Lecture Notes: The Mathematics of Phylogenetics

Elizabeth S. Allman,
John A. Rhodes

IAS/Park City Mathematics Institute
June-July, 2005

University of Alaska Fairbanks
Spring 2009, 2012, 2016, 2018

©2005, Elizabeth S. Allman and John A. Rhodes

Contents

Introduction

The basic problem addressed in these notes is the following: Suppose sequences such as DNA are taken from a number of different organisms. How can we use them to understand evolutionary relationships? Schematically, an instance of the problem and a possible solution might be depicted as in Figure 1:

a: `CATTAGTAGA...`
b: `CATTAGTGGA...`
c: `CATTAGTAGA...`
d: `CCTTAGTGAA...`
e: `CCTTAGTAAA...`

$\rightsquigarrow$

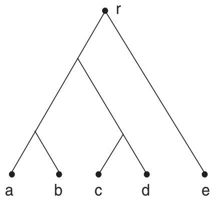

Figure 1: The basic problem of phylogenetics is to use sequences drawn from several different organisms to infer a tree depicting their evolutionary history.

Our goals are to develop some of the many things that the symbol "$\rightsquigarrow$" in this diagram might represent. Although the problem itself is a biological one, our focus will be primarily on how ideas and approaches from the mathematical sciences (mathematics, statistics, computer science) are used to attack it.

The audience we have in mind includes both biologists looking for a deeper understanding of the ideas that underlie the evolutionary analyses they may routinely perform with software, and mathematical scientists interested in an introduction to a biological application in which mathematical ideas continue to play a vital role. It is impossible to write a text that these two different groups will both find exactly to their liking — typical biologists have at most a few undergraduate math and statistics courses in their background, while mathematical scientists may have a similar (or even weaker) background in biology. Nonetheless, we believe that both groups can gain from interactions over this material, reaching a middle ground where each side better understands the other.

You will not find any direct help here on how to use software to perform evolutionary analyses — software changes and improves rapidly enough that such a guide would become quickly outdated. You will also not find a thorough development of all the interesting mathematics that has arisen in phylogenetics (as much as the mathematician authors might enjoy writing that). Rather we hope to focus on central ideas, and better prepare readers to understand the primary literature on the theory of phylogenetics.

Before beginning, there are a few other works that should be mentioned, either for further or parallel study.

- Felsenstein's book [Fel04] is an encyclopedic overview of the field by one of its primary developers, informally presented, and offers many insightful perspectives and historical comments, as well as extensive references to original research publications.

- Yang [Yan06] gives a more formal presentation of statistical phylogenetics, also with extensive references. This book is probably too technical to serve as a first introduction to the field.

- Semple and Steel's text [SS03] provides a careful mathematical development of the combinatorial underpinnings of the field. It has become a standard reference for theoretical results of that sort.

- The collection of papers in [KK10] includes articles addressing species tree estimation from gene trees in both a theoretical and practical setting.

- Steel's book [Ste16] is an excellent up-to-date book focusing on the mathematical underpinnings of phylogenetics. This monograph follows from an NSF-CBMS Regional Research Conference in Phylogenetics in which Mike Steel, a leader in mathematical phylogenetics, gave a series of ten lectures in 2014.

While all of these books are excellent references, the first two lack exercises (which we feel are essential to developing understanding) and they often give an overview rather than attempting to present mathematical developments in detail. The third is an excellent textbook and reference, but its focus is rather different from these notes, and may be hard going for those approaching the area from a biological background alone. In fact, [Yan06] and [SS03] are almost complementary in their emphasis, with one discussing primarily models of sequence evolution and statistical inference, and the other giving a formal mathematical study of trees.

Here we attempt to take a middle road, which we hope will make access to both the primary literature and these more comprehensive books easier. As a consequence, we do not attempt to go as deeply into any topic as a reference for experts might.

Finally, we have borrowed some material (mostly exercises) from our own undergraduate mathematical modeling textbook [AR04]. This is being gradually removed or reworked, as editing continues on these notes.

As these notes continue to be refined with each use, your help in improving them is needed. Please let us know of any typographical errors, or more substantive mistakes that you find, as well as passages that you find confusing. The students who follow you will appreciate it.

Elizabeth Allman
e.allman@alaska.edu

John Rhodes
j.rhodes@alaska.edu

Chapter 1

Sequences and Molecular Evolution

As the DNA of organisms is copied, potentially to be passed on to descendants, mutations sometimes occur. Individuals in the next generation may thus carry slightly different sequences than those of their parent (or either parent, in the case of sexual reproduction). These changes, which can be viewed as essentially random, continually introduce the new genetic variation that is essential to evolution through natural selection.

Depending on the nature of the mutations in the DNA, offspring may be more, less, or equally viable than the parents. Some mutations are likely to be lethal, and will therefore not be passed on further. Others may offer great advantages to their bearers, and spread rapidly in subsequent generations. But many mutations will offer only a slight advantage or disadvantage, and the majority are believed to be selectively neutral, with no effect on fitness. These neutral mutations may still persist over generations, with the particular variants that are passed on being chosen by luck. As small mutations continue to accumulate over many generations, an ancestral sequence can be gradually transformed into a different one, though with many recognizable similarities to its predecessors.

If several species arise from a common ancestor, this process means we should expect them to have similar, but often not identical, DNA forming a particular gene. The similarities hint at the common ancestor, while the differences point to the evolutionary divergence of the descendants. While it is impossible to observe the true evolutionary history of a collection of organisms, their genomes contain traces of the history.

But how can we reconstruct evolutionary relationships between several modern species from their DNA? It's natural to expect that species that have more similar genetic sequences are probably more closely related, and that evolutionary relationships might be represented by drawing a tree. However, this doesn't give much of a guide as to how we get a tree from the sequences. While occasionally simply 'staring' at the patterns will make relationships clear, if there are a large number of long sequences to compare it's hard to draw any conclusions. And 'staring' isn't exactly an objective process, and you may not even be aware of how you are reaching a conclusion. Even though reasonable people may agree, it's hard to view a finding produced in such a way as scientific.

Before we develop more elaborate mathematical ideas for how the problem of phylogenetic inference can be attacked, we need to briefly recount some biological background. While most readers will already be familiar with this, it is useful to fix terminology. It is also remarkable how little needs to be said: Since the idea of a DNA sequence is practically an abstract mathematical one already, very little biological background will be needed.

