15

Phylogenetics

Introduction

One of the purposes of this chapter is to introduce the reader to the new mathematical field of algebraic statistics; cf. [5]. Among the many topics in biology in which algebraic statistics is making an impact, we have chosen phylogenetics as the vehicle for showcasing this new discipline. Our reasons are that

- phylogeny and cladistics are important semiclassical fields in biology (with beginnings in the mid-1950s) quite different from anything we have studied up to now;
- postgenomics phylogeny makes extensive use of algebraic statistics and demonstrates more of its techniques than other branches of biology;
- phylogeny draws heavily on genomic searches, which we studied in the last chapter, and hence reinforces what we investigated there; and
- phylogeny is related to several of the new fields of biology that have arisen with genomics that we outlined in the first section of the genomics chapter, Section 14.1.

Algebraic statistics, as mentioned above, is a new branch of mathematics arising out of the many needs and uses of mathematics in genomics. Not surprisingly, the basic mathematics of algebraic statistics originates in the fields of algebra and statistics, but already new mathematics, inspired by the biology, has been created in the discipline.

This chapter will take us to a higher level of mathematical abstraction, skill, and reasoning than in the other chapters of the book and is likewise more demanding. As in the earlier parts of the book, we make every effort to explain the mathematics we need from first principles, principles that one would encounter in two years of a college mathematics curriculum, one that includes linear algebra. Still, very little abstract algebra makes its way to this level, and so we pay extra attention to illustrate the ideas and terms with examples.

Phylogenetic trees contain a great deal of biological and evolutionary information. Taxa closer together on the tree signify a greater degree of shared evolutionary novelties. The tree shows ancestral relationships among taxa and indicates the geological time the process of evolution has taken step by step. We will see that trees

R.W. Shonkwiler and J. Herod, *Mathematical Biology: An Introduction with Maple and Matlab*, Undergraduate Texts in Mathematics, DOI: 10.1007/978-0-387-70984-0_15,

are constructed using several lines of observation including ontologic, morphologic, physiologic, the fossil record, and finally genomic.

15.1 Phylogeny

Phylogenetics elucidates the history of evolution.

Phylogenetics is the study of the evolutionary relatedness among various groups of organisms, for example among species. Derived from the Greek—*phylon* means tribe or race and *genetikos* means relative to birth—phylogenetics attempts to reconstruct and explain the pattern of events that have led to the distribution and diversity of life as it exists at the present time. The results of a phylogenetic study is a tree diagram graphically depicting ancestor–descendant relationships over evolutionary time. An example, the tree of life, showing all three domains of cellular life, is presented in Figure 15.1.1. This tree was pieced together by Carl Woese and colleagues by comparing base pair sequences of the 16S ribosomal RNA gene.

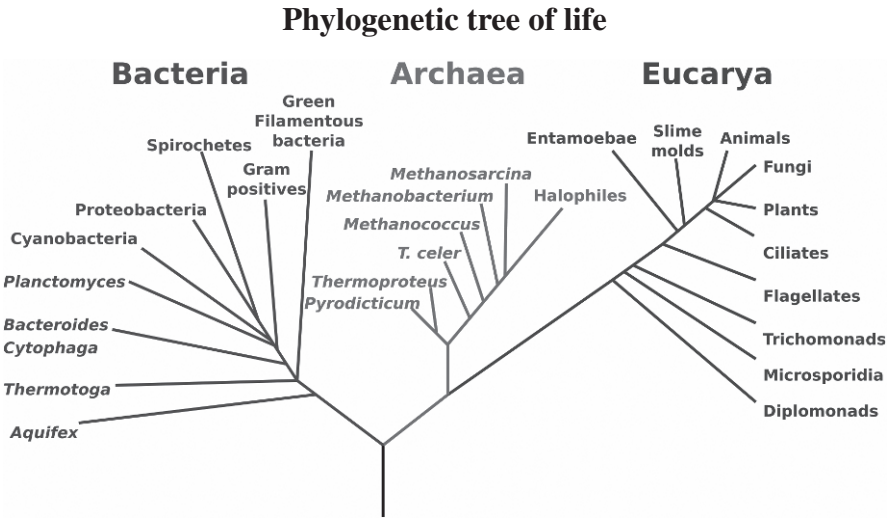


Fig. 15.1.1. The tree of life.

In addition to the ancestor–descendant relationships given in a phylogenetic tree, the tree also implies sets of shared, nested attributes, called *characters*, possessed by organisms farther along each branch of the tree. As a result, constructing a phylogenetic tree resolves two sets of problems. By far the more difficult of the two, without the use of genomics, is establishing the ancestor–descendancy relationship. The science of *cladistics* separates these two problems and concentrates solely on organizing groups of organisms according to nested sets of characters. Similar to

phylogenetic trees, *cladists* portray their results in *cladograms*, which are also tree diagrams. The fundamental units of comparison in cladograms are *taxa*. These are collections of organisms sufficiently distinct from other sets to be given formal names and placed in a Linnaean hierarchy. Often taxa are taken to be species, but, as in the tree of life, they can be more inclusive sets as well.

In this section, we discuss the methodology for constructing and testing cladograms using classical, nongenomic, characters. However, since the end result of the analyses of phylogenetics and cladistics is presented in tree diagrams, it is important to understand how they are used.

Tree diagrams contain a wealth of information.

A *tree* (diagram) is a graph consisting of *nodes* or *vertices* and *edges* or *branches*. The nodes of a tree are labeled in some fashion, for example by the positive integers $1, 2, \ldots, N$. The *size* of a tree is the number of its nodes, N in this case. Each edge connects two nodes, for example an edge might connect nodes 1 and 2; this is indicated by the notation $(1, 2)$. The edge $(1, 2)$ is *incident* on both node 1 and node 2. A *path* from node a to node b is a chain of edges, $(a, a_1), (a_1, a_2), \ldots, (a_n, b)$ connecting a and b. A tree is *connected*; this means that every pair of distinct nodes is connected by a path. There are no circular paths in a tree, that is, paths beginning and ending on the same node. Some examples of trees are shown in Figure 15.1.2. Note that while trees (b) and (c) of the figure appear to be different, they are actually equivalent, since they have the same incidence structure.

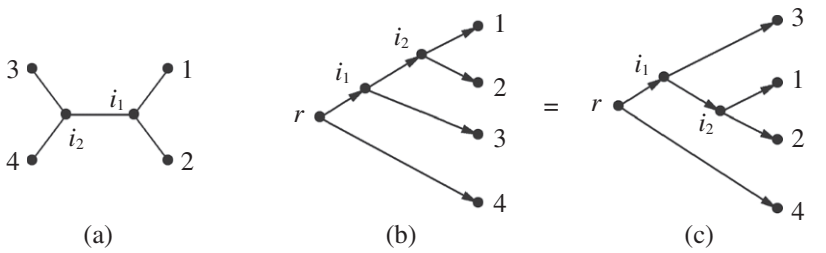


Fig. 15.1.2. (a) Undirected four-leaf tee. (b) Directed four-leaf tree. (c) Tree equivalent to (b).

Given a subset S of nodes of a directed tree, the *subtree containing* S is the part of the tree starting from the nearest common ancestor r to S and including all descendants of r. This concept is termed *monophyletic* in phylogenetics; a monophyletic group is a taxon and all of its descendants.

The number of edges incident at a node is its *degree*. Nodes having degree 1 are *leaf* nodes (except for the root node of a directed tree; see next). Nodes that are not leaf nodes are *interior nodes*. Nodes 1, 2, 3, and 4 are leaf nodes of all three trees in Figure 15.1.2; nodes i_1 and i_2 are interior nodes.

Frequently, a tree indicates the passage of time. In this case the edges are *directed* with the forward direction that of forward in time. The forward direction of a directed tree is indicated by means of arrows. However, when the direction is clear from the context, we will omit the arrows. Trees (b) and (c) of Figure 15.1.2 are directed.

By definition, in a directed tree, only one edge can lead into any given node of the tree. Thus there is a unique node with no edge leading into it; this is the *root node*. A (forward) path in a directed tree must follow the edge directions. A tree having exactly two edges branch out from each interior node is said to be *binary*. Both directed trees of Figure 15.1.2 are binary; the root node is indicated by r.

The leaf nodes of a tree represent taxa for which we have data. The interior nodes of a phylogenetic tree represent hypothetical ancestors. If Figure 15.1.2(b) represents a phylogenetic tree, then i_2 is the nearest common ancestor of taxa 1 and 2. Likewise, i_1 is the nearest common ancestor of 1, 2, and 3. The root node of a phylogenetic tree represents the ancestor of all taxa of the tree.

In contrast, the vertices of a cladogram only represent sister taxa with respect to some shared character. If Figure 15.1.2(b) represents a cladogram, its message is that taxa 1 and 2 share some evolutionarily novel attribute that taxa 3 and 4 do not possess. Likewise, taxa 1, 2, and 3 share some character that 4 does not possess. The term *synapomorphy* means sharing a derived character from an immediate common ancestor. Thus a cladogram expresses a series of synapomorphies.

A tree may be represented in text using nested parentheses to enclose all descendants of a node. The tree shown in Figure 15.1.3(a) has the representation (((2, 4), 1), 3), while that in (b) is written ((1, 3), (2, 4)).

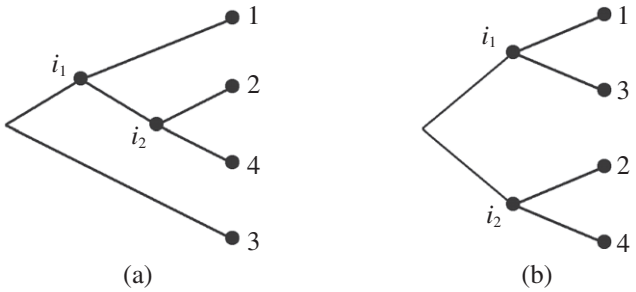


Fig. 15.1.3. (a) (((2, 4), 1), 3) tree. (b) ((1, 3), (2, 4)) tree.

In addition to direction, the edges of a tree can depict other facts about taxa. The length of an edge can show the relative distance, in some sense, separating two taxa, for example, protein sequence distance. This is shown in Figure 15.1.4. Such a tree is called an *additive tree*, and it defines a *tree metric* in that the sum of the branch lengths along the unique path (not necessarily always forward) connecting two nodes gives the distance between them. In this way, we may construct the following distance matrix for the leaf nodes of Figure 15.1.4:

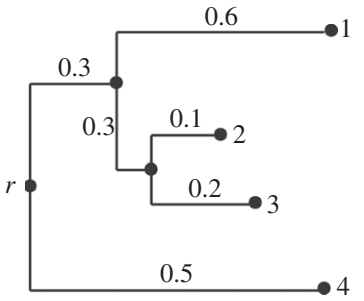


Fig. 15.1.4. Branch lengths give the relative distance between nodes.

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} - & 1.0 & 1.1 & 1.4 \\ 1.0 & - & 0.3 & 1.2 \\ 1.1 & 0.3 & - & 1.3 \\ 1.4 & 1.2 & 1.3 & - \end{pmatrix} \end{array}.$$

Note that the root r of Figure 15.1.4 is superfluous and may be omitted from such a graph.

Cladograms express nested evolutionary novelties.

William Hennig founded the field of cladistics in 1950 based on the principle that the evolutionary process produces, as an expectation, a nested set of evolutionary novelties. To illustrate the ideas and methods of cladistics, we begin with an example—construct a cladogram for the following organisms:

beaver, dolphin, salamander, shark, trout, and turtle.

Because these are extant organisms, many features and lines of evidence are available for distinguishing their similarities and differences. These include ontogenesis and anatomical, physiological, developmental, biochemical, and behavorial characteristics. These are called *intrinsic* characteristics of organisms. *Extrinsic* characteristics are their distribution in space and time. Of course, since the advent of genomics, DNA and protein sequence comparisons are also available, but our emphasis here is the use of nongenomic evidence. We postpone consideration of sequence matching data to the subsequent sections of this chapter.

The first task is to find general similarities shared by all the organisms in the study that could be used to define a group containing all of them. These characteristics are called the *universal set* of the comparison. In our example, these include physical symmetry, possession of an endoskeleton, appendages, a chambered heart, a dorsal nerve cord, a notochord, and visceral or gill pouches at some stage of the life cycle. Establishing a universal sets of characteristics helps in selecting outgroups used to fix evolutionary subgroups within the organisms of the study. We will see how this works below.

Examining the species of this study yields several sets of similarities and differences among them. Each has unique features: One has hair, one has a shell, one has a cartilaginous skeleton, and so on. Some characteristics are common to two or three of the subjects but not to the others: Two possess mammary glands, two have fleshy fins, three have limbs, two lack lungs, three have an amniotic egg, and so on. What principles should be used in constructing the cladogram? Maybe beavers should branch off first as the only one of the group having hair; maybe turtles should branch off first as the only one with a shell.

