Chapter 13
Introduction to Phylogenetic Reconstruction

Phylogenetics, the study of evolutionary relationships in organisms, is one part of the larger field of systematics, which also includes taxonomy. The term **taxonomy** connotes the process and methodology for the naming and classification of organisms. The context of evolutionary biology is phylogeny, the connections between all groups of organisms as understood by ancestor/descendant relationships. The molecular mechanisms of organisms studied strongly suggests that all organisms on earth have a common ancestor. Thus, the species are related to each other by the virtue of having evolved from the same common (now extinct) ancestor. Such a relationship of species is called *phylogeny* and it's graphical representation is called a *phylogenetic tree*.

Computational methods infer these relationships from currently thriving species and reconstruct what their course of evolution might have been. The phylogenetic tree construction help us go back in time and develop a "hypothesis" of how life evolved from the single common ancestor. This hypothesis (a phylogenetic tree) is represented as a cladogram, a branching diagram. Cladograms bear a lot in common with the notion of family trees. In a family tree we trace back our ancestry. For example, in the family tree on the right, the ancestors of all the rest of the family are the initial black dot and yellow square. These ancestors give rise to three children, one of which mates and has two children. We can all trace our lineage back to one set of ancestors.

All species have ancestors too. So, for example, sometime in the past an ancestral species (father) of Homo sapiens walked the earth. This ancestor went extinct (died), but left descendant species (children). In family trees, we can talk coherently about real ancestors. In biology, the ancestors are often gone sometimes without a trace. All we have left are the children. Also unlike family trees, new species form from the splitting of old species, and the formation of the two descendant species is called a splitting event. The ancestor is usually assumed to "die" after the splitting event.

© Springer International Publishing Switzerland 2015

G.B. Singh, *Fundamentals of Bioinformatics and Computational Biology*,
Modeling and Optimization in Science and Technologies 6, DOI: 10.1007/978-3-319-11403-3_13

13.1 Terminology

The nodes of the tree in a cladogram are marked. The stems of the tree end with the taxa under consideration. Each node marks a splitting event. The node therefore represents the end of the ancestral taxon, while the stems represent the species that split from the ancestor. The two taxa that split from the node are called *sister taxas*. They are called sister taxa because they are like the siblings from the parent or ancestor in a family tree. The sister taxa must each be more closely related to one another than to any other group because they share a close common ancestor.

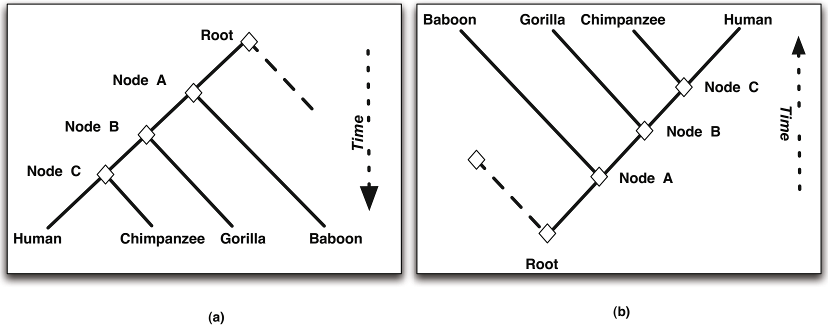


Fig. 13.1 Implicit in the representation of a phylogenetic tree is the passage of evolutionary time shown along the vertical axis. Two commonly utilized representations of the the phylogenetic tree are shown above. In (a) the passage of time begins with the one common ancestor at the top of the tree and continues down to the taxons currently in existence, namely, the species human, chimpanzee, gorilla and baboon. In an equivalent representation shown in (b) the passage of evolutionary time is from the bottom to the top. The time axis is seldom labeled on the tree as it evident from the series of speciation events resulting in the number of species becoming larger with the passage of time.

Two equivalent representations for phylogenetic trees are shown in Figure 13.1. The most closely related species in the cladogram are *humans* and *chimpanzees*. Their common ancestor is represented by *Node C*. At the node, the ancestor goes extinct but leaves two siblings hypothesized to be humans and chimpanzees. The humans and chimpanzees are known as **sister taxas** and are more closely related to each other than to any other taxa on the tree. Working through the tree we come to *Node B* – a node where the ancestor for *human* and *chimpanzees* split from *gorillas*. The *gorilla* is a sister taxa to the *human* and *chimpanzee* ancestor. Similarly, the common ancestor of the now extinct species represented by *Node B* and the *baboon* is another extinct species denoted by *Node A*.

In practice, one does not label the internal ancestor as the phylogenetic tree is usually derived from the known taxas which are labeled on the leaf nodes. Sometimes, when the common ancestor has been identified through fossil remains, an ancestral node may also be labeled. An ancestor plus all its descendants is called a **clade**. A cladogram thus shows us the hypothesized *clades*.

13.1.1 Tree Representation Formats

Trees are often graphically represented and drawn in a two-dimensional space. For example, such a representation for a 6-taxa tree is shown in Figure 13.2(a). Individual taxons (a, b, c, d, e, f) are shown as the leaves of the tree and their distance from their parents are marked on the edges. For example, the taxon-*a* has diverged 2 units from its common parent that it shares with taxon-*b* which has diverged 2.5 units from the same common parent. The units of evolutionary distance is *years*. However, in molecular phylogenetic tree reconstruction methods, often the divergence of a species is estimated by the divergence of a *gene* family. In these instances, which anyway is our focus, the distance annotated on the edge could be a measure or a metric produced by a sequence comparison program.