1.1 DNA structure

The DNA molecule has the form of a double helix, a twisted ladder-like structure. At each of the points where the ladder's rungs meet its upright poles one of four possible molecular subunits appears. These subunits, called *nucleotides* or *bases*, are adenine, guanine, cytosine, and thymine, and are denoted by the letters `A`, `G`, `C`, and `T`. Because of chemical similarity, adenine and guanine are called *purines*, while cytosine and thymine are called *pyrimidines*.

Each base has a specific complementary base with which it can form the rung of the ladder through a hydrogen bond. We always find either `A` paired with `T`, or `G` paired with `C`. Thus the bases arranged on one side of the ladder structure determine those on the other. For example if along one pole of the ladder we have a sequence of bases

$$\texttt{AGCGCGTATTAG},$$

then the other would have the complementary sequence

$$\texttt{TCGCGCATAATC}.$$

Thus all the information is retained if we only record one of these sequences. Moreover, the DNA molecule has a directional sense (distinguished by what are called its 5' and 3' ends) so that we can make a distinction between a sequence like `ATCGAT` and the inverted sequence `TAGCTA`. The upshot of all this structure is that we will be able to think of DNA sequences mathematically simply as finite sequences composed from the 4-letter alphabet $\{\texttt{A}, \texttt{G}, \texttt{C}, \texttt{T}\}$.

Some sections of a DNA molecule form *genes* that encode instructions for the manufacturing of proteins (though the production of the protein is accomplished through the intermediate production of messenger RNA). In these genes, triplets of consecutive bases form *codons*, with each codon specifying a particular amino acid to be placed in the protein chain according to the *genetic code*. For example the codon `TGC` always means that the amino acid cysteine will occur at that location in the protein. Certain codons also signal the end of the protein sequence. Since there are $4^3 = 64$ different codons, and only 20 amino acids and a 'stop' command, there is some redundancy in the genetic code. For instance, in many codons the third base has no affect on the particular amino acid the codon specifies.

Proteins themselves have a sequential structure, as the amino acids composing them are linked in a chain in the order specified by the gene. Thus a protein can be specified by a sequence of letters chosen from an alphabet with 20 letters, corresponding to the various amino acids. Although we tend to focus on DNA in these notes, a point to observe is there are several types of biological sequences, composed of bases, amino acids, or codons, that differ only in the number of characters in the alphabet — 4, 20, or 64 (or 61 if the stop codons are removed). One can also recode a DNA sequence to specify only purines (R) or pyrimidines (Y), using an alphabet with 2 characters. Any of these biological sequences should contain traces of evolutionary history, since they all reflect, in differing ways, the underlying mutations in DNA.

A stretch of DNA giving a gene may actually be composed of several alternating subsections of *exons* and *introns.* (Exons are the part of the gene that will ultimately be *expressed* while introns are *intruding* segments that are not.) After RNA is produced from both exons and introns, a splicing process removes those sections arising from introns, so the final RNA reflects only what comes from the exons. Protein sequences thus reflect only the exon parts of the gene sequence. This detail may be important, as mutation processes on introns and exons may be different, since only the exon is constrained by the need to produce functional gene products.

Though it was originally thought that genes always encoded for proteins, we now know that some genes encode the production of other types of RNA which are the 'final products' of the gene, with no protein being produced. Finally, not all DNA is organized into the coding sections referred to as genes. In humans DNA, for example, about 97% is believed to be non-coding. In the much smaller genome of a virus, however, which is packed into a small delivery package, only a small proportion of the genome is likely to be non-coding. Some of this is non-coding DNA is likely to be meaningless raw material which plays no current role, and is sometimes called *junk* DNA (though it may become meaningful in future generations through further mutation). Other parts of the DNA molecules serve regulatory purposes.The overall picture is quite complicated and still not fully understood.

1.2 Mutations

Before DNA, and the hereditary information it carries, is passed from parent to offspring, a copy must be made. For this to happen, the hydrogen bonds forming the rungs of the ladder in the molecule are broken, leaving two single strands. Then new double strands are formed on these, by assembling the appropriate complementary strands. The biochemical processes are elaborate, with various safeguards to ensure that few mistakes are made in the final product. Nonetheless, changes of an apparently random nature sometimes occur.

The most common mutation that is introduced in the copying of sequences of DNA is a *base substitution.* This is simply the replacement of one base for another at a certain site in the sequence. For instance, if the sequence `AATCGC` in an ancestor becomes `AATGGC` in a descendant, then a base substitution $\texttt{C}\to\texttt{G}$ has occurred at the fourth site. A base substitution that replaces a purine with a purine, or a pyrimidine for a pyrimidine, is called a *transition*, while an interchange of these classes is called a *transversion.* Transitions are often observed to occur more frequently than transversions, perhaps because the chemical structure of the molecule changes less under a transition than a transversion.

Other DNA mutations that can be observed include the deletion of a base or consecutive bases, the insertion of a base or consecutive bases, and the inversion (reversal) of a section of the sequence. While not uncommon, these types of mutations tend to be seen less often than base substitutions. Since they usually have a dramatic effect on the protein for which a gene encodes, this is not too surprising. We'll ignore such possibilities, in order to make our modeling task both clearer and mathematically tractable. (There is much interest in dropping this restriction in phylogenetics, though it has proved very difficult to do so. Current methods that are based on modeling insertion and deletion processes can handle data sets with only a handful of organisms.)

Our view of molecular evolution, then, can be depicted by an example such as the following, in which we have an ancestral sequence S_0, its descendant S_1, and a descendant S_2 of S_1.

$$\begin{aligned} S_0 &: \texttt{ATGTCGCCTGATAATGCC} \\ S_1 &: \texttt{ATGCCGCTTGACAATGCC} \\ S_2 &: \texttt{ATGCCGCGTGATAATGCC} \end{aligned}$$

Notice site 4 underwent a net transition from S_0 to S_1, and then no further net mutation as S_2 evolved. If we were only able to observe S_0 and S_2, we would still see evidence of one net transition in this site. However, while site 8 experienced at least two mutations, if we only saw the sequences S_0 and S_2 we could only be sure that one occurred. Site 12 shows a similar situation, where at least two mutations occurred, yet comparing S_0 and S_2 alone would show no evidence of either.

This illustrates that there may be *hidden mutations*, such as $\texttt{C} \to \texttt{T} \to \texttt{G}$, in which subsequent substitutions obscure earlier ones from our observation when we do not have access to sequences from all generations. A *back mutation* such as $\texttt{T} \to \texttt{C} \to \texttt{T}$ is simply a more extreme case of this. The distinction between observed mutations and the actual mutations including hidden ones will be an important one when considering data.

