.4.3 Node

A node in a phylogeny constitutes a taxon or a common ancestor for a set of taxa. The terminal nodes of the trees are called OTUs. Nodes are connected with each other with vertices also called edges. The number of edges that is connected to a node is known as the degree of that node (Fig. 5.2).

### 5.4.4 Leaf

A leaf is generally a terminal node with degree 1. A leaf generally represents a single present-day taxon in a phylogenetic tree. We generally have nucleotide or amino acid sequences for each leaf to construct a phylogenetic tree.

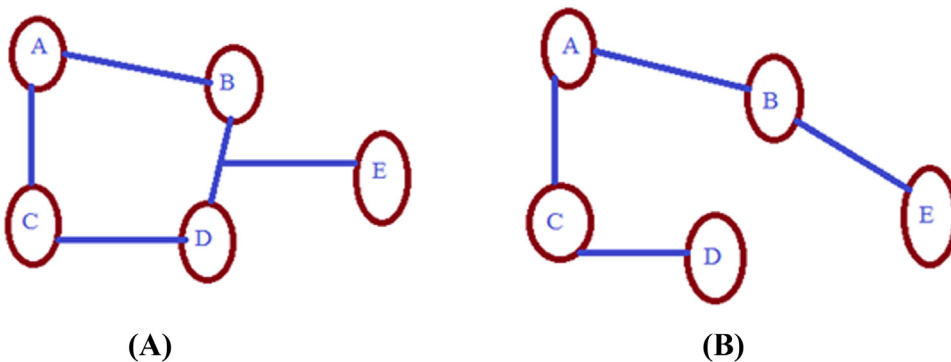


FIGURE 5.1 (A) Cyclic graph and (B) acyclic graph.

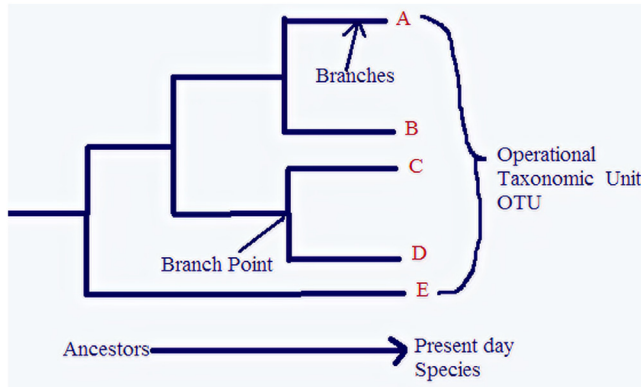


FIGURE 5.2 Phylogenetic tree and its components.

### 5.4.5 Internal node

In a phylogenetic tree, internal nodes represent common ancestors. The degree of an internal node is greater than 1. Generally, we have no sequence data for these internal nodes.

### 5.4.6 Edge

An edge is an entity that connects two nodes in a graph and it is represented by a line. An edge depicts the evolutionary shifting or changes from an ancestral taxon to a descendant taxon. An edge is also called a branch.

### 5.4.7 Topology

A tree topology is a specific structure where many connected elements are arranged like the branches of a tree. The topology represents all of the evolutionary relationships, except time or genetic distance. The directional orientation does not affect the topology of the tree.

### 5.4.8 Root

The root of a tree represents the most recent common ancestor of all taxa in the tree. Hence, it is the oldest known entity in a phylogenetic tree, which tells us the direction of evolution.

### 5.4.9 Rooted tree

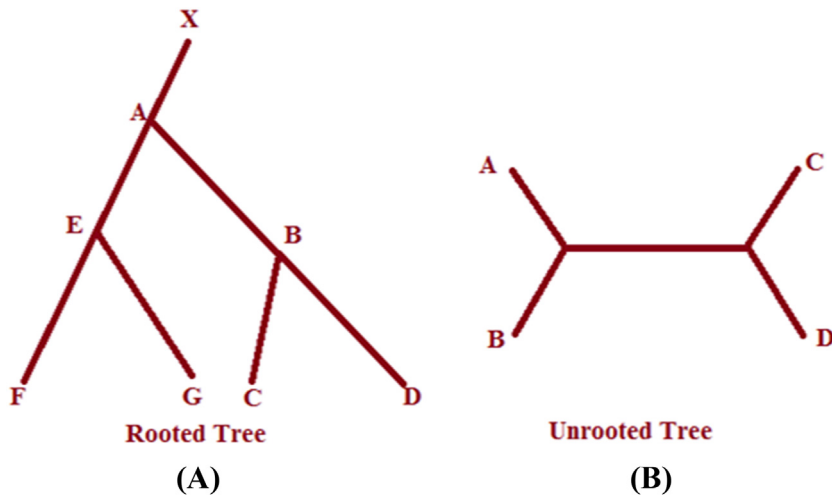
A rooted tree has the last known common ancestor of all taxa. Thus a rooted tree has one vertex labeled as root and every edge is oriented away from this root (Fig. 5.3A).

### 5.4.10 Unrooted tree

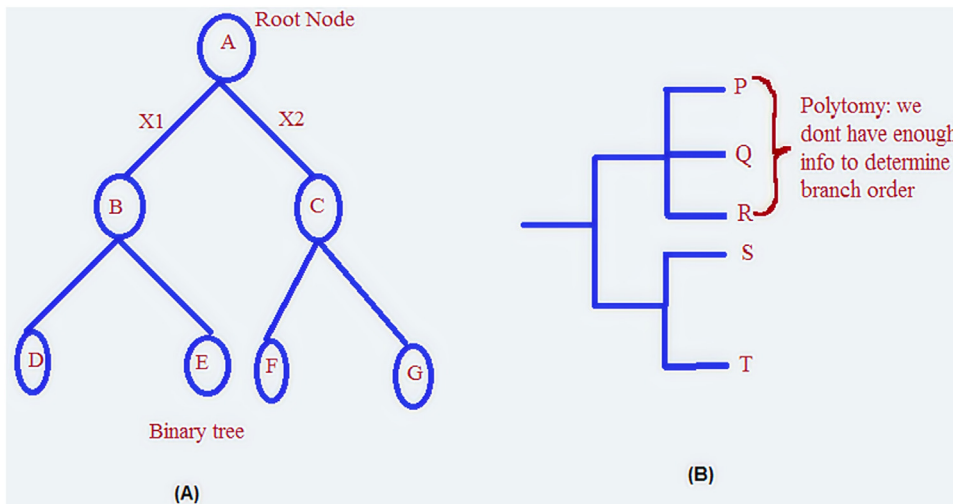
An unrooted tree does not have a root and each vertex has either one or three neighbors. Such trees are drawn without reference to the direction of time (Fig. 5.3B).

### 5.4.11 Binary tree

A binary tree is a tree that has a maximum of two children per parent node except the terminal node (Fig. 5.4A). In a rooted binary tree, leaf nodes have degree 1, the root has degree 2, and all other internal nodes have degree 3. In an unrooted binary tree, all internal nodes have degree 3. If an internal node has more than two descendants, then such trees are known as polytomies (Fig. 5.4B). If there are  $n$  leaves, there are  $n - 1$  internal node in a rooted binary tree and  $n - 2$  internal nodes in an unrooted binary tree.



**FIGURE 5.3** (A) Rooted tree and (B) unrooted tree.



**FIGURE 5.4** Branching conditions in the tree: (A) binary branching and (B) multiple branching.

### 5.4.12 Clade

In a rooted tree, a clade consists of an organism and all of its descendants. A clade is also known as a monophyletic group, meaning that they have a common ancestor that is not a common ancestor for any other leaf in the tree. In an unrooted tree, a clade would be any group of taxa that can be separated from the rest by removing a single edge.

### 5.4.13 Monophyletic taxon

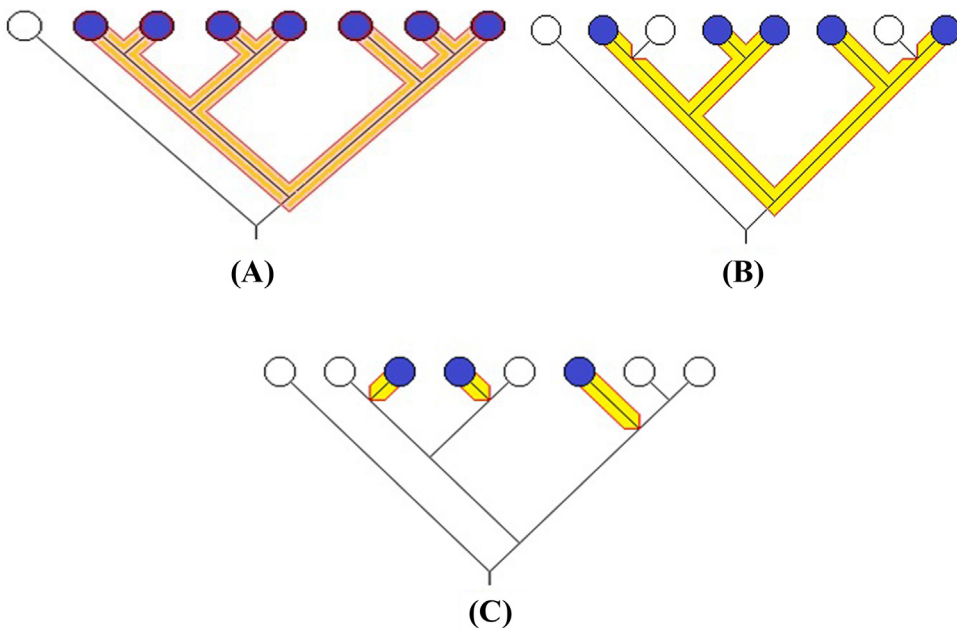
A group of organisms that consists of a common ancestor and all its descendants is known as a monophyletic taxon (Fig. 5.5A). A monophyletic taxon is also called a clade.