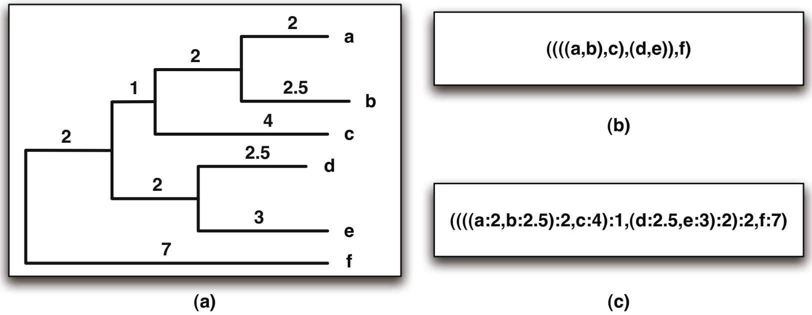


Fig. 13.2 (a) Topology of a tree with edge lengths shown. (b) Textual representation or symbolic expression of the tree without the edge lengths – only topological information is encoded. (c) Symbolic expression for the phylogenetic tree including the encoding of the edge lengths.

As shown in Figure 13.2(b), the tree topology may be represented using a text representation or a **symbolic expression**. The level of nesting of the parenthesis in this representation is equal to the depth of the leaf node. For example, the taxon-*a* and taxon-*b* are at a depth of 4, and they are also at the fourth level of parenthesis nesting in the symbolic representation.

The symbolic representation of tree may also be extended to include the edge length, as shown in Figure 13.2(a).

13.2 Types of Trees

13.2.1 Unrooted and Rooted Trees

Phylogenetic relationships between the genes of organisms are used to form (a) unrooted, or (b) rooted trees. The branching pattern in either case is called the *topology* of the tree for a given number of *taxa*. A *taxa* is defined to be any kind of taxonomic unit such as families, species or DNA sequences. The true biological phylogeny has a "root" – the ultimate ancestor of all the taxas and their ancestors. While some algorithms that construct phylogenetic tree provide information on the what might be the root, others (notably parsimony and probabilistic methods) do not yield any information pertaining to the location of the root.

Figure 13.3 presents both a rooted and an unrooted tree corresponding to a cladogram where ancestors of individual taxas are correspondingly defined. In the rooted tree, the root is generally drawn at the top and all of the taxa is drawn as terminal nodes, or leaves, at the base of the tree. The unrooted tree is a branching structure with no clear top to bottom hierarchy. The internal nodes in an unrooted tree represent the ancestors of the taxa which terminate the branches. The complete definition of a tree includes the edge length along with the topology of the branching patterns.

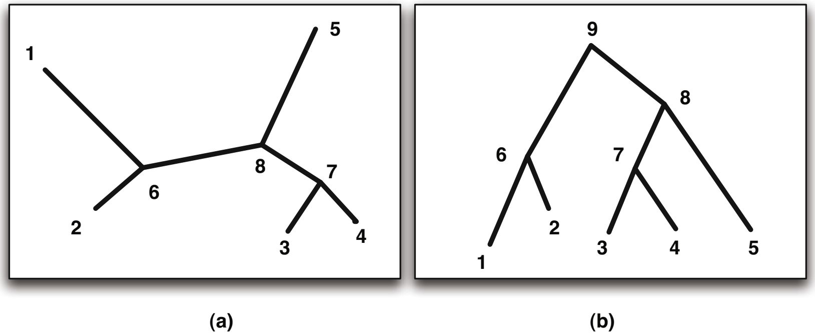


Fig. 13.3 (a) An Unrooted and (b) a Rooted tree. Note that the number of nodes in the unrooted tree is one less than the number of nodes in the rooted tree for the same number of taxas (five in this example shown). This is because the unrooted trees do not designate a root node – the node representing the one common ancestor to all the species in the phylogenetic tree.

13.2.2 Orthologues and Paralogues

Scientists are often interested in studying the phylogenetic tree that represents the evolutionary history of a group of species or populations. In this type of tree, the time of divergence between two species refers to the duration of time that the two species have been reproductively isolated. However when the phylogenetic tree is constructed based on one gene from each species, the tree obtained does not necessarily agree with the species tree. This is observed because of the presence of polymorphic alleles at a given gene locus, which results in the expected time of divergence of genes sampled from different species to be longer than the time for the divergence of species themselves. Thus we can expect the branching patterns obtained from the phylogenetic tree constructed from the gene, often referred to as the **gene tree**, to be different from the branching pattern in species phylogenetic tree.