In reality, we seldom have an ancestral DNA sequence, much less several from different times along a line of descent. Instead, we have sequences from several currently living descendants, but no direct information about any of their ancestors. When we compare two sequences, and imagine the mutation process that produced them, the sequence of their *most recent common ancestor*, from which they both evolved, is unknown. This will produce additional complications as our analysis becomes more sophisticated.

1.3 Aligned Orthologous Sequences

Given a stretch in a DNA sequence from some organism, there are good search algorithms (such as BLAST) to find similar stretches in DNA sequences that have been found from other organisms. Thus if a gene has been identified for one organism, we can quickly locate likely candidate sequences for similar genes in related organisms. Perhaps after experimentally verifying that these are in fact genes and that they have a similar function, we can reasonably assume the sequences are *orthologous*, meaning they descended from a common ancestral sequence.

A complication can arise if an organism has undergone a *gene duplication* in which a gene that was formerly present in the genome appears with multiple copies in descendants. While the copies are still related, at least one of them is likely to be more free to vary, since the other may continue to fulfill its original role. These *paralogous* genes will retain similarities, though, and if they are passed on to several different organisms, it may be easy to mistakenly base an analysis on genes whose most recent common ancestor is the original duplicated one, rather than on those descended from a common copy after the duplication event. Even if one infers the correct evolutionary relationship between such sequences, it may be misleading as to the relationship between the organisms.

Either by trial and error, or using algorithms we will not discuss in these notes, we can *align* the sequences from the different organisms so that many of the bases match across most of the sequences. For some data sets, alignment is a trivial problem, since so much of the sequences match exactly. For other data sets, finding good alignments may be quite difficult. The essential problem is that deletions and insertions may have occurred, and we must decide where to place gaps in the various sequences to reflect this. While many software tools have been developed to do this, the computational burden of finding any sort of a 'best' alignment is so great that heuristic methods are essential. Faced with a large number of sequences, with much variation between them, even the best of current software may not produce an alignment that on human inspection appears very good. In the end, a mix of algorithms (most of which depend on first inferring at least an approximate tree) and *ad hoc* human adjustment is sometimes used to get better results. For those interested in learning more about sequence alignment, [Wat95] is a beginning source.

As the starting point for the techniques we discuss here, we take a collection of *aligned orthologous DNA sequences*. Our goal will be to produce a *phylogenetic tree* that describes their likely descent from a common ancestral sequence. Indeed, Figure 1 of the Introduction shows an example. Note that we will not try to decide if the sequences did evolve along a tree from a common ancestor — that is an assumption we make. Our question is simply to determine the best candidate tree to describe the unknown history.

Chapter 2

Combinatorics of Trees I

2.1 Graphs and Trees

Before we discuss any of the methods of inference of phylogenetic trees, we introduce some background and terminology from graph theory. This simply formalizes the basic notions of the kinds of trees that biologists use to describe evolution, and gives us useful language.

While at first it might seem sufficient to simply draw a few examples of trees rather than define them formally, there are good reasons for being more precise. For instance, when we draw trees depicting the specific evolutionary relationships among a set of taxa, we may do this in many different ways. But while the drawings are different, they are meant to depict the same thing. Thus the idea of a tree is really an abstract one, and by giving it an abstract definition we can legitimately see that two trees are 'the same' even though the diagram we draw may have different forms. If we also look ahead to trying to write software to perform phylogenetic analyses, a computer must contain an internal representation of a tree, even though it has no visual abilities.

Definition. A *graph* G is pair $G = (V, E)$ where V is a set of *vertices* or *nodes*, and E is a set of *edges*. Each edge $e \in E$ is a two-element set $e = \{v_1, v_2\}$ of vertices $v_1, v_2 \in V$.

When $e = \{v_1, v_2\} \in E$, we say v_1 and v_2 are the *ends* of e, that e *joins* v_1 and v_2, that v_1 and v_2 are *incident* to e, and that v_1 and v_2 are *adjacent*. The *degree* or *valence* of a vertex is the number of edges to which it is incident.

Graphs are typically indicated by drawings such as that in Figure 2.1.

Recall that a set, by definition, cannot have a repeated element. This means that by requiring an edge have two vertices $v_1 \neq v_2$ we rule out the possibility of an edge looping from a vertex to itself. We also have ruled out the possibility of two or more edges having the same ends, since E also forms a set. Sometimes a graph is defined in such a way as to allow both of these possibilities, and our concept of graph is technically that of a *simple graph*. However, since we will

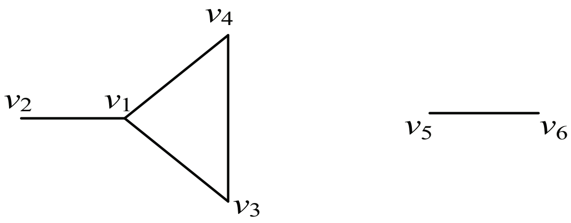

Figure 2.1: A depiction of a graph $G = (V, E)$ with $V = \{v_1, v_2, v_3, v_4, v_5, v_6\}$ and $E = \{\{v_1, v_2\}, \{v_1, v_3\}, \{v_1, v_4\}, \{v_3, v_4\}, \{v_5, v_6\}\}$. The vertices v_1, v_2, and v_3 have respective degrees 3, 1, and 2.

make no use of loops or multiple edges between the same vertices, we choose this more restrictive definition. All of our graphs will also be *finite*, meaning V, and hence E, is a finite set.

Graphs are widely used in mathematical biology to denote relationships between organism. For instance in ecology, a graph may summarize the direct interactions between species in an ecosystem, with nodes representing various species and edges between them their interactions (e.g., predation, or symbiosis). Gene interactions in a single organism can similarly be depicted by a graph, with nodes denoting genes and edges indicating when the product of one gene influences another. Graphs are sometimes referred to as *networks* in these contexts.

In phylogenetics, we often think of each vertex as representing a species, with the edges indicating lines of direct evolutionary relationships (ancestor-descendant pairs of species, along a lineage in which no other species under consideration splits off). Rather than species, however, we may be working with smaller groups of organisms such as subspecies or even individuals, or larger groups such as genera, etc. Thus we adopt the more neutral word *taxon* for whatever group is under consideration. (In scientific literature, these are sometimes referred to as *operational taxonomic units* or *OTUs*.)