### 5.4.14 Paraphyletic taxon

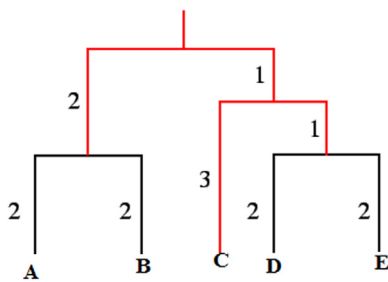
Contrary to the monophyletic group, a paraphyletic taxon does not include all the descendants of the most recent common ancestor (Fig. 5.5B).

### 5.4.15 Polyphyletic taxon

A polyphyletic taxon is one in which all members of the taxon do not share a common ancestor. (Fig. 5.5C). Polyphyletic taxa are considered “unnatural” and usually are reclassified once they are discovered to be polyphyletic.



**FIGURE 5.5** (A) Monophyletic tree; (B) paraphyletic taxon (clade); and (C) polyphyletic taxon (clade).



**FIGURE 5.6** Ultrametric tree in which terminal leaves terminate at same heights.

#### 5.4.16 Split

An event of the partition of the taxa or leaves into two nonempty sets is known as a split. Each edge in a tree represents a split.

#### 5.4.17 Subtree

A subtree of a tree  $X$  is a tree  $Y$  consisting of a node in  $X$  and all of its descendants in  $X$ . The subtree corresponding to the root node is the entire tree; the subtree corresponding to any other node is called a proper subtree.

#### 5.4.18 Edge length

A number associated with an edge in a tree is known as edge length. Generally, it represents divergence time or expected genetic distance among two taxa.

#### 5.4.19 Ultrametric tree

An ultrametric tree follows the molecular clock hypothesis. It is a rooted tree in which the edge lengths of terminal leaves are equidistant from the root (Fig. 5.6). According to the molecular clock hypothesis, here the rate of mutation is the same across all lineages of the phylogenetic tree; as a result, the terminal leaves are at an equal distance from the root (Regan, 1925).

### 5.4.20 Cladogram

A cladogram is a branching tree used for the representation of a phylogeny relationship where the branches or edges are of equal length.

### 5.4.21 Phylogram

A phylogram is a tree used for the representation of a phylogeny relationship. Here, the length of the branch is directly proportional to the time of evolution of the descendent from the parents.

### 5.4.22 Dendrogram

This term is applied to any kind of phylogenetic tree, which is scaled or unscaled. This is the most common output generated from hierarchical clustering.

The total number of the rooted and unrooted tree, which can be constructed using  $n$  taxa, can be calculated with the following formulas.

For given  $n$  taxa, the total number of the unrooted tree can be calculated as:

$$N_{\text{unrooted}} = (2n - 5)! / [2^{n-3} \times (n - 3)!] \quad \text{for } N > 2$$

For given  $n$  taxa, the total number of the rooted tree can be calculated as:

$$N_{\text{rooted}} = (2n - 3)! / [2^{n-2} \times (n - 2)!] \quad \text{for } N > 1$$

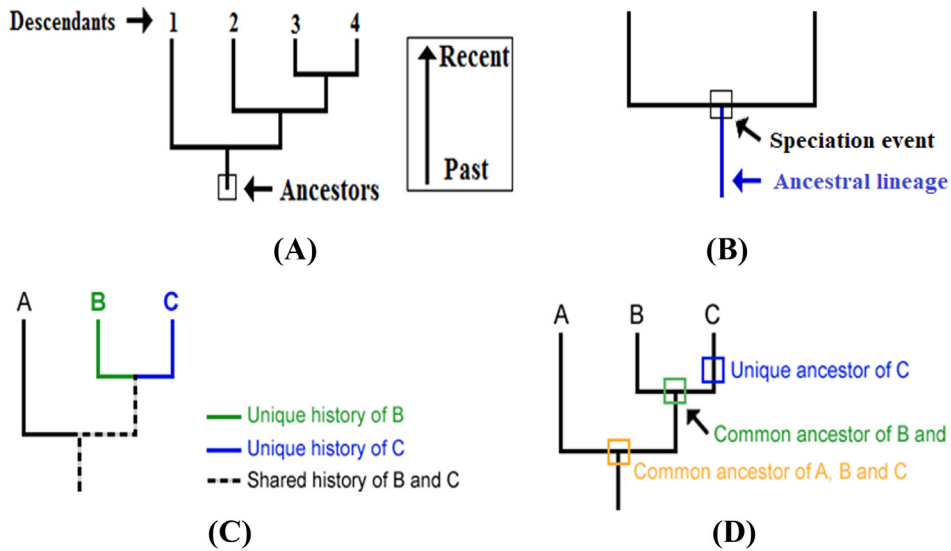
## 5.5 Evolutionary inference of phylogenetic trees

Despite the vast diversity in life forms of this planet, evidence from morphological, biochemical, and gene sequence data of different organisms suggests that all organisms on Earth are genetically related, and this evolutionary relationship of living things can be represented by an extensive evolutionary tree, which is called as the Tree of Life (Forster & Philippe, 1999). The core idea of evolution is that life has a history that has changed over time, which shows different species share common ancestors. The Tree of Life represents the path of how organisms are related evolutionarily. It suggests that different species arise from previous forms via descent and that all organisms, from the smallest microbe to the largest plants and animals, are connected by the passage of genes along the branches of the phylogenetic tree that links all of life (Doolittle, 1999). The core ideas of evolution are that life has a history and it has changed over time showing that different species share common ancestors.

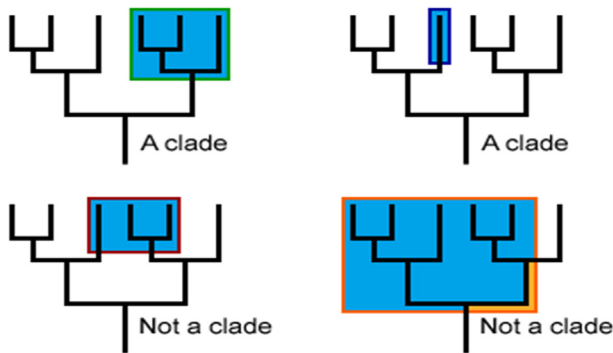
In the general phylogenetic tree, the root of the tree acts for the last known common ancestor, and the tips of the branches represent the descendants of that ancestor (Fig. 5.7A). The rooted phylogenetic tree is directional, which means as we move from the root toward tips, we are supposed to move forward in time. The event of lineage splits introduces branching in a phylogenetic tree and this event is marked as speciation in evolution (Fig. 5.7B). The speciation event gives rise to two or more daughter lineages, resulting in binary or polytomous types of tree. A phylogenetic tree has a unique branching pattern in which each lineage has independent ancestors. Each lineage may either evolve independently or may have some shared ancestry between lineages (Fig. 5.7C). The evolutionary pattern is unique for each lineage, which has unique ancestors that are shared with other lineages (Fig. 5.7D).

Phylogeny can help us to easily assess if a group of lineages forms a clade or not. If we trim a single branch off the phylogenetic tree, all of the organisms on that trimmed branch will make up a new clade. Clades that are snuggled within one another resulted in a nested hierarchy. In a clade, there may be few to many thousands of species. These diagrams represent a different level of hierarchies in a clade (Fig. 5.8). It is also clear from these pictures that how the smaller clades are nested within larger clades.

From the above discussion, it is clear that phylogenetic trees represent patterns of ancestry like family trees. The only difference between a family tree and a phylogenetic tree is that the former can record their own history while the latter cannot do this. This lacuna allows biologists to infer the evolutionary relationship by collecting and analyzing characteristics in terms of morphological, fossils, or molecular sequence data.



**FIGURE 5.7** (A) Direction of evolution; (B) branch split speciation event; (C) tree branching with shared ancestor; and (D) tree branching with unique ancestors.