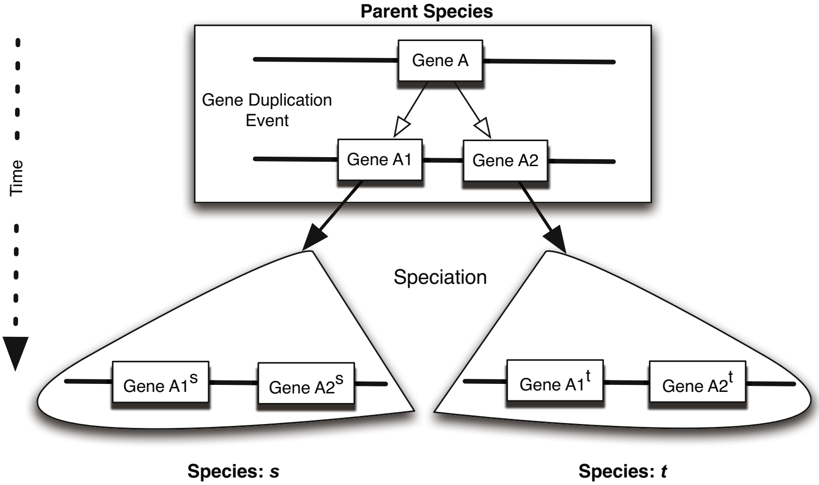


Fig. 13.4 Gene is duplicated in the parent species prior to the speciation. Genes A1 and A2 are paralogs. As the species evolve and two new species s and t are formed, the genes $A1^s$ and $A1^t$ as well as $A2^s$ and $A2^t$ are orthologs. The pairs of genes $A1^s$ and $A2^s$, or $A1^t$ and $A2^t$, or $A1^s$ and $A2^t$, or $A2^s$ and $A1^t$ are also paralogs.

Another problem occurs when the gene studied belongs to a multigene family. For example, consider the situation shown in Figure 13.4. Two related species s and t have evolved from their ancestor where the duplication event occurred before the divergence of the two species. The result of the gene duplication produced genes $A1$ and $A2$. These are known as the **paralogs**. After the speciation event, the genes in the divergent species s and t have been denoted as ($A1^s$, $A2^s$) and ($A1^t$, $A2^t$). In the species s and t, gene pairs ($A1^s$, $A1^t$) and ($A2^s$, $A2^t$) are known as the **orthologues**. The gene pairs ($A1^s$, $A2^s$), ($A1^t$, $A2^t$), ($A1^s$, $A2^t$) and ($A2^s$, $A1^t$) are known as the **paralogues**.

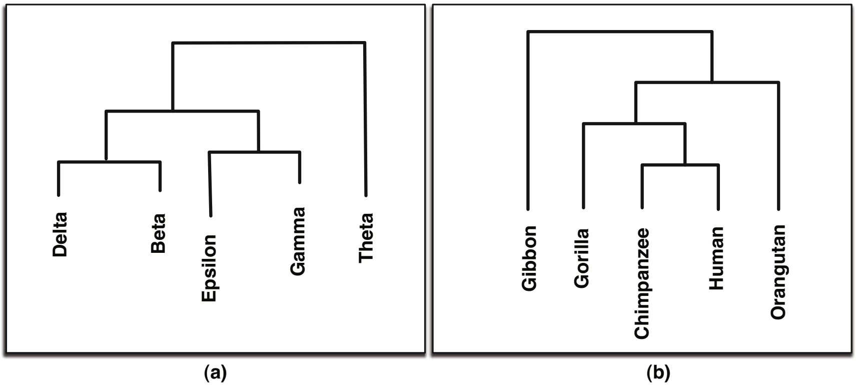


Fig. 13.5 (a) A phylogenetic tree constructed using paralog genes, and (b) A phylogenetic tree construction using ortholog genes. Both these trees are **gene trees**. The gene tree in (a) does not correspond to a species tree but represents the history of duplication events leading to the evolution of the beta-, delta-, epsilon-, gamma- and theta-chains of the human haemoglobin gene family. In contrast, the gene tree in (b) is developed with a kilobase fragment of mitochondrial DNA from five primate species. Here, the diversity of the mitochondrial genes are representative of the evolutionary history of the primate species themselves.

Scientist take great care to use the appropriate type of genes for the construction of appropriate tree. For example, if one is interested in studying the evolutionary history of gene duplication events, as shown in Figure 13.5(a), the paralogues from a single specie(s) will be used in the phylogenetic tree construction process. On the other hand, if the objective is to develop a evolutionary tree of the species, such as the one shown in Figure 13.5(b), one must utilize orthologues for the tree construction algorithm. In practice, the distinction between the orthologous and paralogous genes is not always easy, particularly when there are many copies of the duplicate genes in the genome. Therefore, great care must be taken to infer a species tree from a gene tree.

13.3 Counting Phylogenetic Trees

This section utilizes basic combinatorics to develop a formulation for the total number of possible rooted and unrooted trees that may be constructed given a number of taxas, n. Basic formulations for the total number of vertices and edges is used to inductively develop the formulation for the total number of trees.

Theorem 1. *Every rooted phylogenetic tree with n-taxas is an acyclic graph with (2n-1) vertices and (2n-2) edges.*

Proof: Let us begin by counting the number of vertices in a phylogenetic tree with n taxas (i.e. leaf nodes). Intuitively, we can begin the construction of the tree by linking the two closest sister taxas and denoting them as having been evolved from the same common ancestor. Thus, we add one vertex for the common ancestor in the process of linking the two taxa leaves *(n-2)*. The addition of one parent vertex results in reducing the problem size to the construction of a tree with $(n - 2 + 1) = (n - 1)$ vertices. Continuing further, the addition of each parent vertex will result in the reduction of the problem size by 1. Assuming that a rooted phylogeny is being constructed, the final problem size is 1, corresponding to the one root vertex (the single common ancestor) that is left when this process is complete. This requires the addition of $(n - 1)$ parent vertices. Consequently, the total number of vertices in a rooted phylogenetic tree is $(n + (n - 1))$ or $(2n - 1)$.