If we are to use graphs to model evolution, with edges indicating lines of descent, then we of course need to rule out any taxon being an ancestor of itself. To be precise about this, we need a series of definitions.

Definition. In a graph, a *path of length n* from vertex v_0 to vertex v_n is a sequence of distinct vertices $v_0, v_1, v_2, \dots, v_n$ such that each v_i is adjacent to v_{i+1}.

A graph is *connected* if there is a path between any two distinct vertices.

In Figure 2.1, there are two paths from v_2 to v_3, namely v_2, v_1, v_3 and v_2, v_1, v_4, v_3. However, v_2, v_1, v_2, v_1, v_3 is not a path since it repeats v_2 and v_1. Since there is no path from v_2 to v_5, this graph is not connected.

Definition. A *cycle* is a sequence of vertices $v_0, v_1, \dots, v_n = v_0$ which are distinct (other than $v_n = v_0$) with $n \geq 3$ and v_i adjacent to v_{i+1}

In Figure 2.1, v_1, v_3, v_4, v_1 is a cycle. If we removed the edge $e = \{v_3, v_4\}$ however, the graph would not have a cycle. Note that the path v_1, v_3, v_1 is not a cycle, since the sequence of vertices is too short; this length condition rules out backtracking along a single edge being considered a cycle.

Finally, we can define

Definition. A *tree* $T = (V, E)$ is a connected graph with no cycles.

The graph in Figure 2.1 is not a tree, for two reasons; it is not connected, and it has a cycle. As an example of a tree, we have that depicted in Figure 2.2.

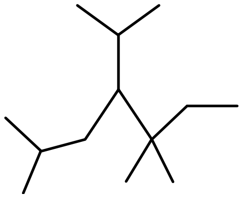

Figure 2.2: A tree.

If a vertex lies in two or more distinct edges of a tree, we say it is an *interior vertex.* If it lies in only one edge, then we call it a *terminal vertex* or a *leaf.* Similarly, an edge joining two interior vertices is an *interior edge*, while one incident to a leaf is called a *pendant edge.*

Though trees will be our primary focus, they are not the only types of graphs that are important for describing evolutionary relationships. For instance, if hybridization of two related ancestral species occurs, a graph with a cycle is needed to depict this, as the two will have both a common ancestor and a common descendant. Other types of lateral gene transfer similarly result in non-tree graphs. Much work has been done in recent years to use networks rather than trees to capture such biological phenomena, as well as for visualizing conflicting phylogenetic signals in data sets. Huson, *et al.* [HRS10] provide an excellent entry into these developments.

A basic property of trees that should be intuitively clear is the following:

Theorem 1. If v_0 and v_1 are any two vertices in a tree T, then there is a unique path from v_0 to v_1.

Sketch of proof. We leave a formalization of the proof as an exercise, but summarize the argument: Given two different paths with the same endpoints, consider a route following first one and then the other backwards, so we return to our starting vertex. Of all such pairs of paths, choose one where this combined route is the shortest.

Now the combined route cannot be a cycle, since a tree has none. So the two paths must have a vertex other than the endpoints in common. But then we can use this to give a pair of paths producing a shorter combined route, which would be a contradiction. □

The length of the unique path between two vertices in a tree is called the *graph-theoretic distance* between the vertices. For example, in Figure 2.3, for the tree on the left the graph-theoretic distance from a to d is 3, while for the other two graphs it is 4.

The trees we have defined are sometimes called *unrooted trees* or *undirected trees*. To those who are familiar with the notion of a directed graph, it may seem odd that we emphasize trees with undirected edges for depictions of evolutionary histories. Indeed, time provides a direction that is of course central to evolutionary processes. However, we will see that many of the mathematical models and inference methods are more naturally associated with undirected trees, and so we make them our basic objects.

With that said, however, sometimes we pick a particular vertex ρ in a tree and distinguish it as the *root* of the tree. More often, we introduce a new vertex to be the root, by subdividing an edge, $\{u, v\}$, into two $\{u, \rho\}$ and $\{\rho, v\}$. Figure 2.3 shows two different rooted versions of the same unrooted tree,obtained by subdividing different edges. We use T to denote an unrooted tree, and T^ρ to denote a tree with a choice of a root ρ.

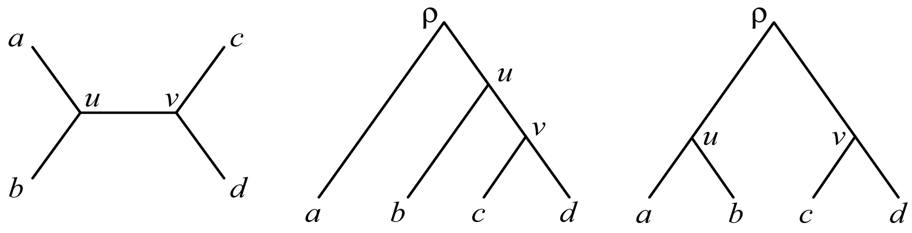

Figure 2.3: An unrooted tree and two rooted counterparts.

Biologically, a root represents the *most recent common ancestor* (MRCA) of all the taxa in the tree. Sometimes, however, roots are chosen for mathematical convenience rather than biological meaning, so one must be careful with such an interpretation.

Viewing ρ as a common ancestor of all other taxa represented in a tree, it is natural to give each edge of the tree an orientation away from the root. More precisely, a *directed edge* is not a set of two vertices, but rather an ordered pair $e = (v_1, v_2)$, where the order is specified by requiring that the path from ρ to v_1 is shorter than that from ρ to v_2. For instance, in passing from the leftmost tree to the middle tree in Figure 2.3, the undirected edge $\{u, v\}$ becomes the directed edge (u, v), since u is closer to ρ than is v.

For a directed edge $e = (v_1, v_2)$, we say v_1 is the *initial vertex* or *tail* of e and that v_2 is the *final vertex* or *head* of e. We also refer to v_1 as the *parent* of v_2, and v_2 as the *child* of v_1. More generally, for a rooted tree we may use the words *ancestor* and *descendant* in the obvious way to refer to nodes. For instance, in the middle tree of Figure 2.3, u is ancestral to d, but not to a.

An unrooted tree is said to be *binary* if each interior vertex has degree three. This terminology of course fits with the biological view of one lineage splitting into two, but this leads to the rather odd situation that a synonym for binary is *trivalent.* We call a rooted tree T^ρ binary if all interior vertices other than ρ are of degree three, while ρ is of degree two.