**FIGURE 5.8** Concept of clades in a tree.

### 5.5.1 Importance of shared derived characters in phylogeny

Some characters may be commonly seen in two lineages and are known as shared characters. Derived characters are derived in the lineage during evolution leading up to a clade. Organisms can be grouped into a clade based on the shared derived characters. These characters are very important to infer the information of evolutionary connections (Fig. 5.9). In vertebrates, four limbs are commonly observed in many organisms indicating that these clades of vertebrates are evolved from a common ancestor.

## 5.6 Tree construction methods

Now we will focus on tree-building methods and algorithms. There are two main approaches to build a phylogenetic tree.

1. Character-based method: this method uses the aligned sequences directly during tree inference; we do not have to take the pairwise distance between sequences.
2. Distance-based methods: this method transforms the multiple sequence alignment data into pairwise distances of sequences and then uses this distance matrix for building a phylogenetic tree, ignoring characters. We will describe the distance- and character-based methods one by one.

### 5.6.1 UPGMA

The UPGMA is an acronym for “Unweighted Pair Group Method with Arithmetic Mean.” This method is very popular among biologists for tree building because of its simplicity (Dawyndt, De Meyer, & De Baets, 2006). UPGMA uses a simple

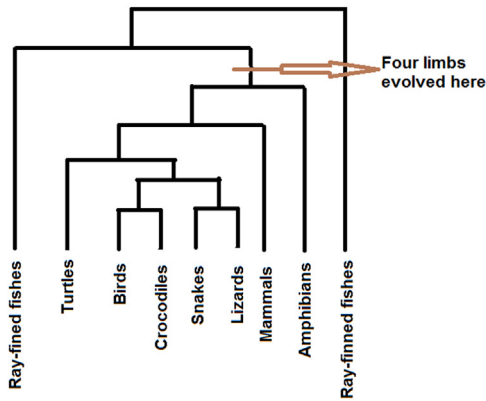


FIGURE 5.9 Shared derived characters.

	A	B	C	D
B	20			
C	60	50		
D	100	90	40	
E	90	80	50	30

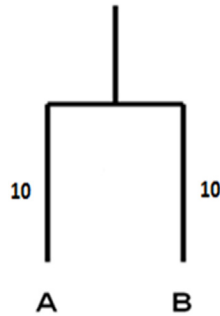


FIGURE 5.10 Joining two operational taxonomic units with minimum distance.

agglomerative clustering approach for phylogenetic tree construction. The method is generally attributed to Sokal and Michener (Sokal, 1958). It was primarily developed for constructing trees that reflect the phenotypic similarities between the OTU. This method is used to construct phylogenetic trees assuming that the rates of evolution among the different lineages are approximately constant and the tree is supposed to follow molecular clock hypothesis. Using the UPGMA approach, we first try to identify two OTUs from the distance matrix, which is closest among all others in terms of the evolutionary distance given in the matrix. The closest pair of the OTUs will be merged as a new single composite OTU (Fig. 5.10). Now a new matrix will be generated in which the distance of all OTUs will be recalculated from the composite OTU. Again the closest pairs will be clubbed together and its distance will be calculated with left-over OTU. This process will be repeated until we are left with only two OTUs. We get the pairwise distances from the multiple sequence alignment of the nucleotide/amino acid sequences.

Now take an example, suppose we have five OTU whose pairwise sequence distances have been given in the distance matrix.

We will apply the following sequential steps to get the final tree:

- First, we must identify the two taxon pairings with the shortest distance, which in this case is A and B, with a distance of 20.
- In the next step, we have to merge A and B and have to calculate the pairwise distance from all other taxa (with A and B). This will result in a new matrix that is one cell smaller than the previous one.
- We will now combine A and B in a binary clad with a common ancestor.
- The distance from parents to A and B will be half the distance between A and B, that is,  $20/2 = 10$ .
- From the parent matrix, we must now generate the next matrix, in which A and B will be fused as (AB). Each distance between AB and the other sequences (C, D, and E) is the mean distance between them and A and B from the original matrix. e.g.  $d(C,AB) = (d(A,C) + d(C,B))/2$ . This way we will calculate all distances to make a new matrix.

Again we will pickup the minimum distance between the two sequences in the new matrix, which is between D and E (30).

We will merge D and E in one group, so the next matrix will be {(AB), C (DE)} of size  $3 \times 3$  and the mean distances among the sequences will be calculated as described above e.g.  $d(AB, DE) = (d(AB,D) + d(AB,E)) / 2$ .

The distance of D and E is joined (Fig. 5.11) from common parents so the distance from parents to D and E will be  $30/2 = 15$ .

Again we will see the minimum distance between the two groups (Fig. 5.12), which is between C and DE (45), we will combine C with DE and calculate the distance among the clade as mentioned above.

Finally, we have a  $2 \times 2$  matrix and we will end at this stage, we have to merge  $\{(AB)\}$  with  $\{(CDE)\}$  Fig. 5.13.

The distance between  $\{(AB)\}$  with  $\{(CDE)\}$  will be divided in equal two half and adjusted as in the previous step so that from both side it becomes  $72.5/2 = 36.25$

UPGMA produce a rooted, ultrametric, binary tree. It follows the molecular clock hypothesis. All branches of the tree terminate at the same height; it means that the rate of evolution is constant along all lineages.

### 5.6.2 Neighbor-joining algorithm

Neighbor joining (NJ) is another distance-based method for tree construction. It uses a clustering approach to construct the tree. It differs from UPGMA in that it does not necessarily produce an ultrametric tree. This method was given by Saitou and Nei (1987). The well-known Clustal package uses this approach to construct a phylogenetic tree (Thompson, Higgins, & Gibson, 1994). In a general preview, NJ can be considered as a special case of star decomposition. For the construction of a phylogenetic tree, NJ keeps track of nodes on a tree rather than taxa or clusters of taxa, which is different from clustering. For the construction of the tree, we need a distance matrix derived from the pairwise hamming distances of the sequences under examination.

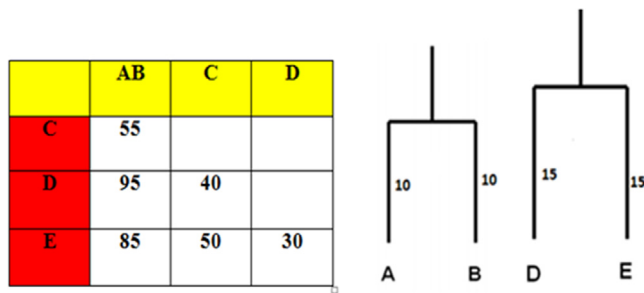


FIGURE 5.11 Joining operational taxonomic units in UPGMA.

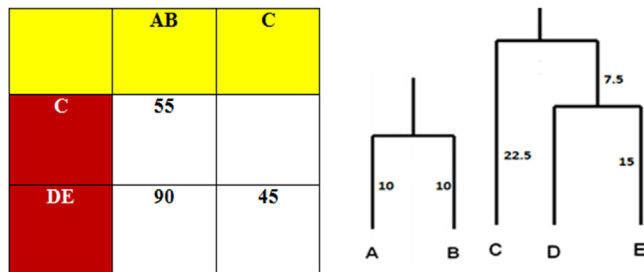


FIGURE 5.12 Joining of other branches.

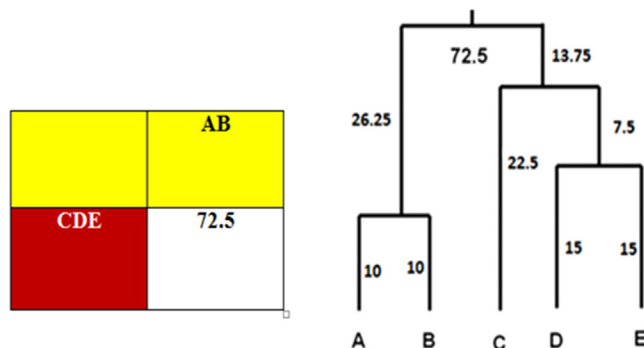


FIGURE 5.13 Final UPGMA tree of five operational taxonomic unit.

### 5.6.2.1 Working example

Suppose, we have to construct the NJ tree using given distance matrix Fig. 5.14. The distance matrix come from the MSA of nucleotide/amino acid sequences from the different OTUs. We will start with a star topology of the tree where all taxa will originate from a common center.

We will construct a modified distance matrix in which the distance between each pair of nodes is assigned, which is calculated from their average divergence to all other nodes.

From this distance table, we will calculate the net divergence  $r_{(i)}$  for each OTU from all other OTUs in the first step:

$$\begin{aligned} r_{(A)} &= 4 + 5 + 2 + 3 = 14 \\ r_{(B)} &= 4 + 7 + 6 + 8 = 25 \\ r_{(C)} &= 5 + 7 + 3 + 4 = 19 \\ r_{(D)} &= 2 + 6 + 3 + 6 = 17 \\ r_{(E)} &= 3 + 8 + 4 + 6 = 21 \end{aligned}$$

Now we will calculate a new distance matrix using the following formula for each OTU pair:  
 $M_{(ij)} = d_{(ij)} - [r_{(i)} + r_{(j)}]/(N - 2)$ .