As for the number of edges in the tree we can reason as follows. Since each of the $(n - 1)$ steps in the construction process results in the addition of 2 edges, the total number of edges in a rooted tree is $2(n - 1)$ or $(2n - 2)$. Alternatively, this result may be arrived at by observing that the number of edges in a tree is one less than the number of vertices. A tree with two vertices has a single edge. A tree with three vertices can only have two edges as a tree may not have any cycles. A rooted phylogenetic tree is a therefore an acyclic graph $G_r = (2n - 1, 2n - 2)$.

Theorem 2. *Every unrooted phylogenetic tree with n-taxas is an acyclic graph with (2n-2) vertices and (2n-3) edges.*

Proof: The proof for an unrooted tree is very similar to the proof for a rooted tree. In an unrooted tree, the construction process is terminated when there are 2 vertices remaining which are then connected with an edge, yielding the final tree. Since there are $(n - 2)$ parent nodes added, the total number of vertices in the final tree, which includes the n taxas is $(n + (n - 2))$ or $(2n - 2)$. Correspondingly, there are are $(2n - 3)$ edges that are utilized in the construction of an unrooted tree. An unrooted tree is therefore an acyclic graph $G_u = (2n - 2, 2n - 3)$.

With that basic background, the task of counting the total number of possible topologies given the number of taxas to be n is considerably simplified. Once again let us first consider the number of possible topologies of rooted trees. The number of rooted trees possible with $n = 2$ is 1. This is shown in Figure 13.6(a). A given two-taxa tree may be extended to a three-taxa tree by adding the third taxon in one of two ways. Firstly, the third taxon could be linked to any of the existing edges of the two-taxa tree. Secondly, the third taxon could directly evolve from the one common ancestor – the root node. Since the two-taxa tree has two edges, the total number of three-taxa tree topologies possible is 3, as shown in Figure 13.6(b). The third taxon C evolves from a common ancestor of A or B, or it evolves from an earlier common ancestor from which an ancestor of both A and B evolved. Continuing a similar analysis, for the total number of topologies possible for four-taxas

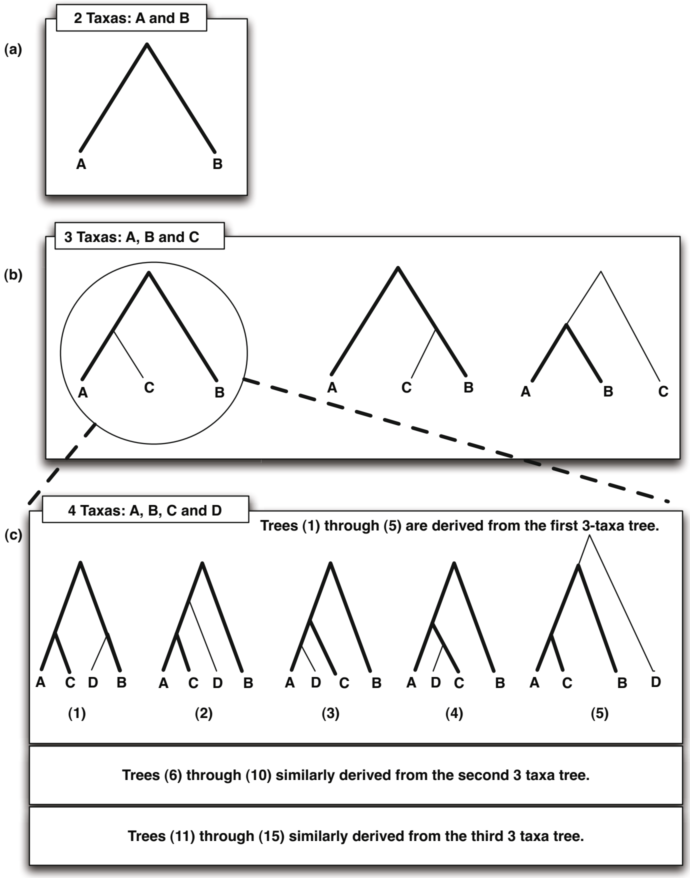

(a)
2 Taxas: A and B
A
B
(b)
3 Taxas: A, B and C
A
C
B
A
C
B
A
B
C
(c)
4 Taxas: A, B, C and D
Trees (1) through (5) are derived from the first 3-taxa tree.
A C D B
A C D B
A D C B
A D C B
A C B D
(1)
(2)
(3)
(4)
(5)
Trees (6) through (10) similarly derived from the second 3 taxa tree.
Trees (11) through (15) similarly derived from the third 3 taxa tree.

Fig. 13.6 Given two taxons, A and B, the total number of rooted trees possible is "1" as shown in (a). The third taxon in (b) may be added from any of the two branches of the two node tree or may be directly linked to a new root node. This yields three possible rooted trees with three taxons A, B and C. Each of the three-taxon trees further yields 5 trees, each resulting in a total of 3 × 5 or 15 possible trees with four taxons A, B, C and D as shown in (c) . In general the number of rooted trees possible with n taxons in $(2n\text{-}3)!!$.

we observe from Figure 13.6(c) that each of the three-taxa topologies may be extended by five different ways, yielding the total count of rooted tree topologies for four-taxas to be (3×5) or 15. The process of counting phylogenetic trees can be inductively generalized as follows:

Theorem 3. *The total number of possible topologies for rooted phylogenetic tree with n-taxas is* $(2n-3) \times (2n-5) \times \ldots \times 1$ = *(2n-3)!!.*

Proof: The correctness of this formula can be shown by induction. Let us denote the count number of possible trees with i taxas to be R_i.