One of the guiding principles of cladistics is evolutionary descent: Organisms are related by descent from a common ancestor. The direction of descent is called *polarity*. Hence cladograms are constructed according to nested sets of evolutionary novelties, that is, synapomorphies, with more inclusive characters appearing nearer the root and more recent novelties shown nearer the leaves. Among the characteristics cited above, how does one decide on their evolutionary descent?

Developmental processes help in deciding polarity.

One line of evidence occurs during *ontogeny*, the development of an organism from a fertilized egg. During development, a trait or attribute similar to that of an ancestral species may be observed for a time, only to have it disappear at a later stage. This is known as *recapitulation*.

The first to state generalized rules of ontogeny, based on detailed studies in the 1820s, was Karl von Baer. They are the following:

1. In development from the egg, the general characteristics appear before the special characteristics.
2. From the more general characteristics, the less general and finally the special characteristics are developed.
3. During its development, an animal departs more and more from the form of other animals.
4. The young stages in the development of an animal are not like the adult stages of other animals lower down on the scale, but are like the young stages of those animals.

An example is the appearance of proto-pharyngeal gill pouches in almost all mammalian embryos at early stages of development. In particular, this applies to our example organisms. Accordingly, organisms in our group having gills, sharks and trout, should be placed toward the root of the cladogram, while lungs will be considered an evolutionary novelty.

The same rationale for ontogeny applies as well to developmental morphology. An evolutionary novelty is often the modification of some preexisting feature within the universal set. During preadult development, the modification might be repeated among derived organisms. In fact, this is the case for our example organisms with respect to skeletal composition. All six start out with cartilaginous skeletons, but except for the shark, the others replace much of this with bone prior to adulthood. Consequently, the bony members of our study are deemed a later subgroup within the group.

Our researches so far take the following form:

(universal set : shark + (bony skeleton : trout
+ (beaver, dolphin, salamander, turtle))).

This leaves four taxa to resolve.

Another striking developmental modification is seen in the salamander. As larvae, salamanders respire with gills; meanwhile, the lungs they will need as terrestrial adults are under development. Since the taxa we have already placed on the cladogram also have gills and the remaining taxa breathe with lungs, we are led to regard lungs as a derived evolutionary novelty and place salamanders next in sequence. The term used by cladists for a derived evolutionary characteristic is *apomorphy*, from the Greek *apo*, meaning away, and *morphy* for form. The term for the opposite is *plesiomorphy* (Greek: close form), meaning a primitive characteristic relative to the study group and therefore more widely shared.

Continuing in this way using ontogeny and developmental morphology, we formulate the cladogram of Figure 15.1.5.

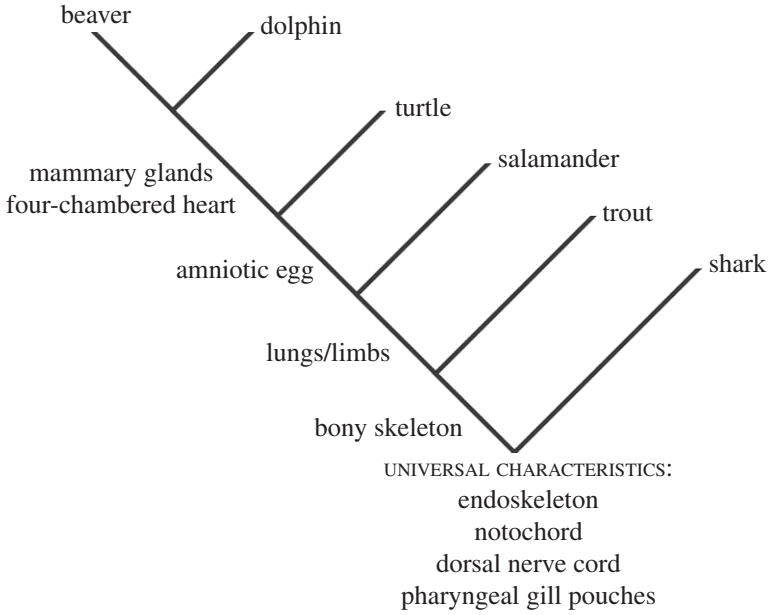


Fig. 15.1.5. Cladogram for example set of six taxa.

Outgroup comparison also provides evidence.

Ontogeny and developmental morphology are not the only criteria used in constructing and corroborating a cladogram. Outgroup comparison is another major source of clues

for determining polarity. An *outgroup* is a taxon related to and believed to be more primitive than the organisms being classified. The perfect outgroup would be a close ancestor of these taxa and would have the primitive form of the characteristics being compared. In the real world, it is impossible to be certain that an outgroup chosen for cladistic analysis is actually the ancestor, or whether its traits are all truly primitive. The existence of *convergent evolution* is a primary complication in cladistics, since it confuses the identification of primitive characteristics. We briefly digress to explain.

In our example, dolphins have been included within the set of animals having evolved terrestrial limbs. But instead, they possess appendages similar to those of sharks. Their assignment therefore has an inconsistency. But the inconsistency is resolved if it were the case that the immediate line of ancestors of the dolphin, having taking up life in the sea, gradually modified their limbs into the finlike structures of today's animal. In fact, this is what cladists believe. *Convergence* is the independent evolution of similar structures to solve the same biological problem, in this case movement through water. Note that despite having returned to the sea, the ancestors of the dolphin did not revert to oxygen exchange via gills but retained their lungs.

To see how outgroups can help, consider the diagrams in Figure 15.1.6. Three taxa A, B, and C present variation with respect to two different traits, a and b. Taxon A shows traits a_2 and b_2, taxon B shows a_2 and b_1, and C shows a_1 and b_2. If trait a_2 is derived from a_1, then Figure 15.1.6(a) is the correct one, but if trait b_2 is derived from b_1, then Figure 15.1.6(b) captures the development. Upon consideration of

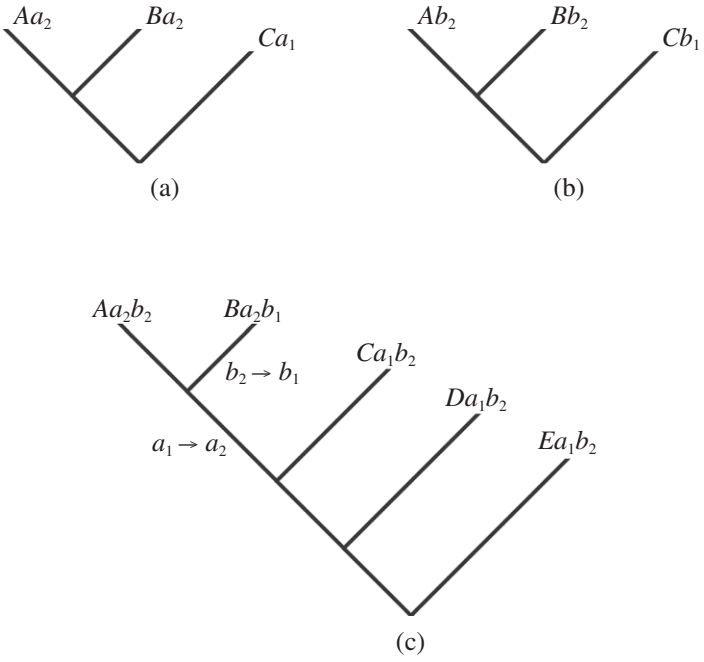


Fig. 15.1.6. Outgroups show that traits a_1 and b_2 are primitive.

outgroups D and E, we find that they possess characteristics a_1 and b_2. Hence these are taken as the primitive forms and we get Figure 15.1.6(c).

Following the discussion above, as an outgroup we seek taxa that could be close ancestors of our study group. This is the purpose of establishing a universal set of characteristics. Evidently, organisms possessing these are the most likely to meet the ancestor criteria. In this example, the lamprey can serve as an outgroup for us, since they possess all the characteristics of the universal set. Since the lamprey does not have a bony skeleton at any stage of its life, we judge that this is a primitive trait. This corroborates our placement of sharks as the first branch of the cladogram.

Having made this determination, we may now include sharks as another outgroup in deciding derived traits for the remaining set: trout, salamander, turtle, dolphin, beaver. Among these, what evolutionary novelties are shared by all but one? We notice that trout lack lungs and fleshy limbs, and, moreover, these are not present in either outgroup, lamprey and sharks. We thus regard lungs as an apomorphy. Continuing to apply these principles, we again arrive at Figure 15.1.5.

Cladograms help in constructing phylogenetic trees.

As previously mentioned, phylogenetic trees go beyond cladograms in that they assert ancestor–descendant relationships. Since biological species are delineated according to the potential of its members to interbreed, species are necessarily the units of evolution.[1]

One can speak of one species as being the ancestor of another. Properly then, the taxa underlying a phylogenetic tree are species. Nevertheless, phylogenetic trees are constructed for higher biological units, for example reptiles as descending from amphibians. The inference is that some amphibian species—*Seymouria* has been cited—is the particular ancestor of the line leading to the reptiles.

Since a cladogram has less information than a phylogenetic tree, each subgraph of a cladogram can be explained by any one of several trees. For example, consider the cladogram shown in Figure 15.1.7. The information here is that taxa A and B share some derived characteristic, a synapomorphy, not possessed by C. Figure 15.1.8 shows six possible ways this could happen.

In the first of these, A and B derive from a common ancestor, and that ancestor and C do likewise. This explains how A and B can have a shared characteristic while C does not have it. In (b), C is itself the ancestor of the common ancestor of A and B. Again, the synapomorphy could derive from this nearest common ancestor and thus not from C. In the other diagrams, one of A or B is the ancestor of the other; for example, A is the ancestor of B in (f). The assumption is that the characteristic in question first appeared, with respect to the diagram, in A and was passed on to B. And so once again, A and B share it, while C does not, as consistent with the cladogram.

[1] Formulating a testable definition of species is a challenge; how can one show that two organisms widely separated in space or time are or were capable of interbreeding? For an in-depth discussion, see [3].

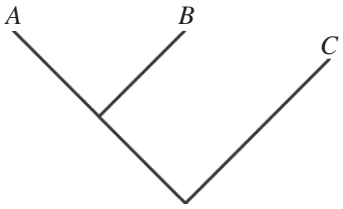


Fig. 15.1.7.

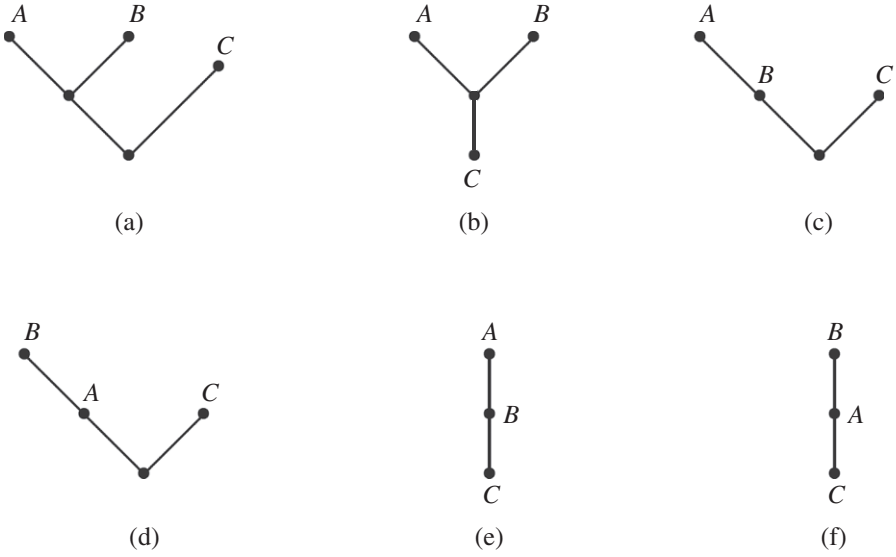


Fig. 15.1.8.

Therefore, additional information is needed to promote a cladogram to a phylogenetic tree. That information is the evolutionary history of speciation events occurring to the organisms being studied. But the sought-after ancestors are, in most cases, extinct organisms. As a result, until the advent of genomics, the information had to come from the fossil record. Continuing to focus on pregenomic methods here, we inquire into the nature of speciation.

In fact, the way in which new species arise is still not well understood; see [6]. We content ourselves with mentioning some of the possibilities. The *transformational* mechanism of speciation is that over time, a species transforms by accumulating modifications, either gradually or all at once (*saltationism*), emerging as another species. If the transformation is gradual and the species is a type that leaves fossils, then the fossil record should contain the multitude of intermediates through time. But the fossil record is generally incomplete, and what we see are merely "snapshots" separated by gaps in the record. Hence there is no need to speculate as to when the new species emerged during the gradual change; the incomplete fossil record has performed this task.

In the saltation theory, evolution proceeds by sudden jumps. Of course, paleontologically, a jump could mean a few hundreds of thousand years. Several mechanisms have been proposed for saltation. For example, a genetic mutation could produce a bifurcation in the apparatus producing hair and instead produce a proto-feather. The mutation does not interfere with interbreeding. In time, those with the mutation do better than those without and come to dominate the gene pool. Eventually, the original gene disappears altogether. There is fossil evidence for this mechanism; see [1].