This will be as follows in the instance of the OTU pair A, B:

$$\begin{aligned} M_{(AB)} &= d_{(AB)} - [(r_{(A)} + r_{(B)})]/(N - 2) = 4 - [14 + 25]/3 = -9 \\ M_{(AC)} &= d_{(AC)} - [(r_{(A)} + r_{(C)})]/(N - 2) = 5 - [14 + 19]/3 = -6 \\ M_{(AD)} &= d_{(AD)} - [(r_{(A)} + r_{(D)})]/(N - 2) = 2 - [14 + 17]/3 = -8.3 \\ M_{(AE)} &= d_{(AE)} - [(r_{(A)} + r_{(E)})]/(N - 2) = 3 - [14 + 21]/3 = -8.7 \end{aligned}$$

We will calculate  $M_{(ij)}$  for all other pairs by using the same iterative approach to construct the following modified matrix (Fig. 5.15).

$$\begin{aligned} M_{(BC)} &= d_{(BC)} - [(r_{(B)} + r_{(C)})]/(N - 2) \\ &= 7 - [25 + 19]/3 = -7.7 \\ M_{(BD)} &= d_{(BD)} - [(r_{(B)} + r_{(D)})]/(N - 2) \\ &= 6 - [25 + 17]/3 = -8 \\ M_{(BE)} &= d_{(BE)} - [(r_{(B)} + r_{(E)})]/(N - 2) \\ &= 8 - [25 + 17]/3 = -7.3 \\ M_{(CD)} &= d_{(CD)} - [(r_{(C)} + r_{(D)})]/(N - 2) \\ &= 3 - [19 + 17]/3 = -9 \\ M_{(CE)} &= d_{(CE)} - [(r_{(C)} + r_{(E)})]/(N - 2) \\ &= 4 - [19 + 21]/3 = -11.3 \\ M_{(DE)} &= d_{(DE)} - [(r_{(D)} + r_{(E)})]/(N - 2) \\ &= 6 - [17 + 21]/3 = -6.7 \end{aligned}$$

	A	B	C	D	E
A	0				
B	4	0			
C	5	7	0		
D	2	6	3	0	
E	3	8	4	6	0

FIGURE 5.14 Distance matrix for calculating neighbor-joining tree.

	A	B	C	D	E
A	0				
B	-9	0			
C	-6	-7.7	0		
D	-8.3	-8	-9	0	
E	-8.7	-7.3	-11.3	-6.7	0

FIGURE 5.15 Modified distance matrix based on net divergence.

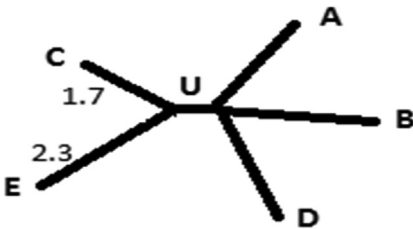


FIGURE 5.16 Merging two operational taxonomic units (C and E) under a common parent U.

Now we will select two OTUs as neighbour for which  $M_{ij}$  is the smallest, which is between C and E (11.3), we will join these two taxa together. Now from our star topology, we have to merge C and E (Fig. 5.16).

We will join C and E with a common ancestor U, and calculate the branch length from the internal node U to the external OTUs C and E with the help of the following formula:

$$S_{(CU)} = d_{(CE)}/2 + [r_{(C)} - r_{(E)}]/2(N - 2)$$

$$S_{(CU)} = 4/2 + [19 - 21]/2 * 3 = 2 + \{[-2]/6\} = 1.7$$

$$S_{(EU)} = d_{(CE)} - S_{(CU)} = 4 - 1.7 = 2.3$$

Now the tree topology will change as given in (Fig. 5.16).

In next step we can define new distances from U to all other terminal node using following formula.:

$$d_{(AU)} = [d_{(AC)} + d_{(AE)} - d_{(CE)}]/2 = [5 + 3 - 4]/2 = 2$$

$$d_{(BU)} = [d_{(BC)} + d_{(BE)} - d_{(CE)}]/2 = [7 + 8 - 4]/2 = 5.5$$

$$d_{(DU)} = [d_{(DC)} + d_{(DE)} - d_{(CE)}]/2 = [3 + 6 - 4]/2 = 2.5$$

The new distance matrix will be created (Fig. 5.17) and will calculate net divergence of all OTUs with this matrix as computed earlier:

$$r_{(U)} = 2 + 5.5 + 2.5 = 10$$

$$r_{(A)} = 2 + 4 + 2 = 8$$

$$r_{(B)} = 5.5 + 4 + 6 = 15.5$$

$$r_{(D)} = 2.5 + 2 + 6 = 10.5$$

We will further create a new matrix  $M_{(ij)}$  as done before with the given formula resulting a new matrix (Fig. 5.18):

$$M_{(ij)} = d_{(ij)} - [r_{(i)} + r_{(j)}]/(N - 2), \text{ here value of } N \text{ will be } 4 - 2 = 2$$

$$M(UA) = 2 - [10 + 8]/2 = -7$$

$$M(UB) = 5.5 - [10 + 15.5]/2 = -7.25$$

$$M(UD) = 2.5 - [10 + 10.5]/2 = -7.75$$

$$M(AB) = 4 - [8 + 15.5]/2 = -7.75$$

$$M(AD) = 2 - [8 + 10.5]/2 = -7.25$$

$$M(BD) = 6 - [15.5 + 10.5]/2 = -7$$

Now we will select and merge the OTU having the lowest  $M_{(ij)}$  value which is 7.75 between UD and AB. We can merge either UD or AB because both have the lowest value, let us join AB this time and make this U1 (Fig. 5.19).

	U	A	B	D
U	0			
A	2	0		
B	5.5	4	0	
D	2.5	2	6	0

FIGURE 5.17 New distance matrix created with merging clad U.

	U	A	B	D
U	0			
A	-7	0		
B	-7.25	-7.75	0	
D	-7.75	-7.25	-7	0

FIGURE 5.18 Modified distance matrix.

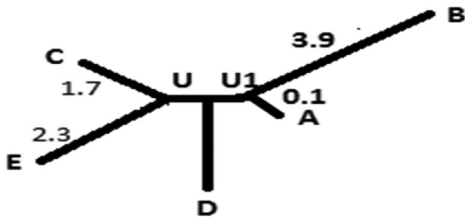


FIGURE 5.19 Merging two branches (A, B) together under common parent U1.

We will merge B and A with common node U1:

$$S_{(AU1)} = d_{(AB)}/2 + [r_{(A)} - r_{(B)}]/2(N - 2) = 4/2 + \{[8 - 15.5]/2 \times 2\} = 2 + \{[-7.5]/4\} = 0.1$$

$$S_{(BU1)} = d_{(AB)} - d_{(AU1)} = 4 - 0.1 = 3.9$$

We will rearrange our tree with the merging of AB under U1 with the calculated distance.

Now we define new distances from U1 to all other terminal node.

$$d_{(UU1)} = [d_{(UA)} + d_{(UB)} - d_{(AB)}]/2 = [2 + 5.5 - 4]/2 = 1.75$$

$$d_{(DU)} = [d_{(DC)} + d_{(DE)} - d_{(CE)}]/2 = [3 + 6 - 4]/2 = 2.5$$

$$d_{(DU1)} = [d_{(DA)} + d_{(DB)} - d_{(AB)}]/2 = [2 + 6 - 4]/2 = 2$$

The new distance matrix will be as follows (Fig. 5.20). We have to calculate net divergence of all OTU with this matrix.

$$r_{(U)} = 1.7 + 2.5 = 4.2$$

$$r_{(U1)} = 1.7 + 2 = 3.7$$

$$r_{(D)} = 2.5 + 2 = 4.5$$

We will create a new matrix  $M_{(ij)}$  as done above with the given formula as done earlier (Fig. 5.21).

$$M_{(ij)} = d_{(ij)} - [r_{(i)} + r_{(j)}]/(N - 2), \text{ here value of } N \text{ will be } 3 - 2 = 1$$

$$M_{(UU1)} = 1.7 - [4.2 + 3.7]/1 = -6.2$$

$$M_{(UD)} = 2.5 - [4.2 + 4.5]/1 = -6.2$$

$$M_{(U1D)} = 2 - [3.7 + 4.5]/1 = -6.2$$

	U	U1	D
U	0		
U1	1.7	0	
D	2.5	2	0

FIGURE 5.20 Distance matrix.

	U	U1	D
U	0		
U1	-6.2	0	
D	-6.2	-6.2	0

FIGURE 5.21 Distance matrix.