Base case: For three taxas $i = 3$, the number of trees possible is $R_3 = 3$. The formula 3!! = 3 is correct.
Basis: Assume the formula is correct for a number of taxas = *(n-1)*. Therefore, the number rooted tree topologies for $(n-1)$ taxas is $2(n-1) - 3 \times 2(n-1) - 5 \ldots 1$
i.e. the number of tree rooted topologies is $R_{n-1} = (2n-5) \times (2n-7) \times \ldots 1$.

Inductive Step: Based on theorem 1, each of the possible *(n-1)*-taxa tree has $(2(n-1)-2)$ or $(2n-4)$ edges. The n^{th} taxon can thus branch off from any of these edges. Alternatively, the n^{th} taxon could be a direct descendant of a *new* common ancestor designated as the new root node. This gives us $(2n-4+1)$ or $(2n-3)$ ways of extending each of the *(n-1)*-taxa trees. Thus, the total number of n-taxa trees is:
$R_n = (2n-3) \times R_{n-1}$
$R_n = (2n-3) \times (2n-5) \times \ldots \times 1 =$ (2n-3)!!.
This concludes the proof. The formula works for the base case. Assuming the formula works for the basis case of $(n-1)$, the formula is shown to work for n in the inductive step. Therefore the formula is correct for $n \geq 3$.

Theorem 4. *The total number of possible topologies unrooted phylogenetic tree with n-taxas is* $(2n-5) \times (2n-7) \times ... \times 1$ = *(2n-5)!!.*

Proof: The proof for this theorem is by inspection. Let us denote the count number of possible trees with i taxas to be U_i.

According to theorem 2, an unrooted tree with n-taxas has $(2n-3)$ edges. Each of these edges may be the site where a root of the rooted tree is inserted. The number of rooted trees is therefore $(2n-3)$ times the number of unrooted trees. This is illustrated in Figure 13.7. This completes the proof.

13.4 Comparing Phylogenetic Trees

Often reconstructed trees need to be compared to measure the extent of topological differences between them. The distance between two unrooted trees is defined using a **topological distance**. This method is based on the comparison of partitions in the two trees. A partition is created by removing an internal edge of an unrooted tree. In general, the method uses the notion of tree partitioning obtained by removing an internal edge from the two trees being compared. Figure 13.8 shows the partitions formed by removing the internal edges of a 6-taxa bifurcating unrooted tree. For a bifurcating tree, which is different from a multifurcating tree, where more than two edges may emanate from a vertex, the number of partitions with n sequences as the leaves is $(n-3)$. This is evident from the observation that there are $(n-2)$ internal nodes in an unrooted tree which in turn must be connected with $(n-3)$ edges. Thus for a bifurcating tree with n-taxa, there are $(n-3)$ possible partitions.

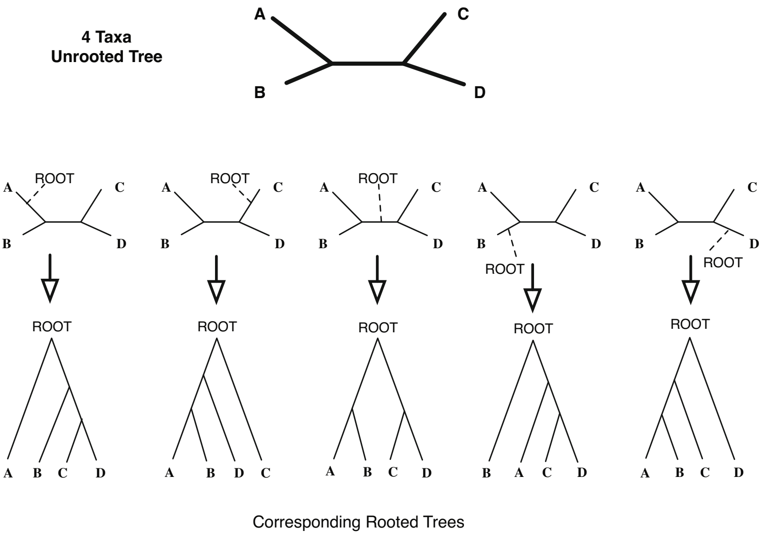
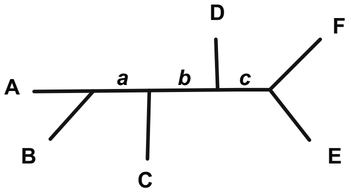


Fig. 13.7 Each n-taxa unrooted tree may be transformed to $(2n-3)$ rooted trees as a root node may be inserted at any one of its edges. In the example shown above 5 rooted trees can be generated from a given 4-taxa unrooted tree corresponding to the inserting of a root at any of the 5 edges of the unrooted tree. Note that since there are three topologies possible for 4-taxa unrooted trees, there are 15 corresponding topologies possible for 4-taxa rooted trees.



Internal Edge Removed	Partitions Created	
a	(A,B)	(C,D,E,F)
b	(A,B,C)	(D,E,F)
c	(A,B,C,D)	(E,F)

Fig. 13.8 A partition of a tree is defined to the two subsets of taxa (leaves) created by removing an internal edge. As there are $(n-2)$ internal vertices of an unrooted n-taxa tree, it will have $(n-3)$ distinct partitions. For example, the 6-taxa tree shown above has 3 partitions.