Transitional speciation, whether gradually or as a jump, produces a phylogenetic tree as exhibited in Figure 15.1.9(a).

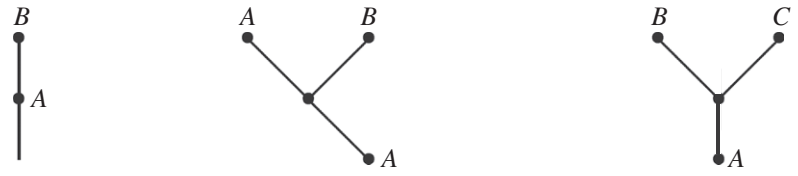


Fig. 15.1.9. Speciation possibilities.

The transitional theory does not produce diversification; the species count remains the same. But the other major theory of speciation, that of splitting, does increase the count. The idea here is that a portion of the population becomes reproductively isolated. This could be the result of geographical circumstances—a mountain range develops, a canyon widens, the subpopulation is transported to a new continent or an island, and so on. Over time, differences accumulate between the subpopulation and the main population. This could stem from natural selection or just by genetic drift. Such an isolated subpopulation already starts off genetically different from the main body. Every genetic trait has a specific distribution of its alleles among an interbreeding population. But any given subpopulation will have profound allelic differences for some subset of traits just by chance alone. Eventually, the subpopulation becomes reproductively incompatible with the original population, and a new species has thus been created.

Reproductive isolation strictly enforced by some geographical barrier is called *allopatric speciation* of type A. In allopatric speciation of type B, the isolation stems from the fact that the population is large and widely separated in distance (relative to the organism). As a result, individuals on the margins never have the opportunity to breed with other parts of the population. As above, these units proceed toward establishing a distinct species. Speciation by splitting results in a phylogenetic tree as exemplified in Figure 15.1.9(b) or (c). In (c), the ancestral species itself produces a new species by transition.

Once a tree has been constructed, its assertions and predictions must stand up to testing. Of course, a phylogenetic tree must be consistent with any cladogram depiction of its taxa. A cladogram that shows species *B* to have derived characteristics

with respect to species A cannot have A as a descendant of B. Thus the evidence used in cladistics is applicable to phylogenetic trees as well.

Paleontology can be a source of falsification, or support, as well. The data here will be extrinsic, that is, information about the distribution of the organisms in space and time. The assertion that one species is the descendant of another would presuppose that fossils of the ancestor should predominately be found in older rock and those of the descendant in younger rock. Or in another case, a speciation event postulated to have taken place at a point in time would suppose the geographical distribution of the ancestor to be more widespread and their numbers to be larger. However, the nature of the fossil record is not so precise. Observed stratigraphic ranges cannot be assumed to be the total life span of a species. What is sampled may only be a portion of the total life span. Similarly, assessing the geographical range of a species is likewise problematical. In many instances, sediments were not even deposited in all areas where a species had been living. And where deposition did occur, there is no guarantee about the fossils that were deposited there or that will remain and be found.

But now a new and powerful tool is available for addressing these issues. That tool is genomics. Even though ancient DNA and protein samples are unavailable, new techniques and lines of attack may be brought to bear on these problems.

15.2 Branch Lengths Estimate the Separation of Species

In this section, we find that natural assumptions about the rate of molecular evolution leads to a quantitative description of the phenomenon via mathematical semigroups. Moreover, in order to estimate mutation rates, we employ the widely used technique of maximum likelihood estimation. We thus begin our first encounter with algebraic statistics. Throughout the remainder of this chapter, we closely follow some of the topics in the groundbreaking text by Pachter and Sturmfels [5]. In the next section, we take time out from our study of phylogenetics to introduce all the additional mathematics that will be needed. For more detail and to learn about the full scope of algebraic statistics applications in genomics, see Pachter and Sturmfels [5].

The molecular clock assumption asserts that molecular evolution is constant over time.

In 1965, Emile Zuckerkandl and Linus Pauling proposed the theory of a *molecular clock*, which states that the rate of molecular evolution is approximately constant over time for all the proteins in all lineages. According to this theory, any time of divergence between genes, proteins, or lineages can be dated simply by measuring the number of changes between sequences. Soon afterward, in 1969, Thomas Jukes and Charles Cantor (1969) proposed a stochastic model for DNA substitution in which all nucleotide substitutions occur at an equal rate, and when a nucleotide is substituted, any one of the other nucleotides is equally likely to be its replacement.

In this section, DNA alignment data is used to compute branch lengths under the assumption of a Markov model for point mutations. These assumptions are patterned after the molecular clock theory:

1. mutations occur at random, dependent only on a mutation rate;
2. mutations occur independently at different sites;
3. (continuous time assumption) at any instant in time, there is a nonzero probability that a mutation will occur.

As in all good science, the assumptions may be oversimplified, but they do capture the essence of the phenomenon and form the basis of a starting point for studying the subject.

While the Markov assumptions may apply to molecular evolution from ancestral species well enough, the requirement of DNA sequence data limits these methods to taxa for which there is such data. Hence at present these methods give rise to unrooted phylogenetic trees among existing species.

Although the model applies to point mutation phenomena in general, for example, protein evolution, we will specialize to DNA mutation. Thus our indices run over the set of bases $\Sigma = \{\text{A, C, G, T}\}$ taken in alphabetical order.

The problem of calculating branch lengths.

In consequence of the independence assumption, we can consider the mutation at each site one by one. At any single site, there is the probability $\theta_{ij}(t)$ that base i will have changed to base j after time t. The path of the change is not considered: i may have changed directly to j, or may have changed to some intermediate k that changed to j, or any other of many possibilities. Let $\theta(t)$ denote the 4×4 matrix of these probabilities. It represents the cumulative effect of changes over a time period t and is called the *substitution matrix*. Thus $\theta(0) = I$, the identity matrix.

A consequence of the Markov assumptions is that the process is "memoryless": Over a period of time $s + t$, and decomposing on base k the process visited at time s (should a base be impossible at time s, then the probability of the transition from i to k will be 0), we have

$$\Pr(i \to j \text{ over time } s+t) = \sum_{k \in \Sigma} \Pr(i \to k \text{ over time } s) \cdot \Pr(k \to j \text{ over time } t).$$

In terms of matrix multiplication, this is exactly

$$\theta(s+t) = \theta(s)\theta(t), \quad s \geq 0, \quad t \geq 0. \tag{15.2.1}$$

A family of matrices, indexed by t, satisfying (15.2.1) is a mathematical *semigroup*. In turn, this implies the existence of an *infinitesimal generator* or *rate matrix* Q having the properties

$$\begin{aligned}
\theta(t) &= e^{Qt} = \sum_{n=0}^{\infty} \frac{1}{n!} Q^n t^n, \\
\theta'(t) &= \theta(t)Q = Q\theta(t), \quad t \geq 0, \quad {}' \text{ signifying the derivative,} \\
\theta^{(k)}(0) &= Q^k, \quad k \geq 0, \quad (k) \text{ signifying the } k\text{th derivative.}
\end{aligned}$$

The off-diagonal elements of Q are the transition rates between bases per unit time. Thus q_{ij}, with $i \neq j$, is the (average) instantaneous rate at which base i mutates into base j. Mathematically, the rows of Q must sum to 0,

$$q_{ij} \geq 0, \quad i \neq j, \qquad q_{ii} < 0,$$
$$\sum_{j \in \Sigma} q_{ij} = 0 \quad \text{for all } i \in \Sigma.$$

Two widely used rate matrices are the *Jukes–Cantor*,

$$Q_{\mathrm{JC}} = \begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix}, \tag{15.2.2}$$

and the *Kimora*-80,

$$Q_{\mathrm{K80}} = \begin{bmatrix} -(\alpha+2\beta) & \beta & \alpha & \beta \\ \beta & -(\alpha+2\beta) & \beta & \alpha \\ \alpha & \beta & -(\alpha+2\beta) & \beta \\ \beta & \alpha & \beta & -(\alpha+2\beta) \end{bmatrix}. \tag{15.2.3}$$

In Jukes–Cantor, the first row says that the rates at which adenine (A) mutates into cytosine (C) or guanine (G) or thymine (T) are the same and equal α (a parameter of the model). Likewise, the other rows say the analogous thing for cytosine, guanine, and thymine.

In Kimora-80, the rate for a purine to a purine or a pyrimidine to a pyrimidine base is α, while from a purine to a pyrimidine or conversely is β. This more accurately reflects actual mutation rates at the expense of introducing a second parameter, β.

Using the Jukes–Cantor rate matrix, the substitution matrix can be computed by MAPLE:

```
MAPLE (symbolic, no MATLAB equivalent)
> with(LinearAlgebra):
> assume(a>0):
> Q:=Matrix([[-3*a,a,a,a],[a,-3*a,a,a],[a,a,-3*a,a],[a,a,a,-3*a]]);
> MatrixExponential(Q)
```

Regarding $a = \alpha t$, this gives

$$\theta(t) = e^{Qt} = \frac{1}{4}\begin{bmatrix} 1+3e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} \\ 1-e^{-4\alpha t} & 1+3e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} \\ 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1+3e^{-4\alpha t} & 1-e^{-4\alpha t} \\ 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1-e^{-4\alpha t} & 1+3e^{-4\alpha t} \end{bmatrix}. \tag{15.2.4}$$

Under the Markov assumptions enumerated on p. 508 above, the course of the ensuing mutation is an instance of a mathematical process known as a *Poisson process*. This means that the distribution of mutation events is given by

$$\Pr(k \text{ events in time } t) = \frac{(\lambda t)^k}{k!} e^{\lambda t},$$

where λ is the Poisson event rate. In turn, this is the average mutation rate imposed

by Q. The event rate for base i is the sum of its mutation rates to the other bases; but this is exactly the negative of the diagonal element, $-q_{ii}$. Therefore, the average event rate is

$$\lambda = -\frac{1}{4}\,\mathrm{trace}(Q),$$

where the *trace* of a matrix is the sum of its main diagonal elements. The expected number of events over time t of a Poisson process is λt, and so the expected number of mutations over time t is

$$\text{branch length} = -\frac{1}{4}\,\mathrm{trace}(Q)\cdot t. \tag{15.2.5a}$$

As indicated, the expected number of events is taken as the branch length of the phylogenetic tree. For the Jukes–Cantor model, we get

$$\text{branch length} = 3\alpha t. \tag{15.2.6}$$

By the diagonalization theorem of Section 2.6, (2.6.2), one can show that $\log\det(e^Q) = \mathrm{trace}(Q)$; hence in terms of the substitution matrix directly,

$$\text{branch length} = -\frac{1}{4}\log\det(\theta(t)). \tag{15.2.5b}$$

Estimating branch lengths leads to the method of maximum likelihood.

Now suppose that we are given an alignment between two DNA sequences and we want to estimate the branch length between them with respect to a rooted tree; see Figure 15.2.1.

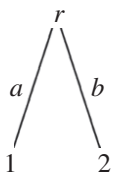


Fig. 15.2.1. Two-claw tree.

The problem is solved by the following theorem.

Theorem 1. *Given an alignment of two sequences of length n, with k differences between their bases, the maximum likelihood estimate of the branch length under the Jukes–Cantor rate model is*

$$\text{branch length} = -\frac{3}{4}\log\left(1-\frac{4k}{3n}\right). \tag{15.2.7}$$

In terms of $c = n - k$, the number of identities between the two sequences, this is

$$\text{branch length} = -\frac{3}{4}\log\left(\frac{4c}{3n}-\frac{1}{3}\right). \tag{15.2.8}$$

Maximum likelihood estimation, as mentioned in the theorem, is the most widely used method for parameter estimation in statistics. It means choosing the value of any unknown parameter in such a way as to make the outcome that was actually observed the most probable. Here is how that works out for branch length estimation.

Let a be the substitution matrix along the left branch of the tree and b that along the right branch. From (15.2.4), these matrices have only two distinct elements: the diagonal elements and the nondiagonal elements. Let

$$a_0 = \frac{1}{4}\left(1 + 3e^{-4\alpha t}\right) \quad \text{and} \quad a_1 = \frac{1}{4}\left(1 - e^{-4\alpha t}\right). \tag{15.2.9a}$$

Similarly, let

$$b_0 = \frac{1}{4}\left(1 + 3e^{-4\beta t}\right) \quad \text{and} \quad b_1 = \frac{1}{4}\left(1 - e^{-4\beta t}\right). \tag{15.2.9b}$$

Then by (15.2.6), the branch length we want to calculate is

$$\begin{aligned} \text{branch length} &= (\text{branch length 1 to } r) + (\text{branch length } r \text{ to } 2) \\ &= 3(\alpha + \beta)t. \end{aligned} \tag{15.2.10}$$

With the abbreviations defined above, we can write

$$a = \begin{bmatrix} a_0 & a_1 & a_1 & a_1 \\ a_1 & a_0 & a_1 & a_1 \\ a_1 & a_1 & a_0 & a_1 \\ a_1 & a_1 & a_1 & a_0 \end{bmatrix}, \qquad b = \begin{bmatrix} b_0 & b_1 & b_1 & b_1 \\ b_1 & b_0 & b_1 & b_1 \\ b_1 & b_1 & b_0 & b_1 \\ b_1 & b_1 & b_1 & b_0 \end{bmatrix}.$$

There are four unknown parameters, a_0, a_1, b_0, and b_1. But since the rows of stochastic matrices must sum to 1, we have

$$a_0 + 3a_1 = 1, \qquad b_0 + 3b_1 = 1. \tag{15.2.11}$$

This is automatically satisfied with the as and bs taken according to (15.2.9).