Although it's conceivable biologically that a tree other than a binary one might describe evolutionary history, it is common to ignore that possibility as unlikely. When non-binary trees arise in phylogenetic applications, they usually have vertices of degree greater than 3 (also called *multifurcations*) to indicate ambiguity arising from ignorance of the true binary tree. This is sometimes referred to as a *soft polytomy.* In contrast, a *hard polytomy* refers to a multifurcation representing positive knowledge of many lineages arising at once, such as a radiation of species. Since even radiations are likely to be modeled well by a succession of bifurcations with short times between them, in most applications, biologists generally prefer to find 'highly resolved' trees, with few multifurcations; a binary tree is the goal.

The trees used in phylogenetics have a final distinguishing feature — the leaves represent known taxa, which are typically currently extant and are the source of the data used to infer the tree. The internal vertices, in contrast, usually represent taxa that are no longer present, and from which we have no direct data. (Even if we have data from ancient organisms, we cannot assume they are direct ancestors of any extant ones; they are more likely to be on offshoots from the lineages leading to our modern samples.) This is formalized by labeling the leaves of the tree with the names of the known taxa, while leaving unlabeled the interior vertices. Thus for either rooted or unrooted trees we are interested primarily in the following objects.

Definition. Let X denote a finite set of taxa, or labels. Then a *phylogenetic X-tree* is a tree $T = (V, E)$ together with a one-to-one correspondence $\phi : X \to L$, where $L \subseteq V$ denotes the set of leaves of the tree. We call ϕ the *labeling map.* Such a tree is also called a *leaf-labeled tree.*

More informally, the labeling map simply assigns each of the taxa to a different leaf of the tree, so that every leaf receives a label.

Often we will blur the distinction between the leaves of a phylogenetic tree and the labels that are assigned to them. For instance, we will use T or T^ρ to denote a phylogenetic X-tree, often without explicitly mentioning either the set X or the labeling function. However, it is important to note that this labeling is crucial. Two different maps from X to the leaves of T usually produce two different phylogenetic trees. In other words, phylogenetic trees can be distinguished from one another either due to different 'shapes' of the underlying trees, or merely by a different labeling of the leaves.

2.2 Counting Binary Trees

Suppose we are interested in relating a collection X of n taxa by a phylogenetic tree. How many different trees might describe the relationship? It's helpful to know this since if we want to relate a group of taxa, the more possible trees there are, the harder the problem may be. Could we, for instance, simply list all the trees that might relate 10 taxa, and consider each one in turn? Or would this list be so long that such an approach is infeasible?

To refine the question, we first need to say what it means for two phylogenetic trees to be the same. Informally, we don't care about the names given to vertices, as long as the shape of the tree and how the labels are placed on leaves agree.

Definition. Two phylogenetic X-trees are *isomorphic* if there is a one-to-one correspondence between their vertices that respects adjacency, and their leaf labelings.

Let $b(n)$ denote the number of distinct (up to isomorphism) unrooted binary phylogenetic X-trees, where $X = \{1, 2, \dots, n\}$. One quickly sees $b(2) = 1$, $b(3) = 1$, and $b(4) = 3$, as the diagrams in Figure 2.4 indicate.

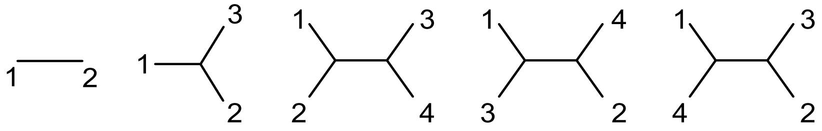

Figure 2.4: All unrooted binary phylogenetic trees for $X = \{1, 2, \dots, n\}$, $n = 2, 3, 4$.

The key observation for counting larger trees now appears clearly when we consider $b(5)$: We can begin with any one of the 3 trees relating the first four taxa, and 'graft' another edge leading to a fifth taxon to any of the 5 edges in that four-taxon tree, so that $b(5) = 3 \cdot 5 = 15$. For $b(6)$, we can begin with any of these 15 trees and graft on another edge leading to a sixth taxon to any of the edges in that tree. To obtain a general formula, then, we also need to obtain a formula for the number of edges in these trees.

Theorem 2. An unrooted binary tree with $n \geq 2$ leaves has $2n - 2$ vertices and $2n - 3$ edges.

Proof. The statement is certainly true for $n = 2$, which is the base case for induction.

Suppose T has $n \geq 3$ leaves, and assume the statement is true for a tree with $n-1$ leaves. If v_1 is one of the leaves of T, then v_1 lies on a unique edge $\{v_1, v_2\}$, while v_2 lies additionally on two other edges $\{v_2, v_3\}$ and $\{v_2, v_4\}$. Removing these three edges and the vertices v_1, v_2 from T, and introducing a new edge $\{v_3, v_4\}$ gives a binary tree T' with $n - 1$ leaves. Since both the number of vertices and the number of edges have been decreased by 2, for T the number of vertices must have been $(2(n-1)-2)+2 = 2n-2$, and the number of edges $(2(n-1)-3)+2 = 2n-3$. □

Theorem 3. If X has $n \geq 3$ elements, there are

$$b(n) = (2n-5)!! = 1 \cdot 3 \cdot 5 \cdots (2n-5)$$

distinct unrooted binary phylogenetic X-trees.

Proof. Again we use induction, with the base case of $n = 3$ clear.

Suppose then that T has n leaves, and let T' be the $(n-1)$-leaf tree constructed from T as in Theorem 2 by 'removing' the leaf v_1 and adjusting edges appropriately. Then with v_1 fixed, the map $T \mapsto (T', \{v_3, v_4\})$ is a bijection from n-leaf trees to pairs of $(n-1)$-leaf trees and edges. In this pair, we think of the edge $\{v_3, v_4\}$ in T' as the one onto which we 'graft' a new edge to v_1 to recover T. Counting these pairs shows

$$b(n) = b(n-1) \cdot (2(n-1)-3) = b(n-1) \cdot (2n-5).$$

But since we inductively assume $b(n-1) = (2(n-1)-5)!! = (2n-7)!!$, we obtain the desired formula. □

This last theorem also yields a count for rooted binary trees, by simply noting that adjoining to any rooted binary tree T^ρ a new edge $\{\rho, v_{n+1}\}$, where v_{n+1} is a new leaf, gives a one-to-one correspondence between rooted binary trees with n leaves and unrooted binary trees with $n+1$ leaves.