While comparing two trees, all possible partitions of the two trees are created. Let there be q_1 partitions for the tree T_1 and q_2 partitions for the tree T_2. Although both the trees are defined on the same set of leaves or taxons, it is possible that q_1 and q_2 are not equal as the two trees may not be strictly bifurcating. Furthermore, let there be p partitions $(p \leq \min(q_1, q_2))$ from the trees T_1 and T_2 that are identical. The topological distance $d_T(T_1, T_2)$ between T_1 and T_2 is defined by Eq. 13.2.

$$d_T(T_1, T_2) = 2 \cdot [min(q_1, q_2) - p] + |q_1 - q_2| \tag{13.1}$$

For strictly bifurcating trees, $q_1 = q_2 + q$ and the above equation reduces to Eq 13.2. This yields $0 \leq d_T(T_1, T_2) \leq 2(n-3)$ as the range for topological distance between two n-taxa bifurcating trees.

$$d_T(T_1, T_2) = 2 \cdot [q - p] \tag{13.2}$$

13.5 Evolution

Many groups of organisms are now extinct, and without their fossils we would not have as clear a picture of how modern organisms are interrelated. We express the relationships among groups of organisms through diagrams called cladograms, which are like genealogies of species. Over the last 3.7 billion years or so, living organisms on the Earth have diversified and adapted to almost every environment imaginable. The diversity of life is truly amazing, but all known living organisms do share certain similarities. All living organisms can replicate, and the replicator molecule is DNA. As well, all living organisms contain some means of converting the information stored in DNA into products used to build cellular machinery from fats, proteins, and carbohydrates.

13.6 Phylogenetic Tree Object in Matlab

MATLAB provides the function phytree(B) for creating an ultrametric phylogenetic tree object, where B is a numeric array of size `[NUMBRANCHES × 2]` where every row represents a branch of the tree and it contains two pointers to the branches or leaves nodes which are its children.Leaf nodes are numbered from 1 to `NUMLEAVES` and branch nodes are numbered from `NUMLEAVES` + 1 to `NUMLEAVES` + `NUMBRANCHES`.

For example, the following array specifies a phylogenetic tree with four leaves and three branches:

```
B = [1 2; 3 4; 5 6];
Tree = phytree (B);
phytreeviewer (Tree);
```

Additionally, a seccond parameter to this function may be used:

```
TREE = phytree(B,C)
```

This creates an additive phylogenetic tree object where branch distances defined by C. C is a numeric array of size `[NUMBRANCHES × 1]` with the distances of every branch added to the tree.

```
C = [0.5 ; 2.0 ; 4.0];
TreeWithDist = phytree (B, C);
phytreeviewer (TreeWithDist)
```

Fig. 13.9 illustrates the result of building a phylogenetic object in MATLAB.

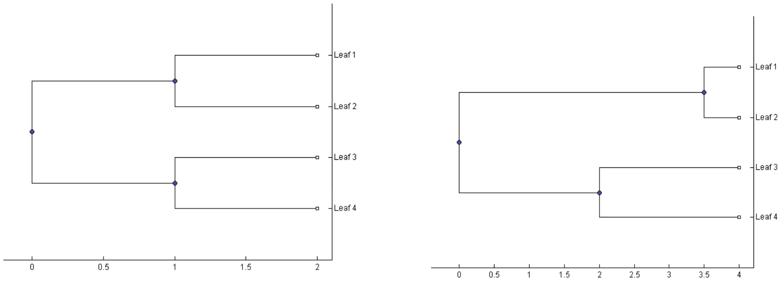


Fig. 13.9 Result of constructing a phylogenetic tree object in MATLAB. The tree on the left is constructed where distances between the nodes have not been specified. The tree on the right includes a specification of distances. Both trees are ultrameric trees where the evolutionary distance between all the leaves and the root node is the same.

MATLAB also provides functions *phytreeread* and *phytreewrite* which are used for reading and writing a tree to a file respectively.

13.6.1 Phylogenetic Trees in BioPerl

PHYLIP is a computational phylogenetics package containing many precompiled programs that are used to build phylogenetic trees. In order to estimate an evolutionary tree from sequence data, data must be passed through various PHYLIP programs, the output of which is then used as input for the next program in line. BioPerl enables a user to automate this pipeline through wrapper classes that interact with the various programs (these programs are not included in BioPerl and must be obtained separately). BioPerl takes each PHYLIP program's output and either stores it in memory or writes it to a supported file type such that BioPerl can then feed this output as input into the next applicable program. The input for such a pipeline consists of some number of supposedly related biological sequences and the output is a tree estimating the sequences' evolution.

An example pipeline can be used to generate a phylogenetic tree is one consisting of, in order, ClustalW, SeqBoot, ProtPars, Consense, and lastly DrawGram. The first step in this pipeline is to use BioPerl to generate a multiple-sequence alignment of the input sequences. This is accomplished by way of BioPerl's wrapper class for ClustalW.

Listing 13.1

```
use Bio::Tools::Run::Alignment::Clustalw;
@params = (’ktuple’ => 2,
    ’matrix’ => ’BLOSUM’);
$factory = Bio::Tools::Run::Alignment::Clustalw->new(@params);
$seq_file = "sequence_file.fasta";
$aln = $factory->align($seq_array_ref);
```

end-listing-13.1

Next, each of the PHYLIP programs is used in turn—piping the output data from the previous program as input data into the next. Each of these programs possesses it’s own wrapper class in BioPerl of the name Bio::Tools::Run::Phylo::Phylip::*, where * is the name of the PHYLIP program. Each program is in turn initialized and run in the following form, where $var_0 is the previous output.