Next, we calculate the probability that two bases at the leaves 1 and 2 of the tree will be the same. Let p_{AA} be the probability that they are both A; then

$$p_{\mathrm{AA}} = \frac{1}{4}a_0 b_0 + \frac{3}{4}a_1 b_1. \tag{15.2.12}$$

This is seen as follows: If the original base at the root is A, with probability $\frac{1}{4}$, then we will have A at leaf 1 if there is no change, and this happens with probability a_0 according to the edge matrix a. Similarly, the A at leaf 2 remains unchanged with probability b_0. This gives the first term in (15.2.12). But if the original base is not A, and this happens with $\frac{3}{4}$ probability—say it is C—then both leaves will be A if C mutates to A along both edges. The mutation from C to A along the left edge happens with probability a_1 according to the substitution matrix, and along the right edge it

is b_1. So the probability that both leaves will be A when the root was not is $(\frac{3}{4})a_1b_1$. And this is the second term.

The probability that both leaves are C or G or T is the same as for A, and so the probability that both leaves are the same is given by

$$\theta = p_{\text{same}} = a_0b_0 + 3a_1b_1; \tag{15.2.13}$$

denote this by θ. By a similar calculation, the probability that the bases at the two leaves are different works out to be (there are 12 ways they could be different)

$$p_{\text{dif}} = 12\left(\frac{1}{4}a_0b_1 + \frac{1}{4}a_1b_0 + \frac{1}{2}a_1b_1\right) = 3a_0b_1 + 3a_1b_0 + 6a_1b_1 = 1 - \theta.$$

As noted, this equals $1 - \theta$, since it is the complementary event to the leaves being the same.

Now return to our original problem; we have two DNA sequences of length n differing in k places. We apply the results derived above to each place. The probability that k bases are different out of n is like getting k heads out of n tosses of a weighted coin (from Sections 2.8), so we have

$$L(\theta) = \Pr(k \text{ differences out of } n) = \binom{n}{k}(1-\theta)^k\theta^{n-k}. \tag{15.2.14}$$

This is called the *likelihood function* for the model, and for emphasis we show it to be a function of θ.

Now suppose that there were $n = 100$ bases and $\frac{3}{4}$ of them, or 75, remained the same, leaving $k = 25$ to mutate. What value of θ would make this outcome the most likely? It would be $\theta = \frac{3}{4}$. For example, if θ, being the probability that a base remains unchanged, were $\frac{1}{2}$, it would be very unlikely to get $\frac{3}{4}$ of 100 bases unchanged by chance; see Figure 15.2.2.

This is how maximum likelihood works. To maximize (15.2.14), we set its derivative to zero and solve for θ. Alternatively, we could take the logarithm of (15.2.14) and set its derivative to zero. Often likelihood functions are products of factors to various powers, and working with the log likelihood function is easier. The calculation is

$$\begin{aligned} 0 &= \frac{d\log L(\theta)}{d\theta} = \frac{d}{d\theta}\left(\log\binom{n}{k} + k\log(1-\theta) + (n-k)\log\theta\right) \\ &= \frac{-k}{1-\theta} + \frac{n-k}{\theta}. \end{aligned} \tag{15.2.15}$$

The solution is $\theta = \frac{n-k}{n}$. From (15.2.13), this gives

$$\frac{n-k}{n} = a_0b_0 + 3a_1b_1.$$

Now substitute (15.2.9) into this to get

$$\frac{n-k}{n} = \frac{1}{4} + \frac{3}{4}e^{-4(\alpha+\beta)t}.$$

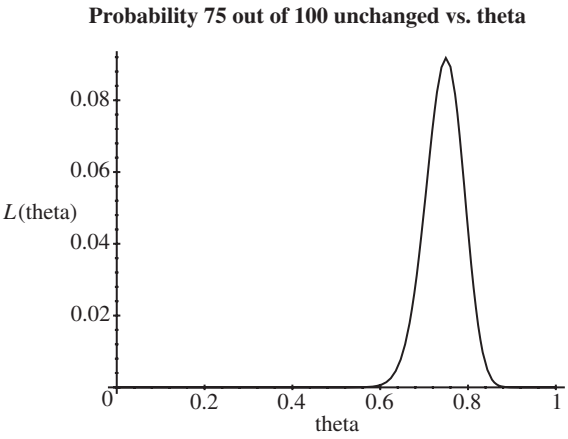


Fig. 15.2.2. Maximizing outcome probability for $n = 100$ and $k = 25$.

Solve this for $\alpha + \beta$ and, remembering (15.2.10), we get the conclusion of Theorem 1,

$$\text{branch length} = -\frac{3}{4}\log\left(1 - \frac{4k}{3n}\right).$$

Keep in mind that, branch length is taken to be the expected number of mutation events over the time between the two observations and is a pure number. If the time between the observations is known, then a mutation rate estimation can be worked out. Conversely, if the mutation rate is known, then the time can be estimated.

We make a last observation on this model. Note that we have not made a determination of α and β individually, only their sum. Likewise, the as and bs occur only combined, as in

$$a_0b_0 + 3a_1b_1 \quad \text{and} \quad a_0b_1 + a_1b_0 + 2a_1b_1. \tag{15.2.16}$$

This is because we took the states of the root to be equally likely; we have information only about differences along the combined link from node 1 through a to r then through b to node 2. The combined edges are governed by the matrix product

$$ab = \begin{bmatrix} a_0b_0 + 3a_1b_1 & a_0b_1 + a_1b_0 + 2a_1b_1 & \dots & \dots \\ a_0b_1 + a_1b_0 + 2a_1b_1 & a_0b_0 + 3a_1b_1 & \dots & \dots \\ \vdots & \vdots & \ddots & \end{bmatrix}.$$

15.3 Introduction to Algebraic Statistics

In order to avoid interrupting the flow of ideas in subsequent sections, we gather together here the concepts and tools of algebra that we will need in those sections.

It all takes place in the ring of polynomials.

A *mathematical ring* is a set S together with two operations, denoted by $+$ and $*$, satisfying the following basic arithmetic laws:

- Associativity:

$$(a+b)+c=a+(b+c), \qquad (a*b)*c=a*(b*c), \quad a,b,c\in S.$$

- Commutativity for $+$ (not necessarily for $*$):

$$a+b=b+a, \quad a,b\in S.$$

- Distributivity:

$$a*(b+c)=a*b+a*c, \quad a,b,c\in S,$$

and

$$(b+c)*a=b*a+c*a, \quad a,b,c\in S.$$

- Existence of an additive identity, 0:

$$0+a=a+0=a, \quad a\in S.$$

- Existence of an additive inverse: For $a\in S$, there is an inverse, denoted by $-a$, such that

$$a+(-a)=(-a)+a=0.$$

Of course, the familiar number systems—the integers $\mathbb{Z}$, the real numbers $\mathbb{R}$, and the complex numbers $\mathbb{C}$—are all rings. Add to the list the rational numbers, $\mathbb{Q}$. A number is *rational* if it is the ratio of two integers, e.g., $\frac{3}{4}$, or $\frac{355}{113}$, or $-\frac{17}{1}$, and so on. But we are interested in rings because the set of polynomials over any one of the number sets above is a ring. Let $\mathbb{Q}[x]$ denote the set of polynomials in the indeterminate x with rational numbers serving as coefficients, that is, expressions of the form

$$a_n x^n + a_{n-1}a^{n-1} + \cdots + a_1 x + a_0,$$

where $a_n, \ldots, a_0$ belong to $\mathbb{Q}$ with $a_n \neq 0$; n is the *degree* of the polynomial. The $+$ and $*$ operations are the usual polynomial addition and multiplication,

$$\begin{aligned}
&(a_n x^n + \cdots + a_1 x + a_0) + (b_m x^m + \cdots + b_1 x + b_0)\\
&\quad = a_n x^n + \cdots + (a_m + b_m)x^m + \cdots + (a_1 + b_1)x + (a_0 + b_0) \quad \text{if } n \geq m,\\
&(a_n x^n + \cdots + a_1 x + a_0) * (b_m x^m + \cdots + b_1 x + b_0)\\
&\quad = a_n b_m x^{n+m} + (a_n b_{m-1} + a_{n-1} b_m)x^{n+m-1}\\
&\qquad + (a_n b_{m-2} + a_{n-1} b_{m-1} + a_{n-2} b_m)x^{n+m-2} + \cdots .
\end{aligned}$$

The additive identity is the zero polynomial, with all coefficients equal to zero, and the additive inverse of $f(x)$ is the polynomial all of whose coefficients are the negatives of those of f. Checking the laws is straightforward but tedious and we omit it.

More generally, the multivariate polynomials form a ring as well. Let $\mathbb{Q}[x_1, x_2, \ldots, x_d]$ denote the set of sums of terms of the form

$$a_{i_1 i_2 \cdots i_d} x_1^{i_1} x_2^{i_2} \cdots x_d^{i_d}, \tag{15.3.1}$$

where the $i_1, i_2, \ldots, i_d$ are nonnegative integers. A polynomial consisting of a single such term, as in (15.3.1), is referred to as a *monomial*. Again $+$ and $*$ are taken as polynomial addition and multiplication. For brevity, we also use the notation $\mathbb{Q}[\mathbf{x}]$ to refer to this space where $\mathbf{x} = (x_1, \ldots, x_d)$.

In the axioms of a ring listed on the previous page, we noted that commutativity for $*$ was not required. But in all of our examples this property is realized. When commutativity for $*$ holds, the ring is *commutative*. In all that follows, we will need only the polynomial rings introduced above, and therefore all our rings are commutative. In what follows, we will refer to $+$ as addition and $*$ as multiplication.

Ideals play a fundamental role in rings.

For a commutative ring S, an *ideal* I is a subset of S that is *closed* under $+$ and closed over multiplication by elements in S. In other words, the following two properties are satisfied

$$a + b \in I \quad \text{if } a, b \in I$$

and

$$a * r = r * a \in I \quad \text{if } a \in I \text{ and } r \in S.$$

We do not consider the empty set an ideal. On the other hand, obviously every ring is an ideal of itself. More importantly, an ideal $I \subset S$ is a ring in its own right. The reason is that the elements of I satisfy the axioms of a ring over I, since they do so over S. It remains only to see that I is closed under $+$ and $*$ and that $0 \in I$. The first is part of the definition. The second follows immediately, since if $a \in I$ and $b \in I$, then $b \in S$ and so $a * b \in I$ by the second part of the definition. Finally, if $a \in I$, then $0 = a * 0 \in I$, since $0 \in S$.

Example. Consider the set I in $\mathbb{Q}[x]$ consisting of all polynomials of the form $f(x) * (x - 1)$, where $f(x) \in \mathbb{Q}[x]$. Obviously, the second property of the definition holds. But also, if $f(x) * (x - 1)$ and $g(x) * (x - 1)$ are two such polynomials, their sum is $(f(x) + g(x)) * (x - 1)$ and so belongs to I, too. Hence this set is an ideal; it is the ideal generated by $(x - 1)$.

The example above is typical. Let $\mathcal{F}$ be a collection of polynomials. Specifically, let $\mathcal{F} \subset \mathbb{Q}[\mathbf{x}]$ be a subset of the ring of polynomials in one or more indeterminates. The ideal *generated by* $\mathcal{F}$, denoted by $\langle \mathcal{F} \rangle$, is the set of all polynomial linear combinations of elements in $\mathcal{F}$,

$$\langle \mathcal{F} \rangle = \{h_1 f_1 + \cdots + h_n f_n : f_1, \ldots, f_n \in \mathcal{F},\ h_1, \ldots, h_n \in \mathbb{Q}[x]\}.$$

It is possible for two subsets $\mathcal{F}$ and $\mathcal{F}'$ to generate the same ideal,

$$\langle \mathcal{F} \rangle = \langle \mathcal{F}' \rangle.$$

In fact, by the Hilbert basis theorem, every ideal is finitely generated.

Theorem 1 (Hilbert basis theorem). *Every infinite set of polynomials $\mathcal{F}$ in $\mathbb{Q}[\mathbf{x}]$ has a finite subset $\mathcal{F}' \subset \mathcal{F}$ such that $\langle \mathcal{F} \rangle = \langle \mathcal{F}' \rangle$.*

The theorem says more than promised. To see that an ideal I is finitely generated, take $\mathcal{F}$ to be the ideal itself.