Corollary 4. The number of rooted binary phylogenetic X-trees relating n taxa is

$$b(n+1) = (2n-3)!! = 1 \cdot 3 \cdot 5 \cdots (2n-3).$$

The formula for $b(n)$ shows it is a very large number even for relatively small values of n. (See exercises.) The implication of this for phylogenetic inference will quickly become clear: Any inference method that relies on considering every possible binary tree will be impossibly slow if the number of taxa is even moderately large.

2.3 Metric Trees

Sometimes it is useful to specify lengths of edges in trees with non-negative numbers, as in Figure 2.5 below. Note that the lengths may either be specified by explicitly writing them next to the edges, or by simply drawing edges with the appropriate lengths, and providing a length scale.

Biologically, these lengths are usually interpreted as some measure of how much change occurred in sequences between the ends of the edge, with larger numbers denoting more change. Occasionally they represent elapsed time. However, it is generally incorrect to assume edge lengths are elapsed times, no matter how intuitively appealing that interpretation is.

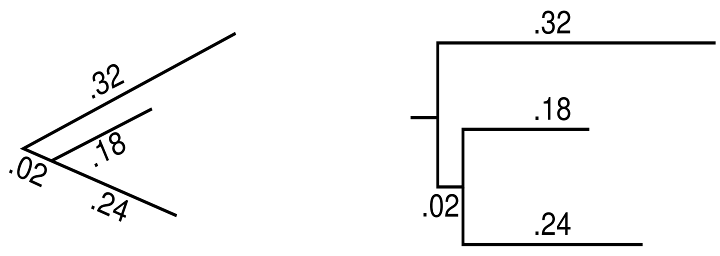

Figure 2.5: Two depictions of the same metric tree, drawn in different styles. In the second, all vertical line segments are considered to have no length, and the short horizontal segment merely indicates the root.

Definition. A *metric tree* (T, w) is a rooted or unrooted tree $T = (V, E)$ together with a function $w : E \to \mathbb{R}^{\geq 0}$ assigning non-negative numbers to edges. We call $w(e)$ the *length* or *weight* of the edge e.

When we wish to emphasize that edge lengths are *not* specified for a tree T, we will sometimes refer to T as a *topological tree.* We can obtain a topological tree from a metric one by simply ignoring the edge lengths. This means several metric trees may have the same underlying topological tree, but differ because edge lengths are not the same.

Notice the terminology conflict between the graph-theoretic notion of length of a path (the number of edges in the path) in a tree, and this new use of the word length. To avoid confusion, graph theorists tend to prefer the term 'weight,' but 'length' is more common in phylogenetics.

A metric tree leads to a way of measuring distances between its vertices. For any $v_1, v_2 \in V(T)$, define

$$d(v_1, v_2) = \sum_{\substack{e \text{ on the path} \\ \text{from } v_1 \text{ to } v_2}} w(e),$$

i.e., the sum of the edge lengths along the path between the two vertices.

Proposition 5. For a metric tree (T, w), the function $d : V \times V \to \mathbb{R}^{\geq 0}$ satisfies

(i) $d(v_1, v_2) \geq 0$ for any v_1, v_2 (non-negativity),

(ii) $d(v_1, v_2) = d(v_2, v_1)$ for any v_1, v_2 (symmetry),

(iii) $d(v_1, v_3) \leq d(v_1, v_2) + d(v_2, v_3)$ for any v_1, v_2, v_3 (triangle inequality).

If all edges of T have positive length, then, in addition,

$$d(v_1, v_2) = 0 \text{ if, and only if, } v_1 = v_2.$$

Proof. See the exercises. □

If all edges of T have positive length, then the proposition above shows d is a *metric* on $V(T)$ in the usual sense of the word in topology or analysis. In this case, we call d a *tree metric.*

If T has some edges of length zero, we can of course remove those edges, identifying the vertices at their two ends, to get a metric tree with all positive edge lengths that carries essentially the same information as our original metric tree. (However, if an edge leading to a leaf has length zero, then the resulting tree may no longer be a phylogenetic X-tree, as a label may now be on an internal vertex.) On this collapsed tree we then have that d is a tree metric. As it is often convenient to work with metric trees that have positive edge lengths, keeping this process in mind we will lose little by making such an assumption at times.

2.4 Ultrametric Trees and Molecular Clocks

If we wish to interpret edge lengths on a rooted metric tree as times, then it is essential that all leaves be equidistant from the root. This is simply because this distance would represent the total elapsed time from the MRCA of the taxa to the present, when the taxa were sampled.

Definition. A rooted metric tree is said to be *ultrametric* if all its leaves are equidistant from its root using the tree metric: For any two leaves a, b and the root ρ, $d(\rho, a) = d(\rho, b)$.

Ultrametric trees are often called *molecular clock* trees in phylogenetics, since if mutation occurs at a constant rate over all time and lineages, *i.e.* is clock-like, then many methods of inference from sequences will produce such trees in idealized circumstances. Edge lengths should be interpreted as measures of how much mutation has occurred, so that with clock-like mutation we simply scale elapsed time by a constant mutation rate to get lengths. But one should be careful — even if a tree is ultrametric, it need not have been produced under a molecular clock. For instance mutation rates could increase in time, uniformly over all lineages, and each would show the same total mutation from root to leaf.

A molecular clock assumption can be biologically reasonable in some circumstances, for instance, if all taxa are reasonably closely related and one suspects little variation in evolutionary processes during the evolutionary period under study. Other times it is less plausible, if more distantly related taxa are in the tree, and the circumstances under which they evolved may have changed considerably throughout the tree.

One attractive feature of a molecular clock hypothesis and ultrametric trees is that even if the location of the root is not known, there is a simple way to find it, as shown by the following.

Theorem 6. Suppose T^ρ is a rooted ultrametric tree with positive edge lengths. Let v_1, v_2 be any two leaves with $d(v_1, v_2)$ maximal among distances between leaves. Then ρ is the unique vertex along the path from v_1 to v_2 with $d(\rho, v_1) = d(\rho, v_2) = d(v_1, v_2)/2$. Thus the placement of ρ can be determined from an unrooted metric version of T.

Proof. If T^ρ has more than one leaf, ρ is an internal vertex. Thus deleting ρ (and its incident edges) from T produces at least two connected components.