We will create a matrix  $M_{(ij)}$ . Because all the values are the same, we can join any one of them. Let us join U with D and make it U2:

$$S_{(DU2)} = d_{(UD)}/2 + [r_{(D)} - r_{(U)}]/2(N - 2) = 2.5/2 + \{[4.5 - 4.2]/2 \times 1\} = 1.25 + \{0.15\} = 1.4$$

$$S_{(UU2)} = d_{(DU)} - S_{(DU2)} = 2.5 - 1.4 = 1.1$$

We will stop at this stage because matrix will reduce to  $[2 \times 2]$  (Fig. 5.22). At this stage, the distance between U1 and U2 will be:

$$d_{(U1U2)} = [d_{(U1D)} + d_{(U1U)} - d_{(DU)}]/2 = [2 + 1.7 - 2.5]/2 = 0.6$$

So this will be our final tree (Fig. 5.23), given the distance matrix at the beginning of this exercise.

### 5.6.2.2 Advantages and disadvantages of the neighbor-joining method

- Advantages
  - It is very fast and thus suited for large datasets and bootstrap analysis.
  - It permits lineages with largely different branch lengths.
  - It permits correction for multiple substitutions.
  - It creates scaled tree.
- Disadvantages
  - The sequence information is reduced.
  - NJ gives only one possible tree.

### 5.6.3 Maximum parsimony

Maximum parsimony is a character-based method of the phylogeny. It means we do not require the distance matrix of the sequences. It is based on Occam's razor problem-solving principle of Monk William of Ockham (1280–1350) which states *Entities should not be multiplied without necessity*. In other words, "The best hypothesis is the one requiring the smallest number of assumptions." The principle of the maximum parsimony approach in phylogeny will identify a tree topology that will require the smallest number of changes to explain the observed differences. In this method, we select the shortest path route to generate the best tree.

	U1	U2
U1	0	
U2	0.6	0

FIGURE 5.22 Distance matrix.

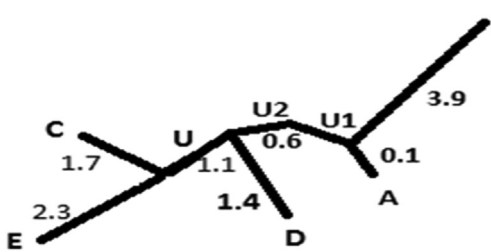


FIGURE 5.23 Final neighbor-joining tree.

	Informative Sites								
	1	2	3	4	5	6	7	8	9
Sequence	-----								
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G
					*	*	*		

FIGURE 5.24 Informative sites in a sequence alignment.

To reduce the complexity of the problem, we use the informative site theory to focus on the partial informative site of the given sequences, which will be used for the phylogeny. It is assumed that not every position of the sequences is informative for inferring the phylogeny or finding the evolutionary relationship between organisms.

An informative site is a position that fulfills two conditions:

- The sites should be occupied by at least two different types of nucleotide.
- The two different nucleotides should present at least twice at that position.

Suppose we have four OTUs having nine sites and we have to construct a phylogenetic tree using the maximum parsimony approach.

First, we have to find the informative sites in the alignment described above. In this alignment, only position numbers 5, 7, and 9 fulfill the information site criteria (Fig. 5.24). On this basis, we can create three possible unrooted trees per site, which are as follows.

We have the following observation.

We have to choose the tree that should have a minimum number of mutations to explain the given tree topology (Fig. 5.25):

- For site 5, tree I requires a total of one mutations and tree II and tree III require only two mutation.
- For site 7, tree I requires a total of one mutation and tree II and tree III require two mutations.
- For site 9, tree I requires a total of two mutations and tree II require one and tree III require only two mutation.

The total for all sites is a minimum (4) for the tree I one so this is the most parsimonious tree (Fig. 5.26).

We can solve this example by taking care of all three information sites together as well:

1. G G A
2. G G G
3. A C A
4. A C G

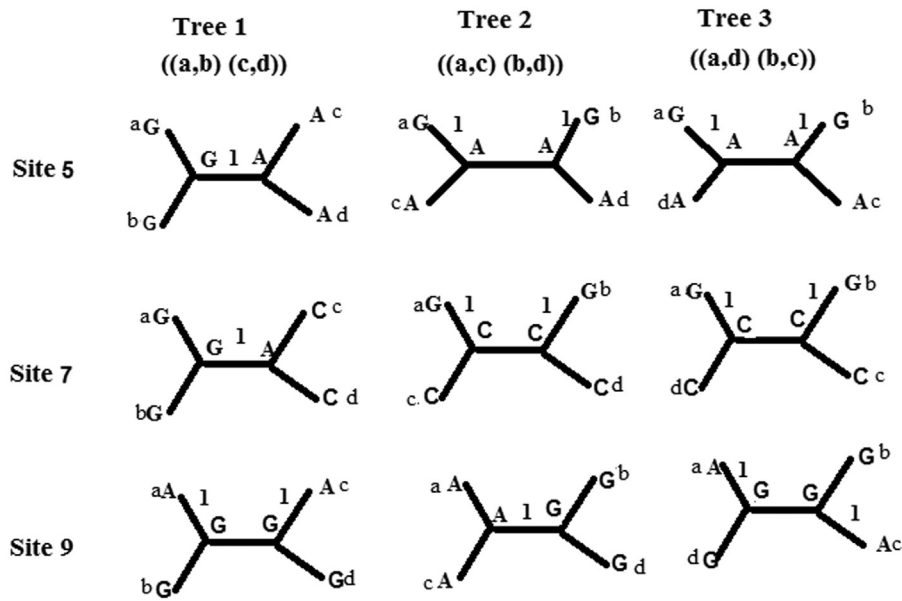


FIGURE 5.25 Total possible mutation over an arm.

Summing changes:

FIGURE 5.26 Total number of mutations required to make the tree.

	site 5	site 7	site 9	Sum
Tree I	1	1	2	4
Tree II	2	2	1	5
Tree III	2	2	2	6

Tree I total sum of changes = 1 + 2 + 1 = 4 (Fig. 5.27A). Tree II total sum of changes = 2 + 1 + 2 = 5 (Fig. 5.27B). Tree III total sum of changes = 1 + 2 + 1 + 1 + 2 = 7 (Fig. 5.27C).

Out of three tree configurations, the total changes required to explain tree configuration are the minimum for tree I; hence, this is the most parsimonious.

In the above example of four OTUs, an informative site can favor only one tree out of the three possible alternatives. This is why site 5 favors tree I over trees II and III, and the best tree is the one that supports the largest number of informative sites.

### 5.6.4 Maximum likelihood phylogeny

The likelihood is a statistical approach of tree construction that provides probabilities of the sequences given a model of their evolution on a particular tree. Given the sequence parameters, the tree with the highest likelihood score is the most desired. It is not possible to consider all possible trees because it is a very computationally intense problem. Since the user can choose a model of evolution, the method can be useful for widely divergent groups or other difficult situations.

This conceptually simple method provides parameter estimates that have good statistical properties. Let us take an example to clear this approach for tree generation. Suppose we have an urn filled with two different color marbles, red and blue. Now we will conduct an experiment where we draw a marble, note down its color, and put it back to the urn. If we assume that the probability of drawing blue marble in such an experiment is  $p$ , then the probability of drawing red marble would be  $q = 1 - p$ . We are interested in estimating probability  $p$  based on the outcome of an independent run of the experiment. For example, one could imagine that 10 independent runs of the experiment result in an observed sequence of B R R B R B R B R R denoted by  $E$ . According to the maximum likelihood method, our best guess of  $p$  is the value that maximizes the probability of observing sequence  $E$ . The probability of observing the sequence  $E$  from the experiment of drawing marble from the urn is:

$$L(p) = \Pr(E; p) = p \times (1 - p) \times (1 - p) \times p \times (1 - p) \times p \times (1 - p) \times p \times (1 - p) \times (1 - p) = 4p \times 6(1 - p) \quad (5.1)$$