A Gröbner basis makes it easier to work with ideals.

It will make a difference in which order the terms of a polynomial are written. The monomial written first is its *leading term*. For a single indeterminate, we write the terms in order of higher to lower degree. For multivariate polynomials we use *lexicographic monomial order*, signified by $\succ$. This means that in order to decide the largest monomial in a set, we use the degree of x_1 without regard for any other indeterminate, unless more than one term has the same highest degree in x_1. In that case, we decide between these according to the highest degree of x_2, and so on. It is like the order of words in a dictionary. Nonzero coefficients are ignored. For example, among the monomials $-5x_1^3x_2^5x_3^4$ and $7x_1^2x_2^7x_3^4$, the former is larger in $\succ$-order. But among $-5x_1^3x_2^5x_3^4$ and $4x_1^3x_2^6$, the latter is larger in $\succ$-order.

We say that $\mathcal{G} = \{g_1, g_2, \ldots, g_r\}$ is a Gröbner basis for an ideal I if $\mathcal{G}$ generates I and if every polynomial f in I has its leading term divisible by the leading term of some g_i. MAPLE can be used to calculate Gröbner bases.

Example. In the polynomial ring $\mathbb{R}[x, y, z]$ (real coefficients allowed although this is not germane to the problem), let I be the ideal generated by $x^2 - y$ and $x^3 - z$. We will use lexicographic monomial order with $x \succ y \succ z$. This is communicated to MAPLE by the order used in listing the indeterminates and by the keyword `plex`:

MAPLE
```
> with(grobner):
> gbasis(x^2-y,x^3-z,[x,y,z],plex);
```

The result is

$$\{-y + x^2, -z + xy, -y^2 + xz, -z^2 + y^3\}$$

(each written in low to high order). Thus every polynomial in I must have its leading term divisible by x^2 or xy or xz or y^3. Note that the last three of these are in the

ideal although they were not among the original generators. For example, we have the following polynomial linear combination:

$$-z^2 + y^3 = (y^2 + yx^2 + x^4)(y - x^2) + (-z - x^3)(z - x^3).$$

Varieties are the zero sets of polynomials.

Let $f \in \mathbb{Q}[x_1, \ldots, x_d]$ be a multivariate polynomial with rational coefficients. The *variety* $\mathcal{V}(f)$ is the set of points $(z_1, \ldots, z_d)$ in d-dimensional complex space where f is zero,

$$\mathcal{V}(f) = \{(z_1, \ldots, z_d) \in \mathbb{C}^d : f(z_1, \ldots, z_d) = 0\},$$

in other words, the roots of f. Complex numbers are used here because polynomials are sure to have roots if complex numbers are allowed, but not if restricted to rational numbers or even real numbers. If S is a subset of $\mathbb{C}^d$, then define $\mathcal{V}_S(f) = \mathcal{V}(f) \cap S$, that is, the roots of f in S.

In applications of varieties studied in this chapter, the functions are often probability calculations and the solutions are expected to satisfy the requirement that they be nonnegative numbers summing to 1; for example,

$$p_1 \geq 0, \ldots, p_m \geq 0, \quad p_1 + \cdots + p_m = 1. \tag{15.3.2}$$

The set of points in m-dimensional space satisfying (15.3.2) is called an *m-dimensional simplex* (or just a *simplex* if the dimension is understood). A simplex in 3-space is a triangular portion of the plane lying in the first octant and passing through the three points (1, 0, 0), (0, 1, 0), and (0, 0, 1); see Figure 15.3.1. The figure was made using the following MAPLE commands:

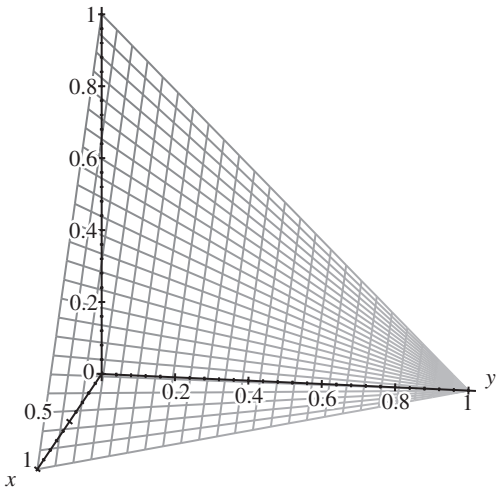


Fig. 15.3.1. Simplex in $\mathbb{R}^3$.

```
MAPLE
> with(plots):
> plot3d(1-x-y,0),x=0..1-y,y=0..1,axes=normal,scaling=constrained,orientation=[10,75]);
```

We will denote the m-dimensional simplex by the notation $\triangle^m$, or just $\triangle$ if m is understood. Note that the m-dimensional simplex is a *hypersurface* in m-space, that is, a surface of dimension 1 less than that of the space. Using the notation introduced above, $\mathcal{V}_\triangle(f)$ is the zeros of f that are meaningful as probabilities.

The variety of a polynomial is generally more than a finite set of points. For example, the variety of $f(z_1, z_2, z_3) = 4z_1z_2 - z_3^2$ is the *hypersurface* of $\mathbb{C}^3$ for which $z_3 = \sqrt{4z_1z_2}$. If $S = \mathbb{R}^3$, the subset of real numbers, then $\mathcal{V}_S(f)$ is the set of points in ordinary 3-space such that $x_3 = \sqrt{4x_1x_2}$ and is reminiscent of a saddle-surface defined only for $x_1 \geq 0$ and $x_2 \geq 0$ or $x_1 \leq 0$ and $x_2 \leq 0$.

Now let $\mathcal{F} \subset \mathbb{Q}[\mathbf{x}]$ be an arbitrary set of polynomials; $\mathbf{x}$ could be a vector of indeterminates, $\mathbf{x} = (x_1, \ldots, x_d)$. By the *variety* $\mathcal{V}(\mathcal{F})$ we mean the intersection of all hypersurfaces $\mathcal{V}(F)$ for all $F \in \mathcal{F}$. Put differently, $\mathcal{V}(\mathcal{F})$ is the set of all points $(z_1, \ldots, z_d) \in \mathbb{C}^d$ that are roots of all $f \in \mathcal{F}$,

$$f(z_1, \ldots, z_d) = 0 \quad \text{for all } f \in \mathcal{F}.$$

Let I be an ideal in $\mathbb{Q}[\mathbf{x}]$ and suppose that $I = \langle \mathcal{F} \rangle$. Then $\mathcal{V}(I) = \mathcal{V}(\mathcal{F})$ because every polynomial $g \in I$ can be written as a polynomial linear combination

$$g = h_1 f_1 + \cdots + h_r f_r, \quad f_1, \ldots, f_r \in \mathcal{F}.$$

So any point $\mathbf{z}$, a zero of $f_1, \ldots, f_r$, is also a zero of g.

The image of a polynomial map is also a variety.

The setting for our next result is that of a vector-valued function $\mathbf{f}(\mathbf{z})$ having m component functions, $f_1, \ldots, f_m$, each of which is defined for a d-dimensional variable $\mathbf{z} = (z_1, \ldots, z_d)$. This sort of setting occurs so often that a special notation is universally used for it; we write

$$\mathbf{f} : \mathbb{C}^d \longrightarrow \mathbb{C}^m.$$

The space $\mathbb{C}^d$ is called the *domain space* of $\mathbf{f}$, and the space $\mathbb{C}^m$ the *range space* or just the *range*.

Theorem 2 (implicitization). *Let* $\mathbf{f} : \mathbb{C}^d \longrightarrow \mathbb{C}^m$ *be a function whose components are multivariate polynomials in* $\mathbf{z}$. *Then the topological closure of the image of* $\mathbf{f}$ *is a variety in* $\mathbb{C}^m$.

The following example involves only polynomials of degree 1, *affine functions*, to simplify the calculations, but nonetheless captures the substance of the theorem.

Example. Define $\mathbf{f} : \mathbb{C}^2 \longrightarrow \mathbb{C}^3$ by

$$\begin{aligned} p_1 &= 2\theta_1 - 3\theta_2 + 1, \\ p_2 &= -\theta_1 + \theta_2 + 5, \end{aligned}$$

$$p_3 = \theta_1 + 2\theta_2.$$

The image is the set of all points

$$\begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \theta_1 \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} + \theta_2 \begin{bmatrix} -3 \\ 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 1 \\ 5 \\ 0 \end{bmatrix}$$

as θ_1 and θ_2 vary. This is a plane defined by two lines. The first has the direction of the vector $[2 \quad -1 \quad 1]^T$, and the second has the direction $[-3 \quad 1 \quad 2]^T$; both lines pass through the point $(1, 5, 0)$. To obtain this image as a variety, we eliminate the parameters θ_1 and θ_2. Add the second and third equations to get

$$3\theta_2 = p_2 + p_3 - 5.$$

Substitute this into the first two equations (multiplying the second by 3 avoids fractions):

$$\begin{aligned} p_1 &= 2\theta_1 - p_2 - p_3 + 6, \\ 3p_2 &= -3\theta_1 + p_2 + p_3 + 10. \end{aligned}$$

Now eliminate θ_1, and we have

$$3p_1 + 7p_2 + p_3 - 38 = 0.$$

Hence the image of $\mathbf{f}$ is the variety of $3p_1 + 7p_2 + p_3 - 38$. As already noted, this shows that the image is a hypersurface.

Note that the theorem does not say that the image itself is necessarily a variety, but that its topological closure is. In another example, consider the mapping $\mathbf{f} : \mathbb{C}^2 \to \mathbb{C}^2$ defined by $p_1 = z_1$ and $p_2 = z_1 z_2$. Name any point in the range $(\hat{p}_1, \hat{p}_2)$; the point $z_1 = \hat{p}_1$, $z_2 = \frac{\hat{p}_2}{\hat{p}_1}$ maps to it, that is, except when $\hat{p}_1 = 0$. So the image of $\mathbf{f}$ is the entire plane except for the p_2-axis itself, and even there the point $(0, 0)$ is in the image.

The *topological closure* of a set is the set itself together with its *boundary points*. In the example above, points on the p_2-axis are boundary points of the image.[2] Therefore, the closure of the image in this example is the entire (p_1, p_2)-plane. This is the variety of the zero polynomial (in (p_1, p_2)-space).

15.4 Algebraic Analysis of Maximum Likelihood

The philosophy of algebraic statistics is that statistical models are algebraic varieties. In this section, we show how the maximum likelihood problem can be cast in these

[2] This is because for any point on the p_2-axis, a sequence of points in the image leading to it can be found.

terms. The development serves as a prototype for other statistical problems that occur in biology.

The methodology goes as follows. Each derivative of the log-likelihood equation, e.g., (15.2.15), can be put into the form of a polynomial in the unknown parameters, θ. One is interested in knowing where these polynomials are simultaneously zero. As we have just seen in Section 15.3, the set of such zeros is called a variety. To analyze a variety, one tries to find the simplest system of polynomials that generate it. This is the purpose of a Gröbner basis. The resulting system of polynomials is easier to solve.

Note that the computations of this section are algebraic and often symbolic. So MATLAB will be able to do the calculations only if the symbolic package has been purchased. Although this package is from the people who produce MAPLE, the MAPLE code of this section has not been tested within MATLAB.

Interior nodes lead to a hidden Markov model.

The two-claw tree problem from before is too simple to illustrate the ideas and techniques of the algebraic method. Instead, we will analyze the rooted three-leaf tree of Figure 15.4.1.

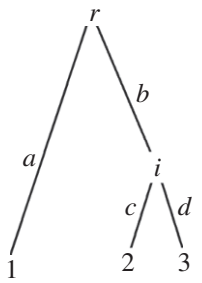


Fig. 15.4.1. Rooted three leaf tree with internal node.

It is important to note that this problem is fundamentally different from the two-claw problem because of the interior node i. We cannot know its state, only those of the leaf nodes. For this reason, interior nodes are said to be *hidden*, and the model is referred to as a *hidden Markov model*. Outcomes at the leaf nodes depend on the state of the hidden nodes, but the data will not be able to pin down those states directly; their values will have to be inferred. In calculating probabilities at the leaf nodes, allowance will have to be made for all possible states of the hidden nodes.

As before, we will assume Jukes–Cantor rates. Hence matrices a and b are defined in terms of their mutation rates α and β by (15.2.9); likewise, c and d are given by similar equations in terms of their rates γ and δ, respectively. And, as before, we will calculate probabilities in terms of the parameters a_0, a_1, b_0, b_1, and now c_0, c_1 and d_0, d_1.