Suppose v_1 and v_2 are in different connected components of $T \smallsetminus \{\rho\}$. Then the path from v_1 to v_2 must pass through ρ, and we see by Exercise 12 that $d(v_1, v_2) = d(v_1, \rho) + d(\rho, v_2)$. Since all leaves are equidistant from the root, $d(v_1, v_2) = 2d(v_1, \rho)$. Thus ρ lies the specified distance from v_1 and v_2. Since edge lengths are positive, there can be only one such vertex on the path with this property.

If v_1 and v_2 are in the same connected component of $T \smallsetminus \{\rho\}$, then the path between them does not pass through ρ and so by Exercise 12 and the triangle inequality we have $d(v_1, v_2) < d(v_1, \rho) + d(\rho, v_2)$. But since all leaves are equidistant from the root, this shows $d(v_1, v_2) < 2d(v_1, \rho)$, and so the maximal distance between leaves cannot occur between v_1 and v_2. □

Sometimes when a root must be added to an unrooted metric tree, either as a rough approximation of the MRCA of the taxa or simply for convenience in drawing the tree, the root location is chosen as the midpoint of the path between the most distant pair of leaves. The above theorem justifies this as a valid way of locating the most recent common ancestor for a molecular clock tree, or a tree that appears to be 'close' to being one. For other trees, it is simply a heuristic method of locating a root, and cannot be reliably expected to give the true location of the MRCA.

2.5 Rooting Trees with Outgroups

In the absence of a molecular clock hypothesis, locating the root of a tree is generally not possible from mathematical considerations alone. Instead, a simple biological idea can often be used: We include an extra taxon in our study, beyond those we are primarily interested in relating. If this taxon is chosen so it is more distantly related to each of the taxa of primary interest than any of those are to each other, we call it an *outgroup*. For instance, if we are primarily interested in relating a number of species of duck, we might include some non-duck bird as an outgroup.

Provided we infer a good unrooted evolutionary tree for all our taxa including the outgroup, the vertex at which the outgroup is joined to the taxa of primary interest represents the MRCA of those, and hence serves as a root for the subtree relating only the taxa of primary interest. Thus we use prior biological knowledge of the relationship of the outgroup to the other taxa to solve our rooting problem. This is illustrated in Figure 2.6, where the taxon o designates the outgroup.

Though using an outgroup is the standard way of rooting phylogenetic trees when a molecular clock hypothesis is not reasonable, and the method usually

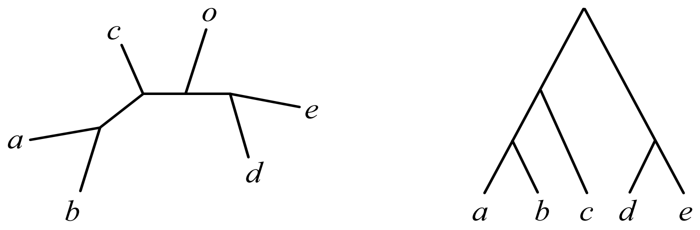

Figure 2.6: Using an outgroup o to infer a root

seems to work well, two features should be noted. First, this approach requires prior knowledge that the outgroup is distantly related to the other taxa. This may be obvious in some situations, such as that mentioned with ducks. But in others, particularly when one is attempting to infer very ancient relationships between very different taxa, it may not be clear what can serve as an outgroup. Second, even if it is clear how to pick an outgroup, there may be difficulties in inferring a tree that has it correctly placed. By definition, an outgroup is distantly related, and so whatever method is used to infer the tree may have more difficulties placing it correctly than the other taxa. If we lack knowledge of an outgroup, or doubt that we can place it on a tree correctly, then we may be forced to accept an unrooted tree as the best we can say about evolutionary relationships between the taxa of interest.

2.6 Newick Notation

There is a convenient and intuitive way of specifying a phylogenetic tree, which is simpler than the formal definition (as a collection of vertices and edges, with labelled leaves). It also has the advantage over a drawn figure of using standard typographic symbols. It is commonly used for input and/or output to computer programs, as well as in theoretical research papers.

Newick notation uses parenthetical grouping of taxon labels to specify the clustering pattern of a tree. This is best seen through examples. For instance the rooted tree on the right of Figure 2.6 is denoted $(((a,b),c),(d,e))$. However, it could also be designated $((d,e),(c,(b,a)))$, or in a number of other ways where we change the order of the items inside a pair of matched parentheses. This non-uniqueness of the Newick specification of a tree can make it hard to recognize by eye when when two large trees are the same.

If we instead wanted to designate an unrooted tree, we introduce a root arbitrarily, and use the rooted notation, simply specifying in words that we mean the tree should be unrooted. For instance the tree on the left of Figure 2.6 is the unrooted version of $(((a,b),c),((d,e),o))$, or equivalently of $(((((d,e),o),c),b),a)$, or a host of other possibilities.

In the case of a metric tree, we append each edge length to the taxon or group below it. For example the trees in Figure 2.5 with taxa a,b,c listed from top to bottom would be designated (a:0.32, (b:0.18, c:0.24):0.02). Since removing the root effectively joins two edges, adding their length, the unrooted version of this metric tree could also be designated as ((a:0.34, b:0.18):0, c:0.24), or in other ways.

2.7 Exercises

1. Give sets of vertices and edges defining the first tree in Figure 2.3. Do the same for the second tree, but since it is rooted, give a set of directed edges (*i.e.* order the vertices of each edge from parent to child.) Which edge of the first tree was subdivided to create the root in the second?

2. Draw two different figures representing the tree $T = (V, E)$ with $V = \{a, b, c, d, e, f, g, h\}$ and $E = \{\{a, g\}, \{b, c\}, \{c, f\}, \{c, g\}, \{d, g\}, \{e, f\}, \{f, h\}\}$. Do this so neither drawing can be obtained from the other by just stretching or shrinking edges or angles, rotating the full figure, or by taking a mirror image of the full figure.

3. An unrooted tree has vertices $v_1, v_2, \dots, v_9$ with edges

   $$\{v_1, v_2\}, \{v_1, v_6\}, \{v_1, v_9\}, \{v_2, v_7\}, \{v_2, v_8\}, \{v_3, v_9\}, \{v_4, v_9\}, \{v_5, v_9\}.$$

   a. Without drawing the tree, determine the degree of each vertex.

   b. Use your answer to part (a) to determine the leaves of the tree.

   c. Use your answer to part (a) to determine whether the tree is binary.

   c. Draw the tree to check your work.