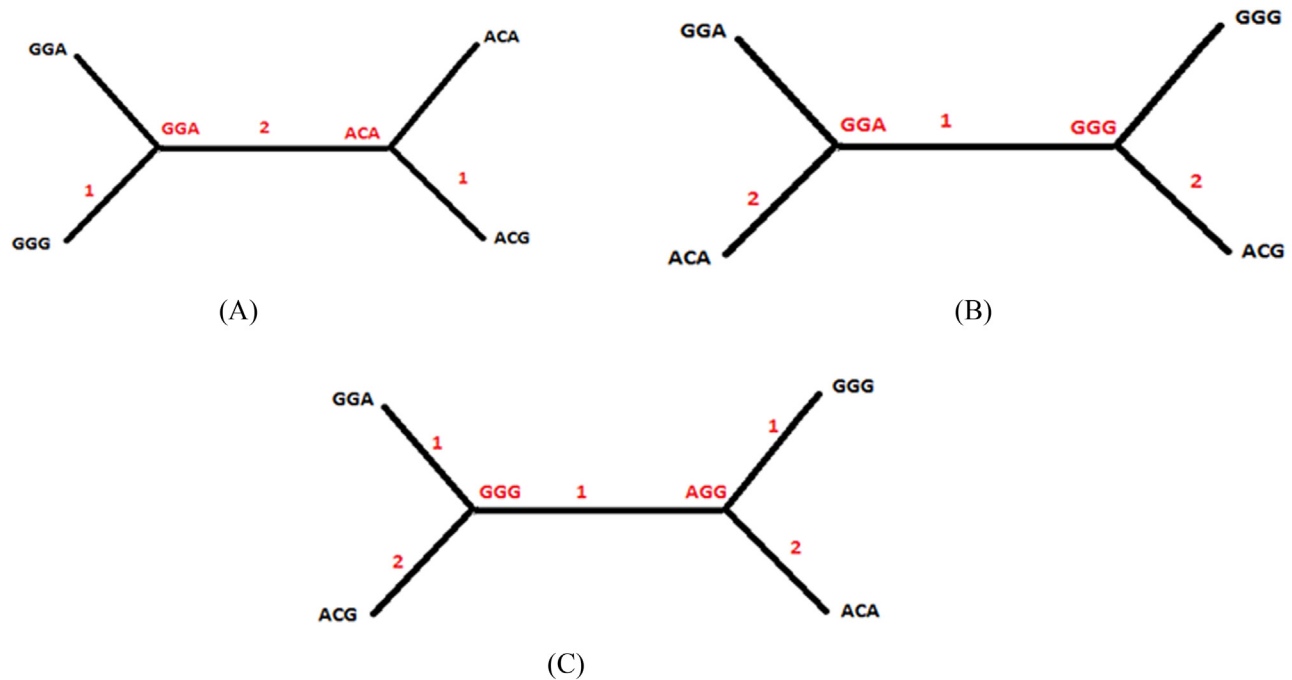


FIGURE 5.27 (A) Branched tree I. (B) Branched tree II. (C) Branched tree III.

Because of the independence of each trial, the probabilities  $p$  and  $1 - p$  in Eq. (5.1) represent the likelihood function of parameter  $p$ . In general, we have to choose the value of  $p$  that maximizes  $L(p)$ . Calculating  $p$  that maximizes a differentiable function with respect to  $p$  can be deduced by solving the equation  $L'(p) = 0$ , which is:  $2(-1 + p)5p^3 - 2 + 5p = 0$

The solution to the above equation results in three possible values (0, 0.4, and 1) of  $p$  that satisfies the equation. From a quick inspection of the likelihood function against the possible values of  $p$ , we can conclude that the maximum likelihood estimate of  $p$  is 0.4. There are four blue marbles observed in the sequence BRBRBRBR; hence, this is a natural estimate of  $p$ . As we repeat the experiment (drawing a marble for the urn) a large number of times, we notice that this maximum likelihood estimator (MLE) of  $p$  tends to the truth. Consistency and efficiency for a large class of MLEs hold under general and reasonable conditions (van der Vaart, 1998). Because of its simplicity and easy-to-understand estimation principle, maximum likelihood has become an important choice among computational biologists for inferring phylogenetic relationships from molecular data from nucleotide or amino acid sequence.

In phylogenetic tree construction using the maximum likelihood approach, we estimate the likelihood function of the molecular sequence as a function of unknown tree topology with branch length, similar to the marble draw experiment (Otu & Sayood, 2003). Substitution models can be used to obtain state transition probabilities given the tree topology with branch lengths. For example, if  $l$  denotes a branch length on a tree, substitution models allow us to calculate  $P_{ij}(l)$ , which denotes the transition probability from state  $i$  to state  $j$  on a branch of length  $l$ , where  $i, j \in \{A, G, C, T\}$ . To estimate the substitution rates and branch lengths, we need some other source of information about speciation/branching and sampling time (Olsen, Matsuda, & Hagstrom, 1994). The followings are the two assumptions that greatly affect the result of phylogenetic analysis:

1. Evolution at different sites (on a given tree) is independent.
2. Evolution proceeds independently on different branches of the phylogeny, which is conditional on the internal node states.

### 5.6.5 Bayesian phylogeny

Bayes' theorem explains how to calculate inverse probabilities. Let us suppose we have an earthen (E), a metal (M), and an ivory (I) urn, each filled with two types of coins: the gold and the other silver (Fig. 5.28). Let us assume given an urn, a coin is chosen uniformly at random.

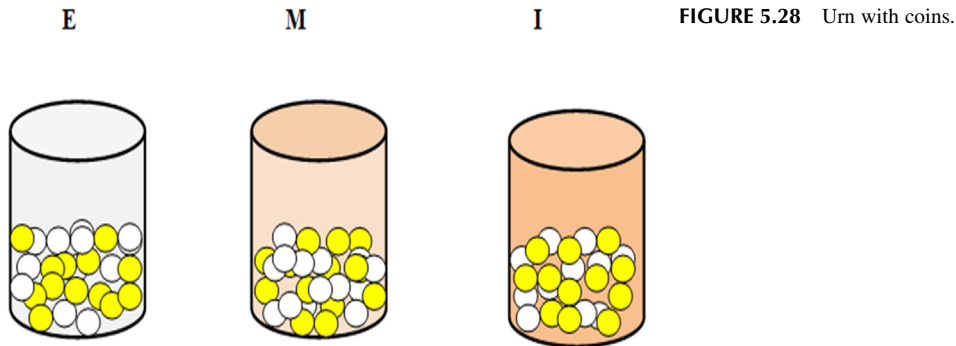


FIGURE 5.28 Urn with coins.

For example, if a coin is chosen from earthen urn E, which has a total of 40 coins and out of which 30 are gold coins, then there is a  $3/4$  chance that the chosen coin will be a gold coin. The inverse problem states if a gold coin is chosen, how likely is it that it came from urn E? To address this problem, we need to estimate the prior distribution for the selection of the urns. Our answer may vary if we believe a priority that urn E is 10% likely to be the chosen than if we believe that all three urns are equally probable.

Bayes' theorem states that "if a complete list of mutually exclusive events  $E_1, E_2, \dots, E_n$  has prior probabilities  $\Pr(E_1), \Pr(E_2), \dots, \Pr(E_n)$  and if the likelihood of the event G given event  $E_i$  is  $\Pr(A|E_i)$  for each I," then  $\Pr(E_i|A)$  is calculated as given below:

$$\Pr(E_i|A) = \frac{\Pr(A|E_i)\Pr(E_i)}{\sum_i \Pr(A|E_i)\Pr(E_i)}$$

The posterior probability of  $E_i$  given A, written  $\Pr(E_i|A)$ , is proportional to the product of the likelihood  $\Pr(A|E_i)$  and the prior probability  $\Pr(E_i)$  where the normalizing constant  $\Pr(A) = \sum_j \Pr(A|E_j) \Pr(E_j)$  is the prior probability of A.

When we apply the Bayesian approach to construct a phylogenetic tree from molecular sequence data, the tree topology may be compared with different urns. The coins represent site patterns that are likely to mutate, of course, there may be several site patterns in a given tree. A prior distribution is a probability distribution on parameters before any data are observed. A posterior distribution is a probability distribution on parameters after data are observed.

Let us say we want to find the posterior probability of a clade. We would need to sum the posterior probabilities of all trees with the clade.

$$\begin{aligned} \Pr(\text{clade}|\text{data}) &= \sum \text{Tree with clade} \frac{\Pr(\text{tree}|\text{data})}{\Pr(\text{data})} \\ &= \sum \text{tree with clade} \frac{\Pr(\text{data}|\text{tree})\Pr(\text{tree})}{\Pr(\text{data})} \end{aligned}$$

## 5.7 Estimating reliability of phylogenetic tree

Inferring the ancestral relationship among organisms is a routine exercise among evolutionary biologists. There are several ways, through which we can make a phylogenetic tree based on either distance- or character-based criteria. We know that there are several trees possible (rooted and unrooted) for N taxon. The challenging part is the reliability of tree topology and branching, that is, estimating who are the parents of whom and who are the siblings. To solve this problem, we take the help of statistics. There are two sampling techniques, through which we can access the reliability of the internal branching of a tree.

We use resampling to reuse data to generate new, hypothetical samples called resample and these resampled data are representative of an underlying population. It is used when:

- No idea of underlying distribution for the population;
- The situation when traditional formulas are difficult to apply; and
- This approach is also used as a substitute for traditional approaches.

There are two popular tools, a bootstrap and jackknife, which are used in estimating the reliability of the phylogenetic tree. Although they have many similarities, they do have a few notable differences also.

### 5.7.1 Bootstrapping

Bootstrapping is any test or metric that uses random sampling with replacement and falls under the broader class of resampling methods. It uses sampling with replacement to estimate the sampling distribution for the desired estimator. This approach is used to assess the reliability of sequence-based phylogeny.