Another difference between this problem and the two-claw problem is that there are now three leaf nodes. Observed outcomes here are triples of the nucleotides A, C, G, and T. Consequently, the number of possible outcomes is $4^3 = 64$, making for a vector of length 64,

$$p = [p_{\mathrm{AAA}} \;\; p_{\mathrm{AAC}} \;\; \cdots \;\; p_{\mathrm{TTT}}].$$

But by symmetries of the Jukes–Cantor model, many of the components are the same. We can see what they are by noting that the probabilities are invariant under any shuffling of the letters A, C, G, and T. That gives us the following equalities:

$$\begin{aligned}
p_{\mathrm{AAA}} &= p_{\mathrm{CCC}} = p_{\mathrm{GGG}} = p_{\mathrm{TTT}}, && 4 \text{ terms}, \\
p_{\mathrm{AAC}} &= p_{\mathrm{AAG}} = \cdots = p_{\mathrm{TTG}}, && 12 \text{ terms}, \\
p_{\mathrm{ACA}} &= p_{\mathrm{AGA}} = \cdots = p_{\mathrm{TGT}}, && 12 \text{ terms}, \\
p_{\mathrm{CAA}} &= p_{\mathrm{GAA}} = \cdots = p_{\mathrm{GTT}}, && 12 \text{ terms}, \\
P_{\mathrm{ACG}} &= p_{\mathrm{ACT}} = \cdots = p_{\mathrm{CGT}}, && 24 \text{ terms}.
\end{aligned} \tag{15.4.1}$$

Accounting for symmetries leaves only the five output probabilities shown in (15.4.1). They are p_{123} for all three nucleotides the same, p_{12} for only the first two the same; similarly define p_{13} and p_{23}. Finally, p_{dis} denotes the case in which all three are distinct.

We may now calculate the output probabilities. As noted above, p_{123} can occur in four ways; pick one, say AAA, compute it, and multiply by 4. As before, we assume that the root node could be A or C or G or T with equal probability, $\frac{1}{4}$; hence

$$p_{123} = \frac{4}{4}(a_0b_0c_0d_0 + 3a_0b_1c_1d_1 + 3a_1(b_1c_0d_0 + b_0c_1d_1 + 2b_1c_1d_1)). \tag{15.4.2$_1$}$$

In the same way, we calculate the others:

$$\begin{aligned}
p_{12} = \frac{12}{4}[&a_0(b_0c_0d_1 + b_1c_1d_0 + 2b_1c_1d_1) + a_1(b_0c_1d_0 + b_1c_0d_1 + 2b_1c_1d_1) \\
&+ 2a_1(b_1c_0d_1 + b_1c_1d_0 + b_0c_1d_1 + b_1c_1d_1)],
\end{aligned} \tag{15.4.2$_2$}$$

$$\begin{aligned}
p_{13} = \frac{12}{4}[&a_0(b_0c_1d_0 + b_1c_0d_1 + 2b_1c_1d_1) + a_1(b_1c_1d_0 + b_0c_0d_1 + 2b_1c_1d_1) \\
&+ 2a_1(b_1c_1d_0 + b_1c_0d_1 + b_0c_1d_1 + b_1c_1d_1)],
\end{aligned} \tag{15.4.2$_3$}$$

$$\begin{aligned}
p_{23} = \frac{12}{4}[&a_1(b_0c_0d_0 + 3b_1c_1d_1) + a_0(b_1c_0d_0 + b_0c_1d_1 + 2b_1c_1d_1) \\
&+ 2a_1(b_1c_0d_0 + 2b_1c_1d_1 + b_0c_1d_1)],
\end{aligned} \tag{15.4.2$_4$}$$

$$\begin{aligned}
p_{\mathrm{dis}} = \frac{24}{4}[&a_0(b_0c_1d_1 + b_1c_0d_1 + b_1c_1d_0 + b_1c_1d_1) \\
&+ a_1(2b_1c_1d_1 + b_0c_0d_1 + b_1c_1d_0) + a_1(2b_1c_1d_1 + b_1c_0d_1 + b_0c_1d_0) \\
&+ a_1(b_1c_1d_1 + b_1c_0d_1 + b_1c_1d_0 + b_0c_1d_1)].
\end{aligned} \tag{15.4.2$_5$}$$

Before continuing, we note that these equations can be significantly simplified by invoking the observation made at the end of the previous section, that because the states of the root are equally likely, we can determine only the product matrix ab and not a and b individually. Therefore, we should be able to simplify this equation by using (15.2.16) and defining the matrix $e = ab$; then

$$e_0 = a_0b_0 + 3a_1b_1, \qquad e_1 = a_0b_1 + a_1b_0 + 2a_1b_1. \tag{15.4.3}$$

Of course, the product ab is also stochastic, so its rows sum to 1:

$$e_0 + 3e_1 = 1. \tag{15.4.4}$$

The equivalent (unrooted) tree is shown in Figure 15.4.2.

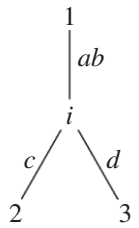


Fig. 15.4.2. Equivalent phylogenetic tree.

To incorporate (15.4.3) into (15.4.2), we let MAPLE do the work. Note that the new variables are in uppercase and we use the MAPLE command `algsubs` in place of `subs`.

Code 15.4.1.

MAPLE

```
> #first define the old variables
> p123:=(a0*b0*c0*d0+3*a0*b1*c1*d1+3*a1*(b1*c0*d0+b0*c1*d1+2*b1*c1*d1));
> p12:=(12/4)*(a0*(b0*c0*d1+b1*c1*d0+2*b1*c1*d1)+a1*(b0*c1*d0+b1*c0*d1+2*b1*c1*d1)
                +2*a1*(b1*c0*d1+b1*c1*d0+b0*c1*d1+b1*c1*d1));
> p13:=(12/4)*(a0*(b0*c1*d0+b1*c0*d1+2*b1*c1*d1)+a1*(b1*c1*d0+b0*c0*d1+2*b1*c1*d1)
                +2*a1*(b1*c1*d0+b1*c0*d1+b0*c1*d1+b1*c1*d1));
> p23:=(12/4)*(a1*(b0*c0*d0+3*b1*c1*d1)+a0*(b1*c0*d0+b0*c1*d1+2*b1*c1*d1)
                +2*a1*(b1*c0*d0+2*b1*c1*d1+b0*c1*d1));
> pdis:=(24/4)*(a0*(b0*c1*d1+b1*c0*d1+b1*c1*d0+b1*c1*d1)+a1*(2*b1*c1*d1+b0*c0*d1+b1*c1*d0)
                 +a1*(2*b1*c1*d1+b1*c0*d1+b0*c1*d0)+a1*(b1*c1*d1+b1*c0*d1+b1*c1*d0+b0*c1*d1));

> #now compute the new variables
> P123:=algsubs(a0*b0=e0-3*a1*b1,p123): P123:=expand(P123):
  P123:=algsubs(a0*b1=e1-a1*b0-2*a1*b1,P123);
> P12:=expand(p12): P12:=algsubs(a0*b0=e0-3*a1*b1,P12): P12:=expand(P12):
  P12:=algsubs(a0*b1=e1-a1*b0-2*a1*b1,P12);
> P13:=expand(p13): P13:=algsubs(a0*b0=e0-3*a1*b1,P13): P13:=expand(P13):
  P13:=algsubs(a0*b1=e1-a1*b0-2*a1*b1,P13);
> P23:=expand(p23): P23:=algsubs(a0*b0=e0-3*a1*b1,P23): P23:=expand(P23):
  P23:=algsubs(a0*b1=e1-a1*b0-2*a1*b1,P23);
> Pdis:=expand(pdis): Pdis:=algsubs(a0*b0=e0-3*a1*b1,Pdis): Pdis:=expand(Pdis):
  Pdis:=algsubs(a0*b1=e1-a1*b0-2*a1*b1,Pdis);
```

Alternatively, one may argue directly from Figure 15.4.2. Either way, the simplified probabilities are these:

$$\begin{aligned}
P_{123} &= e_0c_0d_0 + 3e_1c_1d_1,\\
P_{12} &= \frac{12}{4}(e_0c_0d_1 + e_1c_1d_0 + 2e_1c_1d_1),\\
P_{13} &= \frac{12}{4}(e_0c_1d_0 + e_1c_0d_1 + 2e_1c_1d_1),\\
P_{23} &= \frac{12}{4}(e_0c_1d_1 + e_1c_0d_0 + 2e_1c_1d_1),\\
P_{\text{dis}} &= \frac{24}{4}(e_0c_1d_1 + e_1c_0d_1 + e_1c_1d_0 + e_1c_1d_1).
\end{aligned}$$

On the surface it would appear that these five probabilities are functions of six variables, e_0, e_1, c_0, c_1, d_0, and d_1. But in reality, (15.4.4) holds between e_0 and e_1. Since matrices c and d are also stochastic, similar relationships hold for the cs and ds. These dependencies could be used to eliminate, say, e_0, c_0, and d_0 throughout,

$$e_0 = 1 - 3e_1, \qquad c_0 = 1 - 3c_1, \qquad d_0 = 1 - 3d_1. \tag{15.4.5}$$

With these substitutions, we see that the five probabilities are a function of a three-dimensional parameter vector θ,

$$\theta_1 = e_1, \qquad \theta_2 = c_1, \qquad \theta_3 = d_1. \tag{15.4.6}$$

But it is preferable to retain the homogeneous coordinates, that is, both es, both cs, and both ds, as long as possible and invoke (15.4.5) as the last step. When we refer to θ, it will be as if the substitutions (15.4.5) and (15.4.6) had been carried out.

Our development so far can be summarized in terms of a vector-valued function, $\mathbf{f}$, mapping the three-dimensional parameter space of θ into a five-dimensional outcome space with the five probabilities as its component functions. In the notation of the last section, we have $\mathbf{f} : \mathbb{C}^3 \to \mathbb{C}^5$, with component functions $f_1 = p_{123}$, $f_2 = p_{12}$, and so on,

$$\mathbf{f}(\theta) = (f_1,\ f_2,\ f_3,\ f_4,\ f_5) = (p_{123},\ p_{12},\ p_{13},\ p_{23},\ p_{\text{dis}}).$$

Notice that each component function $f_i(\theta)$ is linear as a function of the homogeneous coordinates and that these components sum to 1.

Now suppose we have three aligned DNA sequences, each n bases long, that correspond to the three leaves of the tree. Out of the n places, suppose that u_1 places match in all three sequences, u_2 places match in the first two sequences only, u_3 match in the first and third only, u_4 match in the second and third only, and in u_5 places all three sequences are different. Thus the vector $\mathbf{u} = [u_1\ u_2\ u_3\ u_4\ u_5]$ constitutes the

observed data. The likelihood function for this outcome is[3]

$$L(\theta) = \left(\frac{n!}{u_1!u_2!u_3!u_4!u_5!} \right) f_1^{u_1} f_2^{u_2} f_3^{u_3} f_4^{u_4} f_5^{u_5}.$$

To maximize this, take the logarithm of both sides, set its derivative with respect to each parameter to zero, and solve the resulting system,

$$\frac{\partial \log(L(\theta))}{\theta_1} = \frac{\partial \log(L(\theta))}{\theta_2} = \frac{\partial \log(L(\theta))}{\theta_3} = 0,$$

where, for $i = 1, 2, 3$,

$$\frac{\partial \log(L(\theta))}{\partial \theta_i} = \frac{u_1}{f_1}\frac{\partial f_1}{\partial \theta_i} + \frac{u_2}{f_2}\frac{\partial f_2}{\partial \theta_i} + \frac{u_3}{f_3}\frac{\partial f_3}{\partial \theta_i} + \frac{u_4}{f_4}\frac{\partial f_4}{\partial \theta_i} + \frac{u_5}{f_5}\frac{\partial f_5}{\partial \theta_i}. \tag{15.4.7}$$

The likelihood variety.

Recall that each function f_i is multilinear as a function of the homogeneous coordinates. It follows that by combining the terms of (15.4.7) with a common denominator, the result is a ratio of polynomials in the homogeneous coordinates or in the θ_i as well. For example, invoking (15.4.5) and (15.4.6), we have

$$\begin{aligned} f_1 &= (1 - 3e_1)(1 - 3c_1)(1 - 3d_1) + 3e_1c_1d_1 \\ &= (1 - 3\theta_1)(1 - 3\theta_2)(1 - 3\theta_3) + 3\theta_1\theta_2\theta_3 \end{aligned}$$

and

$$\frac{\partial f_1}{\partial \theta_1} = -3(1 - 3\theta_2)(1 - 3\theta_3) + 3\theta_2\theta_3.$$

Similar results hold for the other components.

The *critical points* of the problem are the points θ in three-dimensional space where the functions (15.4.7) vanish, that is, equal zero, but the denominators of these equations are not zero. Then to be a solution to our problem, a critical point $\hat{\theta}$ must also be a vector of probabilities, that is, each component must lie between 0 and 1, the 3-simplex.

Discounting, temporarily, points where the denominators are zero, the set of critical points is a variety in 3-space called the *likelihood variety*. The maximum likelihood solution we want, the solution in terms of θ, is the computation of this variety.