Listing 13.2

```
$factory = ClassName->new(%params);
$var_1 = $factory->run($var_0);
```

end-listing-13.2

So the first PHYLIP program that is run is SeqBoot, which is a bootstrapping program. Next, we pass the bootstrapped alignment to ProtPars, a parsimony program that estimates phylogenies. The results from ProtPars are then passed to Consense, which computes the consensus trees. Finally, the Consense output is drawn as a rooted tree by DrawGram.

Listing 13.3

```
# SeqBoot
my $seqboot_factory =
Bio::Tools::Run::Phylo::Phylip::SeqBoot->new();
my $aln_ref= $seqboot_factory->run($aln);
# ProtPars
$tree_factory = Bio::Tools::Run::Phylo::Phylip::ProtPars-> new();
$tree = $tree_factory->run($aln_ref);
# Consense
my $con_factory = Bio::Tools::Run::Phylo::Phylip::Consense->new();
my $tree = $con_factory->run($tree);
# Drawgram
my $draw_factory = Bio::Tools::Run::Phylo::Phylip::DrawTree->new();
my $image_filename = $draw_factory->draw_tree($tree);
```

end-listing-13.3

If this sample is run on the following Fast input file, the produced tree will look like Figure 13.10.

Listing 13.4

```
>sequence_1
ACGTACGTACGT
>sequence_2
ACGTACGTACGTACGT
>sequence_3
ACGTACGTACGTTGCA
>sequence_4
ATCGATCGATCG
>sequence_5
ATCGATCGATCGATCG
```

end-listing-13.4

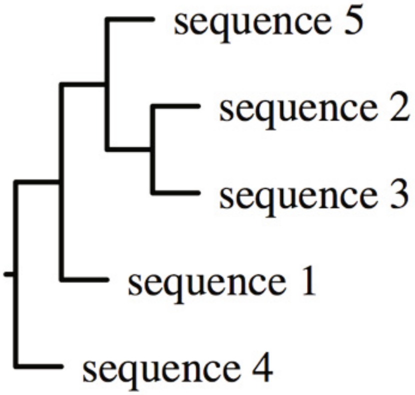


Fig. 13.10 The resulting phylogenetic tree produced by the PHYLIP pipeline

13.7 Significance of Trees Constructed

The accuracy and validity of comparative genomic conclusions drawn are directly related to the accuracy and validity of the phylogenetic tree reconstructed inferred using a phylogeny reconstruction method. It is therefore desirable to be able to assign a quantitative "quality" parameter to the tree as the phylogenetic reconstruction methods are often prone to making errors resulting from inferring wrong branches, complex biological processes such as recombinant and horizontal gene transfers which can not be modeled as a single tree, as well as due to issues related with inaccuracy and incompleteness of data. Consequently, various methods have been introduced for quantitatively estimating *confidence* in the branch of an reconstructed phylogenetic tree.

Of the two common ones, namely the bootstrap method and Bayesian inference techniques, the former is discussed in somewhat greater detail below.

13.7.1 Bootstrapping

The bootstrap method is usually used for trees generated by maximum parsimony (MP) or maximum likelihood (ML) methods. The confidence in a brach is computed by estimating many trees over subsamples of the dataset and using the the percent of trees containing that branch as a measure of its support.

Original sequence data is sub-sampled to to produce **new input data** of the same length. The result of the sampling process may create duplicates of the original sites (columns in the multiple sequence alignment) or may completely eliminate certain sites. Notwithstanding, the sub-sampling process maintains statistical similarity between the new dataset and the original input data. A phylogenetic tree is subsequently constructed with each new data set using the particular method of interest.

The support or confidence in a labeled tree is constructed by taking the majority consensus of the set of trees created during the bootstrapping iterations. Essentially then the support for a branch is its likeness to the a majority rule consensus tree created after the bootstrapping analysis.

The bootstrapping method may be applied for evaluating the significance of trees generated by distance based methods as well. An extra step is need to first compute the distance matrix from each of the replicate data sets produced as a result of sub-sampling the original data set.

Further Readings

1. Schuh, R.T., Brower, A.V.Z.: Biological systematics: principles and applications, 2nd edn. Comstock Pub. Associates/Cornell University Press, Ithaca (2009)
2. Baum, D.A., Smith, S.D.: Tree thinking: an introduction to phylogenetic biology. Roberts, Greenwood Village (2012)
3. Forster, P., Renfrew, C.: Phylogenetic methods and the prehistory of languages. McDonald Institute monographs. McDonald Institute for Archaeological Research, University of Cambridge, Cambridge (2006)
4. Huerta-Cepas, J., Capella-Gutierrez, S., Pryszcz, L.P., Denisov, I., Kormes, D., Marcet-Houben, M., Gabaldón, T.: Phylomedb v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. Nucleic Acids Res. 39(Database issue), D556–D560 (2011)
5. Planet, P.J.: Tree disagreement: measuring and testing incongruence in phylogenies. J. Biomed. Inform. 39(1), 86–102 (2006)
6. Felsenstein, J.: Evolutionary trees from dna sequences: a maximum likelihood approach. J. Mol. Evol. 17(6), 368–376 (1981)