Bootstrap values in a phylogenetic tree indicate that out of 100, how many times the same branch is observed when repeating the generation of a phylogenetic tree on a resampled set of data. If we get this observation 100 times out of 100, then this supports our result. In this condition, we are confident that the observed branch in the relationship is not due to a single extreme data point. If we recover the same node through 95 of 100 iterations of taking out one character and resampling of our tree, then we have a good idea that the node is well supported with a bootstrap value of 95%. Bootstrap values less than 50% are not taken into account for tree construction.

In bootstrap resampling, we generate artificial sequences by choosing randomly sampling sites of the original sequences with replacement. This exercise produces a different composition of the data with the same sequence length as the original data sets. To understand it let us take the following example. Suppose we have a simple sequence with six sites. Now we have chosen randomly a site that lets site number 3 from the original sequence. Now in each sequence, site 3 is placed in the first position and the rest of the position is randomly exchanged in repeated steps until the bootstrap sequence is also 6-bp long as shown in Fig. 5.29.

This whole process is repeated at least 100 times and the number of times a clade is seen among the bootstrap trees is reported. Higher the bootstrap value, the more confident we are that the observed branch is not due to a single extreme data point.

## 5.8 Phylogenetic tools

There are a large number of tools available online to analyze sequences and to construct phylogenetic trees (Table 5.1). Among those tools, clustal (<http://www.clustal.org/>) is used for multiple alignments of nucleic acid and protein sequences; the recent version of clustal is clustal omega, which is very fast and scalable for large datasets (Sievers, Wilm, & Dineen, 2011). Clustal is available for windows mac and Linux machines. The second tool that is very useful for phylogenetic analysis is MEGA, which is used for sequence alignment, generating phylogenetic trees, database mining, and estimating rates of molecular evolution (<http://www.megasoftware.net/>) (Tamura, Dudley, Nei, & Kumar, 2007). Phylogeny.fr (<http://www.phylogeny.fr/>) is used for the generation and analysis of phylogenetic relationships between molecular sequences, such as protein and DNA (Dereeper, Guignon, & Blanc, 2008). PHYLogeny Inference Package (PHYLP) is a free package of programs for inferring phylogenies (Felsenstein, 1993). PHYLP is based on methods of parsimony, distance matrix, and likelihood methods of tree construction. Phylogenetic Analysis Using Parsimony is commercial software for the analysis of evolutionary relationships using maximum parsimony, maximum likelihood, and distance methods (<http://paup.csit.fsu.edu/>) (Wilgenbusch & Swofford, 2003).

Phylemon2.0 (<http://phylemon.bioinfo.cipf.es>) consists of a collection of 30 tools for various applications of bioinformatical analysis (Tárraga, Medina, & Arbiza, 2007). The recent version of Phylemon uses Web 2.0 technology features, with a new user interface. PhyloGene server has the facility to analyze, for identifying and visualizing the coevolving proteins using normalized phylogenetic profile data (<http://genetics.mgh.harvard.edu/phylogene/>). It predicts protein function and also identifies new members of the pathway and disease genes (Sadreyev, Ji, & Cohen, 2015). DIVEIN requires a set of aligned sequences as an input and has the facility to input many basic parameters related to

	Original sequence	Bootstrap sequence
Human	A T G A C C	G T A A C A
Rat	A T A A C T	A T A A C A
Mouse	A T A A C T	A T A A C A
Chimpanzee	A T G A C T	G T A A C A
	↓ Site 3	↓ Position-1

FIGURE 5.29 A sequence with six sites.

Site-3 is placed at position one in bootstrap sequence and next five randomly chosen sites: 2, 1, 1, 5, 4 are placed in next five position.

**TABLE 5.1** List of tools and software used for phylogenetic analysis.

Tools name	Description	Reference
Clustal package	This tool is used for multiple sequence alignment. It is available both for command line as well as graphical mode ( <a href="http://www.clustal.org">http://www.clustal.org</a> )	Thompson, Toby, and Des (2003)
MEGA	It is widely used for sequence alignment, generating phylogenetic trees, database mining, and estimating rates of molecular evolution. ( <a href="http://www.megasoftware.net">http://www.megasoftware.net</a> )	Tamura et al. (2007)
PHYLIP	This is a free program for inferring phylogeny using several available approaches like maximum likelihood, maximum parsimony, and distance matrix ( <a href="https://evolution.genetics.washington.edu/phylip.html">https://evolution.genetics.washington.edu/phylip.html</a> )	Felsenstein (1993)
PAUP	PAUP is a commercial software for the analysis of evolutionary relationships using maximum parsimony, maximum likelihood, and distance methods ( <a href="http://paup.csit.fsu.edu">http://paup.csit.fsu.edu</a> )	Posada (2003)
Phylemon2.0	This server consists of a collection of 30 tools for various applications of bioinformatical analysis ( <a href="http://phylemon.bioinfo.cipf.es">http://phylemon.bioinfo.cipf.es</a> )	Sánchez et al. (2011)
DIVEIN	It is widely used for tree construction, evolutionary models, and bootstrapping ( <a href="https://indra.mullins.microbiol.washington.edu/DIVEIN/">https://indra.mullins.microbiol.washington.edu/DIVEIN/</a> )	Deng et al. (2010)
UGENE	It is a free bioinformatics software for multiple sequence alignment, genome sequencing data analysis, and amino acid sequence visualization ( <a href="http://ugene.net/">http://ugene.net/</a> )	(Okonechnikov, 2012)
PhyML	Fast and accurate estimation of phylogenies using maximum likelihood	Guindon et al. (2010)
QuickTree	Tree construction optimized for efficiency using neighbor-joining approach ( <a href="https://github.com/khowe/quicktree">https://github.com/khowe/quicktree</a> )	Howe, Bateman, and Durbin (2002)
TOPALi	Statistical and evolutionary analysis of multiple sequence alignments ( <a href="http://www.topali.org/">http://www.topali.org/</a> )	Milne et al. (2004)
Treefinder	Fast ML tree reconstruction, bootstrap analysis, model selection, hypothesis testing, tree calibration, tree manipulation and visualization, computation of site-wise rates, and sequence simulation ( <a href="https://www.treefinder.de/">https://www.treefinder.de/</a> )	Abby, Tannier, Gouy, and Daubin (2010)
DAMBE	It is a general-purpose package for DNA and protein sequence phylogenies, and also gene frequencies	Xia and Xie (2001)
SeaView	It is a multiplatform graphical user interface for sequence alignment and phylogenetic tree building	Gouy, Guindon, and Gascuel (2010)

*DAMBE*, Data Analysis in Molecular Biology and Evolution; *PAUP*, Phylogenetic Analysis Using Parsimony; *PHYLIP*, PHYLogeny Inference Package.

tree construction, evolutionary models, and bootstrapping (<https://indra.mullins.microbiol.washington.edu/DIVEIN/>) (Deng et al., 2010).

## 5.9 Application of molecular phylogeny

Phylogenetics study is important as it improves our knowledge and understanding of the evolution of genes, genomes, species, and other molecular sequences. Phylogenetics help us to understand how nature has shaped the path of evolution in terms of molecular sequences. It reveals how mutations have changed the sequence, which we have today, and how they will change its future. There are many core area of biology, which is solely dependent on the evolution and to understand the evolution at the molecular level we have to take the help of phylogeny.

### 5.9.1 Classification

Phylogenetics-based studies provide us much more accurate information of patterns of relatedness or closeness and differences among the organisms (Carbayo, Álvarez-Presas, & Olivares, 2013; Field et al. 1988). With the advent of molecular sequencing and more advanced algorithms, molecular phylogeny gives us a more reliable classification than

was available before. Molecular phylogenetics now reshape the Linnaean classification with the addition of new closely related species.

### 5.9.2 Identifying the origin of pathogens

Molecular sequencing information and phylogenetic analysis can be used to get more knowledge about a new pathogen as well as its outbreak (Li, Wang, & Evans, 2019). The species or taxa of a pathogen can be identified as well as its source; host and mode of transmission can also be identified using the bioinformatics analysis. The recent outbreak of the Covid-19 virus and traceback to its origin from Wuhan is a classical example of this (Guo, Cao, & Hong, 2020). The correct identification and origin of the pathogens can lead to new recommendations for public health policy.

### 5.9.3 Species conservation

Phylogenetics can be used for the conservation of endangered species. In this way, a species can be prevented to reach the stage of extinction (Moritz, 1995). The genetic data of Asiatic tigers and African lions helped a lot the conservationist to design strategies for their population growth (Wilting, Courtiol, & Christiansen, 2015). Bioinformatics approaches play a central role in phylogenetic analysis. It starts with downloading homologous sequences from the databases to aligning the sequences using various software. The development of more robust and powerful algorithms to handle large sequences is also a challenge that is tackled by bioinformaticians and computational biologists (Andreatta & Ribeiro, 2002).