[3] The point of showing this equation is to note the relationship between the components f_i and the data u_i. How does the ratio of factorials come about? The argument is the same as our derivation of the combinations factor in Section 2.8. For example, there are u_1 places in the three DNA sequences where the bases match, say, all are As. One first imagines that these As are different, say, $A_1, \ldots, A_{u_1}$. This contributes to the $n!$ in the numerator, but too much so because these As are, in fact, not distinct. Since there are $u_1!$ ways to rearrange the As, dividing by it corrects the overcount. Argue similarly for the other factors.

In the following MAPLE code, we enter the component probabilities in terms of `t1`, `t2`, `t3` representing θ_1, θ_2, and θ_3. MAPLE calculates the derivatives and combines terms over a common denominator to form the three numerators `p1`, `p2`, `p3`. The denominator is not needed for the critical points of the likelihood derivatives, but is needed for the *Hessian* (see next) and to check that a root is admissible. We also compute the Hessian, or the second derivative matrix, in order to check the nature of a critical point. If the eigenvalues of the Hessian are all negative, then the critical point is a maximum point of the surface. If the eigenvalues are all positive, then the critical point is a minimum. If the eigenvalues are of mixed sign, then the critical point is a saddle-point and corresponds to a saddle-surface in parameter space there. (Like the flat spot on a saddle, this means that there are nearby points of the surface both higher and lower than at the flat spot.) The numerical root finder, `fsolve`, is used to solve the system but, owing to the multiplicity of roots, is shown to need additional help.

In the following, assume that the observed data are $u_1 = 31$, $u_2 = 5$, $u_3 = 7$, $u_4 = 11$, $u_5 = 13$.

Code 15.4.2.

```
  MAPLE
> f1:=(1-3*t1)*(1-3*t2)*(1-3*t3)+3*t1*t2*t3;
> f2:=3*(1-3*t1)*(1-3*t2)*t3+3*t1*t2*(1-3*t3)+6*t1*t2*t3;
> f3:=3*(1-3*t1)*t2*(1-3*t3)+3*t1*(1-3*t2)*t3+6*t1*t2*t3;
> f4:=3*(1-3*t1)*t2*t3+3*t1*(1-3*t2)*(1-3*t3)+6*t1*t2*t3;
> f5:=6*(1-3*t1)*t2*t3+6*t1*(1-3*t2)*t3+6*t1*t2*(1-3*t3)+6*t1*t2*t3;
> cden:=f1*f2*f3*f4*f5;
  #fix t3 to reduce computation time, then p3 won't be needed
> t3:=1/10;
> u1:=31: u2:=5: u3:=7: u4:=11: u5:=13:
> p1:=u1*diff(f1,t1)*cden/f1 + u2*diff(f2,t1)*cden/f2 + u3*diff(f3,t1)*cden/f3 + u4*diff(f4,t1)*cden/f4
                             + u5*diff(f5,t1)*cden/f5;
> p2:=u1*diff(f1,t2)*cden/f1 + u2*diff(f2,t2)*cden/f2 + u3*diff(f3,t2)*cden/f3 + u4*diff(f4,t2)*cden/f4
                             + u5*diff(f5,t2)*cden/f5;
  #the next for the Hessian matrix
> h11:=cden*diff(p1,t1)-p1*diff(cden,t1);
> h12:=cden*diff(p1,t2)-p1*diff(cden,t2);
> h21:=cden*diff(p2,t1)-p2*diff(cden,t1);
> h22:=cden*diff(p2,t2)-p2*diff(cden,t2);
> simplify(h12-h21); #check that the mixed partials are equal
  #use fsolve to find a root
> S1:=fsolve({p1,p2},{t1,t2},{t1=0..1,t2=0..1});
  #see if this is a root of cden, check size of cden against p1 and p2
> assign(S1); p1; p2; cden;
  #if this is a root of cden, try another
> t1:='t1'; t2:='t2'; #reset t1 and t2
> S2:=fsolve({p1,p2},{t1,t2},{t1=0..1,t2=0..1},avoid={S1});
  #and avoid={S1,S2} if another round necessary, etc.,
  #also check if the eigenvalues of the Hessian are negative
> assign(S2);
> h:=array([[h11,h12],[h21,h22]]);
> evalf(Eigenvals(h));
```

As the above shows, MAPLE's numerical root finder might find a root that is also a root of the denominator. Then it is necessary to search for another. By contrast, an algebraic root finder proceeds in a very different way. The public domain computer algebra system *SINGULAR* is specialized to deal with these kinds of problems. (SINGULAR is obtainable free of charge from the website www.singular.uni-kl.de.)

By casting the roots as the variety of an ideal, a more suitable basis, a Gröbner basis, may be used in place of the original polynomials. Furthermore, there is a clever way to avoid roots of the common denominator.

Let z_i play the role of $\frac{1}{f_i}$; then $z_i f_i - 1 = 0$. Adjoin the z_i to our ring and work in the space $\mathbb{Q}[\theta_1, \ldots, \theta_d, z_1, \ldots, z_m]$. Let J_u be the ideal generated by the maximum likelihood polynomials and these reciprocal relations for the z_i,

$$J_u = \left\langle z_1 f_1 - 1, \ldots, z_m f_m - 1, \sum_{j=1}^{m} u_j z_j \frac{\partial f_j}{\partial \theta_1}, \ldots, \sum_{j=1}^{m} u_j z_j \frac{\partial f_j}{\partial \theta_d} \right\rangle.$$

A point $(\theta, z) \in \mathbb{C}^{d+m}$ belongs to the variety $\mathcal{V}(J_u)$ of J_u if θ is a root of the maximum likelihood equations and if $f_i(\theta) z_i = 1$, for all i; therefore, $f_i(\theta) \neq 0$. Since we are not interested in the zs, only the θs, put

$$I_u = J_u \cap \mathbb{Q}[\theta_1, \ldots, \theta_d]$$

to eliminate the zs; I_u is the likelihood ideal. Here is the SINGULAR code for this problem.

Code 15.4.3.

```
  SINGULAR
> ring bigring = 0, (t1,t2,z1,z2,z3,z4,z5), dp; number t3 = 1/10;
> poly f1 = (1-3*t1)*(1-3*t2)*(1-3*t3)+3*t1*t2*t3;
> poly f2 = 3*(1-3*t1)*(1-3*t2)*t3+3*t1*t2*(1-3*t3)+6*t1*t2*t3;
> poly f3 = 3*(1-3*t1)*t2*(1-3*t3)+3*t1*(1-3*t2)*t3+6*t1*t2*t3;
> poly f4 = 3*(1-3*t1)*t2*t3+3*t1*(1-3*t2)*(1-3*t3)+6*t1*t2*t3;
> poly f5 = 6*(1-3*t1)*t2*t3+6*t1*(1-3*t2)*t3+6*t1*t2*(1-3*t3)+6*t1*t2*t3;

> int u1=31; int u2=5; int u3=7; int u4=11; int u5=13;
> ideal Ju = z1*f1-1, z2*f2-1, z3*f3-1, z4*f4-1, z5*f5-1,
    u1*z1*diff(f1,t1)+u2*z2*diff(f2,t1)+u3*z3*diff(f3,t1) +u4*z4*diff(f4,t1)+u5*z5*diff(f5,t1),
    u1*z1*diff(f1,t2)+u2*z2*diff(f2,t2)+u3*z3*diff(f3,t2) +u4*z4*diff(f4,t2)+u5*z5*diff(f5,t2);
> ideal Iu = eliminate(Ju,z1*z2*z3*z4*z5);
> ring smallring = 0, (t1,t2), dp;
> ideal Iu = fetch(bigring,Iu); Iu;
  // dim(G)=dimension of G, vdim(G)= #roots if dim(G)=0
> ideal G = groebner(Iu); dim(G); vdim(G);
  // 20 digits of precision
> ideal G = groebner(Iu); LIB "solve.lib"; solve(G,20);
```

Of the 16 solutions, only three are in the range $0 < \theta_1, \theta_2 < 1$. And only one of those has a negative definite Hessian (making it a maximizing point), as shown in the MAPLE calculation above.

15.5 Characterizing Trees by Their Variety, Phylogenetic Invariants

We now view the problem of the last section in a completely different way. Instead of studying the problem of maximizing the log-likelihood function from the perspective of parameter space θ, one can analyze it from the standpoint of the range space of probabilities. In our three-leaf problem, this is the five-dimensional space of the points

$(p_{123}, p_{12}, p_{13}, p_{23}, p_{\text{dis}})$. In this section, we no longer regard these as probabilities but merely as points $(p_1, p_2, p_3, p_4, p_5)$ in 5-space. In this analysis, we would like to characterize the *image* of $\mathbf{f}$,

$$\text{image}(\mathbf{f}) = \{\mathbf{p} = (p_1, p_2, \ldots, p_5) : \mathbf{p} = \mathbf{f}(\theta) \text{ for some } \theta\}.$$

The rationale is that different tree structures will give rise to images in $\mathbb{C}^5$ having different surface structures no matter what the values of the parameters may be. We seek to characterize these surface structures. Our best hope for this is to regard them as varieties and compute their generating polynomials. Polynomials that are zero on the image of $\mathbf{f}$ are called *phylogenetic invariants*.

Phylogenetic invariants characterize the tree without having to solve it.

As we saw in the algebra section, the closure of the image of $\mathbf{f}$ is an algebraic variety. Let $I_{\mathbf{f}}$ denote the ideal of polynomials in $p_1, \ldots, p_m$ that vanish on this variety. If h is one such polynomial, then

$$h(p_1, \ldots, p_m) = 0, \quad \text{where } \mathbf{p} = \mathbf{f}(\theta), \quad \theta \in \mathbb{C}^d.$$

The problem of finding generating polynomials for $I_{\mathbf{f}}$ is the problem of implicitization. As we saw earlier, it amounts to eliminating the θs from the component functions f_i. We illustrate the method for the three-leaf tree of the previous section, (15.4.2).

One begins by transforming the equations to a simpler form as prescribed by *Fourier analysis*. The theory underlying the *Fourier transformation* is well known, but its study is beyond the scope of this text. The transformation for this problem and for all other small trees is given at the *small trees website*,

http://www.math.tamu.edu/∼lgp/small-trees.

In this example, the transformed coordinates are simple products of the factors $(e_0 - e_1)$ and $(e_0 + 3e_1)$ and the same in c and d.

In fact, it can be verified that

$$q_{111} = p_{123} - \frac{1}{3}p_{12} - \frac{1}{3}p_{13} - \frac{1}{3}p_{23} + \frac{1}{3}p_{\text{dis}} = (e_0 - e_1)(c_0 - c_1)(d_0 - d_1). \tag{15.5.1$_5$}$$

Denote this combination by q_{111}. Similarly, it can be verified that

$$q_{000} = p_{123} + p_{12} + p_{13} + p_{23} + p_{\text{dis}} = (e_0 + 3e_1)(c_0 + 3c_1)(d_0 + 3d_1), \tag{15.5.1$_1$}$$

$$q_{011} = p_{123} - \frac{1}{3}p_{12} - \frac{1}{3}p_{13} + p_{23} - \frac{1}{3}p_{\text{dis}} = (e_0 + 3e_1)(c_0 - c_1)(d_0 - d_1), \tag{15.5.1$_2$}$$

$$q_{101} = p_{123} - \frac{1}{3}p_{12} + p_{13} - \frac{1}{3}p_{23} - \frac{1}{3}p_{\text{dis}} = (e_0 - e_1)(c_0 + 3c_1)(d_0 - d_1), \tag{15.5.1$_3$}$$

$$q_{110} = p_{123} + p_{12} - \frac{1}{3}p_{13} - \frac{1}{3}p_{23} - \frac{1}{3}p_{\text{dis}} = (e_0 - e_1)(c_0 - c_1)(d_0 + 3d_1). \tag{15.5.1$_4$}$$

The coordinates $q_{000}, q_{011}, \ldots, q_{111}$ are called the *Fourier coordinates*, and the equations (15.5.1) constitute the *Fourier transform*. They have been indexed by the subgraphs of the tree; q_{000} corresponds to the empty tree, q_{111} to the full tree, q_{011} to the span of leaves 2 and 3, and so on. An excluded edge corresponds to a factor such as $(e_0 + 3e_1)$, an included edge to a factor such as $(e_0 - e_1)$.