4. Consider the trees in Figure 2.7.

   a. Which of them are the same, as rooted metric trees?

   b. Which of them are the same, as unrooted metric trees, provided the root is deleted and its two incident edges joined into one?

   c. Which of them are the same, as rooted topological trees?

   d. Which of them are the same, as unrooted topological trees, provided the root is deleted and its two incident edges joined into one?

   e. Which trees are ultrametric?

5. Draw the three rooted binary topological trees that could describe the relationship between 3 taxa a, b, c. How do they naturally relate to the three unrooted 4-taxon trees in Figure 2.4?

6. Draw the 15 unrooted binary topological trees that could describe the relationship between 5 taxa a, b, c, d, e. Group them naturally according to the relationship they display for the first 4 taxa a, b, c, d.

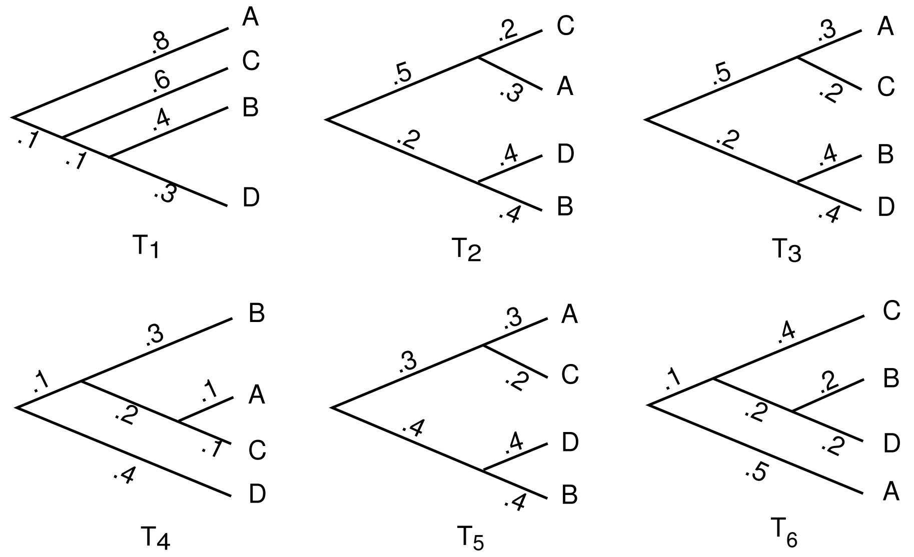

Figure 2.7: Trees for Problem 4

7. Make a table of values of the function $b(n)$, giving the number of unrooted binary X-trees for sets X of size up to 10.

8. Show that $b(n) = \dfrac{(2n-5)!}{2^{n-3}(n-3)!}$.

9. Mitochondrial DNA in humans is inherited solely from the mother. If it is used to construct a tree relating a large number of humans from different ethnic groups, then its root would represent the most recent ancestral female from which we all inherited our mitochondria. The clustering pattern of ethnic groups on such a tree might give insight into the physical location of this woman, who is sometimes called Mitochondrial Eve.

   In 1987 a work by Cann, Stoneking, and Wilson, claimed to locate Mitochondrial Eve in Africa, supporting the 'out of Africa' theory of human origins. A rooted tree was constructed that showed relationships between 147 individuals. Estimate how many topologically different trees would need to be looked at if every possibility was really examined. (You will need to use the statement in problem 8 and Stirling's formula: $n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}$. If you are not familiar with the asymptotic symbol '$\sim$' you can loosely interpret it as meaning 'is approximately.')

10. Give a complete proof of Theorem 1, that if T is a tree and $v_1, v_2 \in V(T)$ then there is a unique path from v_1 to v_2. (Note: You need to be careful with the fact that the term 'cycle' can only apply to sequences of at least 3 vertices.)

11. Prove Proposition 5.

12. Suppose T is a metric tree with all positive edge lengths. Prove that v_3 lies on the path from v_1 to v_2 if, and only if, $d(v_1, v_2) = d(v_1, v_3) + d(v_3, v_2)$. Show by example that this may be false if zero-length edges exist.

13. The phylogeny of four terminal taxa A, B, C, and D are related according to a certain metric tree with positive branch lengths. The tree metric distances between taxa are given in Table 2.1

	A	B	C	D
A		.5	.4	.8
B			.3	.5
C				.6

Table 2.1: Distances between taxa for Problem 13

a. Using any approach you wish, determine the correct unrooted metric tree relating the taxa, as well as all edge lengths. Explain how you rule out other topological trees.

b. Can you determine the root from these data? Explain why or why not.

c. Suppose you only had a table of numbers that were approximately those coming from a tree metric. Would whatever method you used in (a) still work? (This is the situation for real phylogenetic problems.)

14. The term ‘ultrametric’ originally was not applied to trees, but rather to a function $d : V \times V \to \mathbb{R}^{\geq 0}$ that satisfied not only the properties of a metric listed in Proposition 5, but also a strong form of the triangle inequality:

$$d(v_1, v_3) \leq \max(d(v_1, v_2), d(v_2, v_3)) \text{ for all } v_1, v_2, v_3.$$

a. Show this property implies the usual triangle inequality (iii) of Proposition 5.

b. Show that for a tree metric arising from an ultrametric tree, the strong triangle inequality holds when v_1, v_2, v_3 are leaves.

c. Show by a 3-leaf example that the strong triangle inequality on leaves does not hold for all tree metrics.

d. Show that if the strong triangle inequality holds, then for all choices of v_1, v_2, v_3 the two largest of the numbers $d(v_1, v_2)$, $d(v_1, v_3)$, and $d(v_2, v_3)$ are the same. (This is sometimes stated as: An ultrametric implies all triangles are isoceles.)

e. Show that if a tree metric from an unrooted 3-leaf tree satisfies the strong triangle inequality on the leaves, then there is a placement of a root for which the underlying tree is ultrametric. (This holds more generally for n-leaf tree metrics satisfying the strong triangle inequality on leaves; with proper placement of a root, they all arise from ultrametric trees.)

15. Suppose in a graph $G = (V, E)$ an edge $e = (\alpha, \beta)$ is subdivided by the introduction of a new node γ, to obtain a graph $G' = (V', E')$. What are V' and E'?

16. A rooted tree T is specified in Newick notation by $((a, b), c), d))$. List the 7 other Newick specifications for it.

17. How many different Newick specifications can be given for a rooted binary tree relating n taxa? Explain your formula.

18. How many different (rooted) Newick specifications can be given for an unrooted binary tree relating n taxa? Explain your formula.