13.8 Exercises

1. Phylogenetics is...

   A) The grouping of organisms by their physical characteristics
   B) The study of evolutionary relationships in organisms
   C) The study of gene expression in organisms
   D) The extraction of phylo from genetic sequences

2. What is the difference between a rooted and unrooted tree?

   A) Rooted trees are unable to reveal information about evolutionary relationships
   B) Rooted trees are considered more reliable for establishing evolutionary relationships
   C) Rooted trees attempt to establish the relatedness of organisms to a common ancestor
   D) Rooted trees are significantly easier to construct than unrooted trees

3. What is the term for an individual species in a phylogenetic tree?

   A) Leaf
   B) Operational Tanonomic Unit (OTU)
   C) External Node
   D) All of the Above

4. At the Tree of Life website, how are organisms connected?

   A) A single species will link directly to all related species
   B) Organisms are linked by the root page, will contains direct links to all species and groupings
   C) Each branch connects to related groupings via other branch pages, terminating in individual species
   D) The tree is divided into several root pages that link to a specific type of living creature, which possess links to sub-groupings that terminate in individual species

5. What is the difference between using an optimality criterion and a clustering algorithm for reconstructing phylogenetic trees?

   A) Optimality criterions function by determining the number of steps necessary to transform one tree into another, while clustering algorithms function by creating hereditary groupings
   B) In the steps necessary to reconstruct a phylogenetic tree, the optimality criterion must be applied first in order to provide valid input data for the clustering algorithm
   C) In the steps necessary to reconstructing a phylogenetic tree, the clustering algorithm must be applied first in order to provide valid input data for the optimality criterion method

   D) Clustering algorithms function by determining the number of steps necessary to transform one tree into another, while optimality criterions function by determining the number of steps necessary to transform one tree into another

6. How does a branch and bound optimization render exhaustive searches more efficient?

   A) It generates all possible trees
   B) It discovers and prunes unproductive paths
   C) It guarantees the discovery of an optimal tree
   D) It makes use of breadth-first search

7. What is the correct order of the steps the ClustalW algorithm uses for Multiple Sequence Alignment:

   1. ClustalW constructs a distance matrix of N(N-1)/2 pairs of sequences by pairwise alignment of the sequences
   2. ClustalW builds a guide tree from the distance matrix using the clustering method (neighbor-joining) by Saitou and Nei
   3. ClustalW will convert the similarity scores to evolutionary distances based on the model by Kimura

   Correct order of processing steps:

   A) 1,2,3
   B) 2,3,1
   C) 1,3,2
   D) 2,1,3

8. In topological distance, the more related the trees the larger the value resulting from the partition metric will be. True or False. Justify.
9. Consider the two unrooted phylogenetic trees shown in Fig. 2b. Recall that the topological distance between two trees, dT is defined as $dT = 2 \times [min(q_1, q_2) - p] + |q_1 - q_2|$ where q_1 and q_2 is the number of partitions in each of the two trees and p is the number of common partitions. Find the topological distance between the following two unrooted trees. Note that edges a, b and c are internal edges.
10. Consider the following set of sequences. Perform the following analysis using the functions provided in MATLAB Bioinformatics toolbox:

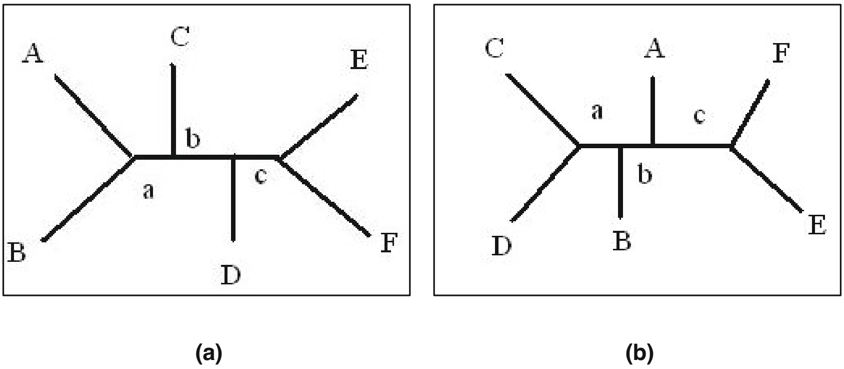


Fig. 13.11 Comparison of Phylogenetic Trees

```
>sequence_1
ACGTACGTACGT
>sequence_2
ACGTACGTACGTACGT
>sequence_3
ACGTACGTACGTTGCA
>sequence_4
ATCGATCGATCG
>sequence_5
ATCGATCGATCGATCG
```

(a) Compute the pairwise distance between each sequence using MATLAB's global alignment function *nwalign*.

(b) Select the pair of sequences with the smallest distance and construct the alignment. Construct a sequence structure with closest neighbors and compute the consensus sequence *cons*. The following example assumes that *seqA* and *seqB* are to be merged into an internal node. Note down the score.

```
 [score aln] = nwalign (seqA, seqB, 'Alphabet', 'NT');
seqStruct(1).Sequence = aln(1,:);
seqStruct(2).Sequence = aln(3,:);
cons = seqconsensus (seqStruct);
```

(c) Redo the distance computations with the merged sequences replaced with their consensus sequence. Continue till all sequences have been merged into a single node noting down the score at each step.

(d) Use the order in which the sequences were merged and the distances of the nodes merged to construct a phylogenetic tree using MATLAB function *phytree*.

(e) Display and print the tree.

(f) Save the tree to a disk file, read the saved tree and display it.