## 5.10 Conclusion

The biotic diversity on this planet is tremendous; nature has introduced an enormous variety of morphological differences among organisms for better adaptation to their habitat. Since the origin of life on this earth, about 3.7 billion years back until date evolution has shaped organisms in various ways and it is still on and will continue in the future too. The biologist has tried to infer the path, relationship, and mode of this evolution among organisms using various approaches. Because we cannot see in the past directly, we need to develop some models and assumptions based on the indirect molecular pieces of evidence we have. There are several tools to draw phylogenetic relationship among organisms; the list is neither exhaustive nor complete; the search for a better algorithmic approach to creating a better model will always be open for the new researchers to give their input; and this is the way nature brings changes and opens a new dimension for the organisms to adapt and survive.

## Conflict of interest

The authors declare that they have no conflict of interest.

## References

- Abby, S. S., Tannier, E., Gouy, M., & Daubin, V. (2010). Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. *BMC Bioinformatics*, *11*, 324.
- Adams, B. J., & Nguyen, K. B. (2002). Taxonomy and systematics. *Entomopathogenic nematology*, *1*.
- Andreatta, A. A., & Ribeiro, C. C. (2002). Heuristics for the phylogeny problem. *Journal of Heuristics*, *8*(4), 429–447.
- Carbayo, F., Álvarez-Presas, M., Olivares, C. T., et al. (2013). Molecular phylogeny of Geoplaninae (Platyhelminthes) challenges current classification: Proposal of taxonomic actions. *Zoologica Scripta*, *42*(5), 508–528.
- Dawyndt, P., De Meyer, H., & De Baets, B. (2006). UPGMA clustering revisited: A weight-driven approach to transitive approximation. *International Journal of Approximate Reasoning*, *42*(3), 174–191.
- Deng, W., Maust, B. S., Nickle, D. C., Learn, G. H., Liu, Y., Heath, L., . . . Mullins, J. I. (2010). DIVEIN: A web server to analyze phylogenies, sequence divergence, diversity, and informative sites. *Biotechniques*, *48*(5), 405–408.
- Dereeper, A., Guignon, V., Blanc, G., et al. (2008). Phylogeny.fr: Robust phylogenetic analysis for the non-specialist. *Nucleic Acids Research*, *36* (suppl\_2), W465–W469.
- Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science*, *284*(5423), 2124–2128.
- Dupré, J. (2006). Scientific classification. *23*(2–3), 30–32.
- Felsenstein, J. (1993). *PHYLIP (phylogeny inference package), version 3.5c*. Joseph Felsenstein.
- Field, K. G., Olsen, G. J., Lane, D. J., et al. (1988). Molecular phylogeny of the animal kingdom. *Science*, *239*(4841), 748–753.
- Fitch, W. M. (2000). Homology: A personal view on some of the problems. *Trends in Genetics*, *16*(5), 227–231.

- Forterre, P., & Philippe, H. (1999). Where is the root of the universal tree of life? *Bioessays*, 21(10), 871–879.
- Gould, S. J. (1986). Evolution and the triumph of homology, or why history matters. *American Scientist*, 74(1), 60–69.
- Gouy, M., Guindon, S., & Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, 27(2), 221–224.
- Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3), 307–321.
- Guo, Y.-R., Cao, Q.-D., Hong, Z.-S., et al. (2020). The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak – an update on the status. *Military Medical Research*, 7(1), 1–10.
- Hall, B. K. (2012). *Homology: The hierarchical basis of comparative biology*. Academic Press.
- Hennig, W. (1999). *Phylogenetic systematics*. University of Illinois Press.
- Hillis, D., & Huelsenbeck, J. J. (1992). Signal, noise, and reliability in molecular phylogenetic analyses. *The Journal of Heredity*, 83(3), 189–195.
- Howe, K., Bateman, A., & Durbin, R. (2002). QuickTree: Building huge Neighbour-Joining trees of protein sequences. *Bioinformatics (Oxford, England)*, 18(11), 1546–1547.
- Hubbs, C. L. (1944). Concepts of homology and analogy. *The American Naturalist*, 78(777), 289–307.
- Lee, M. S. Y. (2004). The molecularisation of taxonomy. *Invertebrate Systematics*, 18(1), 1–6.
- Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, 50(6), 913–925.
- Li, J., Wang, T., Evans, J. D., et al. (2019). The phylogeny and pathogenesis of Sacbrood virus (SBV) infection in European honey bees, *Apis mellifera*. *Viruses*, 11(1), 61–67.
- Margulis, L., & Chapman, M. J. (2009). *Kingdoms and domains: An illustrated guide to the phyla of life on Earth*. Academic Press.
- McKelvey, B. (1982). *Organizational systematics—Taxonomy, evolution, classification*. University of California Press.
- Milne, I., Wright, F., Rowe, G., Marshall, D. F., Husmeier, D., & McGuire, G. (2004). TOPALI: Software for automatic identification of recombinant sequences within DNA multiple alignments. *Bioinformatics (Oxford, England)*, 20(11), 1806–1807.
- Mora, C., Tittensor, D. P., Adl, S., et al. (2011). How many species are there on Earth and in the ocean? *PLoS Biology*, 9(8), e1001127.
- Moritz, C. (1995). Uses of molecular phylogenies for conservation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 349(1327), 113–118.
- Okonechnikov, K, et al., & UGENE team. (2012). Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics*, 28(8), 1166–1167.
- Olsen, G. J., Matsuda, H., Hagstrom, R., et al. (1994). FastDNAMl: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Bioinformatics*, 10(1), 41–48.
- Oparin, A. I. J. T. (1957). *The origin of life on the earth* (3rd ed.).
- Otu, H. H., & Sayood, K. J. (2003). A new sequence distance measure for phylogenetic tree construction. *Bioinformatics*, 19(16), 2122–2130.
- Perelman, P., Johnson, W. E., Roos, C., et al. (2011). A molecular phylogeny of living primates. *PLoS Genetics*, 7(3), e1001342.
- Posada, D. (2003). Using MODELTEST and PAUP\* to select a model of nucleotide substitution. *Current Protocols in Bioinformatics*, 6–5.
- Regan, C. T. (1925). *Organic evolution I*. Nature Publishing Group.
- Rieppel, O. (2020). Morphology and phylogeny. 53(2), 217–230.
- Sadreyev, I. R., Ji, F., Cohen, E., et al. (2015). PhyloGene server for identification and visualization of co-evolving proteins using normalized phylogenetic profiles. 43(W1), W154–W159.
- Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. 4(4), 406–425.
- Sellés-Martínez, J. (1996). Concretion morphology, classification and genesis. 41(3–4), 177–210.
- Sidow, A., & Bowman, B. H. (1991). Molecular phylogeny. 1(4), 451–456.
- Sievers, F., Wilm, A., Dineen, D., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. 7(1), 539.
- Simpson, G. G. (1961). Principles of animal taxonomy.
- Sneath, P. H., & Sokal, R. R. (1962). Numerical taxonomy. 193(4818), 855–860.
- Sokal, R. R. (1966). Numerical taxonomy. 215(6), 106–117.
- Sokal, R. R. (1958). A statistical method for evaluating systematic relationships. 38, 1409–1438.
- Stern, D. L. (2013). The genetic causes of convergent evolution. 14(11), 751–764.
- Szalay, F. S. (1977). Ancestors, descendants, sister groups and testing of phylogenetic hypotheses. 26(1), 12–18.
- Sánchez, R., Serra, F., Tárraga, J., Medina, I., Carbonell, J., Pulido, L., . . . Dopazo, H. (2011). Phylemon 2.0: A suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. *Nucleic Acids Research*, 39(Web Server issue), W470–W474.
- Tajima, F., & Nei, M. (1984). Estimation of evolutionary distance between nucleotide sequences. 1(3), 269–285.
- Tamura, K., Dudley, J., Nei, M., & Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, 24(8), 1596–1599.
- Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. 22(22), 4673–4680.
- Thompson, J. D., Toby, J. G., & Des, G. H. (2003). Multiple sequence alignment using ClustalW and ClustalX. *Current Protocols in Bioinformatics*, 2–3.
- Tárraga, J., Medina, I., Arbiza, L., et al. (2007). Phylemon: A suite of web tools for molecular evolution, phylogenetics and phylogenomics. 35 (suppl\_2), W38–W42.

- Wilgenbusch, J. C., & Swofford, D. (2003). Inferring evolutionary trees with PAUP. (1), 6.4.1–6.4.28.
- Wilkins, J. S. (2003). How to be a chaste species pluralist-realist: The origins of species modes and the synapomorphic species concept. *18*(5), 621–638.
- Wilting, A., Courtiol, A., Christiansen, P., et al. (2015). Planning tiger recovery: Understanding intraspecific variation for effective conservation. *1*(5), e1400175.
- Woese, C. R. (2000). Interpreting the universal phylogenetic tree. *97*(15), 8392–8396.
- Xia, X., & Xie, Z. (2001). DAMBE: Software package for data analysis in molecular biology and evolution. *The Journal of Heredity*, *92*(4), 371–373.
- Yang, Z. (2006). *Computational molecular evolution*. Oxford: Oxford University Press.