With this simplification we work to eliminate the parameters. First, eliminate $e_0 + 3e_1$; from (15.5.1$_1$) and (15.5.1$_2$),

$$\frac{q_{000}}{(c_0 + 3c_1)(d_0 + 3d_1)} = e_0 + 3e_1 = \frac{q_{011}}{(c_0 - c_1)(d_0 - d_1)}.$$

Next eliminate $c_0 + 3c_1$. Solve the first and third members of this for $c_0 + 3c_1$ and use (15.5.1$_3$) to get

$$\frac{q_{000}(c_0 - c_1)(d_0 - d_1)}{q_{011}(d_0 + 3d_1)} = c_0 + 3c_1 = \frac{q_{101}}{(e_0 - e_1)(d_0 - d_1)}.$$

Use the first and third members of this to solve for d_0+3d_1 and combine with (15.5.1$_4$),

$$\frac{q_{000}(c_0 - c_1)(d_0 - d_1)^2(e_0 - e_1)}{q_{011}q_{101}} = d_0 + 3d_1 = \frac{q_{110}}{(e_0 - e_1)(c_0 - c_1)}.$$

And finally incorporate (15.5.1$_5$) into this; we get

$$q_{000}q_{111}^2 = q_{011}q_{101}q_{110}.$$

Therefore, the ideal $I_{\mathbf{f}}$ is generated by the homogeneous third-degree polynomial

$$I_{\mathbf{f}} = \langle q_{000}q_{111}^2 - q_{011}q_{101}q_{110} \rangle.$$

The molecular clock assumption yields a different ideal.

Recall from Section 15.2 the molecular clock assumption: For any subtree and each path from the root of that subtree to any leaf, the products of the transition matrices corresponding to the edges of the path are identical. As applied to the three-taxa problem we have been following, Figure 15.4.1, this means that

$$c = d, \qquad a = bc = bd.$$

Or, in terms of the individual parameters,

$$\begin{aligned} d_0 &= c_0, & d_1 &= c_1, \\ a_0 &= b_0c_0 + 3b_1c_1, & a_1 &= b_0c_1 + b_1c_0 + 2b_1c_1. \end{aligned}$$

In calculating the five probabilities, the assumption that a and b cannot be distinguished is no longer valid (since their edges are different distances from the root).

There are now only two independent matrices, b and c, and four homogeneous parameters or two independent parameters. Also note that p_{12} and p_{13} will be the same, because $c = d$, and the one chosen must be counted twice. Hence the data space is only four-dimensional here. Starting from (15.4.2), we use MAPLE for the calculation. Denote the output probabilities by p_1, p_2, p_3, and p_4:

```
MAPLE
#recall that p123,p12,...were defined in Code 15.4.1
> p1:=subs({d0=c0,d1=c1,a0=b0*c0+3*b1*c1,a1=b0*c1+b1*c0+2*b1*c1},p123): p1:=simplify(p1);
> p2:=subs({d0=c0,d1=c1,a0=b0*c0+3*b1*c1,a1=b0*c1+b1*c0+2*b1*c1},p12):
> p2:=2*simplify(p2); #for emphasis
> p3:=subs({d0=c0,d1=c1,a0=b0*c0+3*b1*c1,a1=b0*c1+b1*c0+2*b1*c1},p23): p3:=simplify(p3);
> p4:=subs({d0=c0,d1=c1,a0=b0*c0+3*b1*c1,a1=b0*c1+b1*c0+2*b1*c1},pdis): p4:=simplify(p4);
```

This gives

$$
\begin{aligned}
p_1 &= b_0^2c_0^3 + 3b_0^2c_1^3 + 6b_0b_1c_0^2c_1 + 6b_0b_1c_0c_1^2 + 12b_0b_1c_1^3 + 3b_1^2c_0^3 \\
&\quad + 6b_1^2c_0^2c_1 + 6b_1^2c_0c_1^2 + 21b_1^2c_1^3, \\
p_2 &= 6b_0^2c_0^2c_1 + 12b_0c_0^2b_1c_1 + 84b_1c_1^2b_0c_0 + 102b_1^2c_0c_1^2 + 84b_1^2c_1^3 \\
&\quad + 6b_0^2c_1^2c_0 + 48b_0c_1^3b_1 + 30b_1^2c_1c_0^2 + 12b_0^2c_1^3, \\
p_3 &= 3b_0^2c_0^2c_1 + 42b_0c_1^3b_1 + 6b_1c_0^3b_0 + 21b_1^2c_0c_1^2 + 12b_0c_0^2b_1c_1 + 60b_1^2c_1^3 \\
&\quad + 3b_0^2c_1^2c_0 + 12b_1c_1^2b_0c_0 + 21b_1^2c_1c_0^2 + 6b_0^2c_1^3 + 6b_1^2c_0^3, \\
p_4 &= 24b_0c_0^2b_1c_1 + 18b_0^2c_1^2c_0 + 60b_1c_1^2b_0c_0 + 114b_1^2c_0c_1^2 + 60b_0c_1^3b_1 \\
&\quad + 78b_1^2c_1^3 + 24b_1^2c_1c_0^2 + 6b_0^2c_1^3.
\end{aligned}
$$

From the small trees website, we can look up the Fourier transformation for this problem; the Fourier coordinates are

$$
\begin{aligned}
q_{0000} &= p_1 + p_2 + p_3 + p_4 = (b_0 + 3b_1)^2(c_0 + 3c_1)^3, \\
q_{0011} &= p_1 - \frac{1}{3}p_2 + p_3 - \frac{1}{3}p_4 = (c_0 + 3c_1)(b_0 + 3b_1)^2(c_0 - c_1)^2, \\
q_{0111} &= p_1 + \frac{1}{3}p_2 - \frac{1}{3}p_3 - \frac{1}{3}p_4 = (c_0 + 3c_1)(b_0 - b_1)^2(c_0 - c_1)^2, \\
q_{1111} &= p_1 - \frac{1}{3}p_2 - \frac{1}{3}p_3 + \frac{1}{3}p_4 = (b_0 - b_1)^2(c_0 - c_1)^3.
\end{aligned}
$$

As before, the Fourier coordinates are indexed according to the portions of the subtree included and excluded. It is easy to check that $q_{0011}q_{0111}^2 = q_{0000}q_{1111}^2$; therefore, the ideal is generated by

$$
I_{\mathbf{f}} = \langle q_{0011}q_{0111}^2 - q_{0000}q_{1111}^2 \rangle.
$$

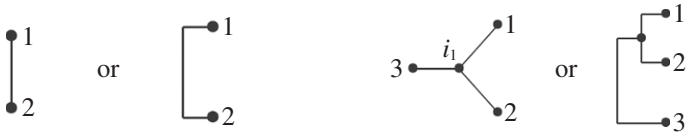
15.6 Constructing the Phylogenetic Tree

Previously, we have seen how to compute branch lengths for an existing phylogenetic tree using genomic alignments. In this section, we take up the study of how to

construct the tree itself. For this we need a matrix of pairwise distances, called a *dissimilarity map*, between the taxa of the tree. The map can be the result of a multiple alignment between genomes. For example, it could be calculated by the public domain software, MAVID, written for this purpose.

The number of possible trees for n taxa is exponential in n.

In Figure 15.6.1(a), we show a two-leaf unrooted tree. Also shown is an equivalent form more commonly seen in phylogenetic studies. The tree may be described by the notation (1, 2), indicating the leaves of the tree and their connectivity. This tree has $n = 2$ leaves, no interior nodes, and one edge. There is only one such tree.

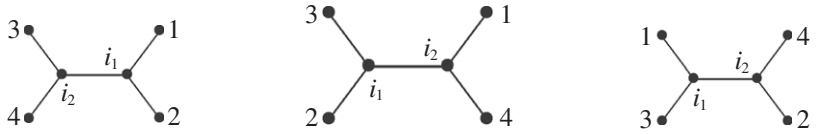


(a) (1, 2) (b) (1, 2, 3)

Fig. 15.6.1.

By adding a new leaf, attaching it to the single existing edge, we obtain the $n = 3$ leaved tree of Figure 15.6.1(b). The notation (1, 2, 3) indicates that the leaves share a common interior node, i_1. By adding the new leaf, we have also added an interior node. In general, one new interior node will be added for each new leaf, and so inductively, the number of interior nodes is given by $n - 2$, equal to 1 here. It is also seen that by dividing an edge for the leaf addition, we have created two new edges. In general, the number of new edges created by the addition of a new leaf is two, and so inductively, the number of edges of a tree is given by $2n - 3$, equal to 3 here. The number of such three-leaved trees is one.

Once again we add a new leaf. This time there are three edges at which to make the attachment, giving rise to three different tree structures. The possibilities are shown in Figure 15.6.2. In the first case, the leaf is added to the edge between the interior node and leaf 3, in the second case to the edge between the interior node and leaf 1, and in the third case to the edge between the interior node and leaf 2. Each has $n - 2 = 2$ interior nodes and $2n - 3 = 5$ edges. As noted, for this $n = 4$ case there are three different trees.



(a) ((1, 2), (4, 3)) (b) ((1, 4), (3, 2)) (c) ((1, 3), (2, 4))

Fig. 15.6.2.

In general, moving from the n to the $n+1$ case, the number of different possibilities is equal to the number of existing edges. Hence the number of tree structures increases according to

$$1 \cdot 3 \cdot 5 \cdot 7 \cdots (2n-5) = (2n-5)!!.$$

The right-hand side of this equation defines a notation for the left-hand side. These relationships between the number of leaves, interior nodes, edges, and tree structures hold for unrooted phylogenetic trees having three edges adjacent at each interior node.[4]

On the basis of this analysis, we see that the number of trees that have to be searched in calculating a phylogenetic tree grows very rapidly with n; for $n = 6$, it is 105 trees, but for $n = 10$ the number is 2,027,025. As a result, maximum likelihood is usually infeasible for tree construction. Instead, methods are available utilizing branch lengths, or more generally tree metrics.

Distance functions build trees two leaves at a time.

A *dissimilarity map*, d, on the first n integers, denoted by $[n] = \{1, 2, \ldots, n\}$, is a symmetric nonnegative-valued function satisfying

$$d(i, j) = d(j, i) \geq 0, \qquad d(i, i) = 0.$$

The matrix of a dissimilarity map is the $n \times n$ matrix D whose (i, j)th element is d_{ij}.

A dissimilarity map is a *metric* on $[n]$ if it satisfies the *triangle inequality*,

$$d(i, j) \leq d(i, k) + d(k, j), \quad i, j, k \in [n].$$

A dissimilarity map d on $[n]$ is a *tree metric* if there is a tree T with n leaves and a nonnegative length for each edge such that for every pair of leaves i and j, the length of the unique path from i to j equals $d(i, j)$. An example is given in Figure 15.6.3 with corresponding distances presented in (15.6.4). For a tree metric, fix two leaves i and j and let k be some leaf. It is easy to see that $d(i, j) \leq d(i, k) + d(k, j)$ by considering the subtree spanned by i and k. So a tree metric is a metric.

A *cherry* of a tree is a pair of leaves both adjacent to the same node, their common ancestor. Let (x, y) be a cherry of a tree, let v be their common ancestor, and let k be any other leaf. Then[5]

$$\begin{aligned} d(v, k) &= \frac{1}{2}(d(v, k) + d(v, k)) \\ &= \frac{1}{2}(d(k, v) + d(v, x) + d(k, v) + d(v, y) - (d(x, v) + d(v, y))) \\ &= \frac{1}{2}(d(x, k) + d(y, k) - d(x, y)). \end{aligned} \tag{15.6.1}$$

[4] Since a rooted tree can be created from an unrooted tree by attaching the root to any edge, the number of rooted binary trees on n leaves is $(2n-3)!!$.

[5] Use equality here because v is on the unique path from k to either x or y.

Theorem 1 (Saitou and Nei). *Let d be a tree metric on $[n]$. For every pair $i, j \in [n]$, put*

$$Q_d(i, j) = (n-2)d(i, j) - \sum_{k \neq i} d(i, k) - \sum_{k \neq j} d(j, k). \tag{15.6.2}$$

The pair $x, y \in [n]$ that minimizes $Q_d(i, j)$ is a cherry of the tree. Note that $Q_d(i, j)$ will be negative if d is a metric, since $d(i, j) \leq d(i, k) + d(k, j)$.

Given a dissimilarity map on n taxa, one can invoke the theorem to identify the two taxa, x and y, most related, i.e., most likely to be a cherry of a phylogenetic tree on these taxa. Let v denote their common node. To continue the construction, we use (15.6.1) to define the distance from v to the other leaves of the tree. The construction is now continued as if v were a leaf of the reduced set. We are led to the *neighbor-joining algorithm.*

Neighbor-joining algorithm.

Step 1. Compute the $\binom{n}{2}$ values $Q_d(i, j)$ of (15.6.2); let x and y give the minimum. Add x, y, and their common node v to the tree.

Step 2. Remove x and y from the list $[n]$, but add v to the list. Extend the dissimilarity map to v by defining

$$d(v, k) = \frac{1}{2}(d(x, k) + d(y, k) - d(x, y)) \tag{15.6.1}$$

for all remaining leaves k in the list.

Step 3. If the reduced list is of length 3 or more, return to Step 1; otherwise, join the last two with a common edge, completing the tree.

A tree metric, d_T, can be created for the tree recursively. Add Step 1.5 between Steps 1 and 2 as follows:

Step 1.5 Pick an arbitrary element of the list, r, different from x and y, and define

$$\begin{aligned} d_T(x, v) &= \frac{1}{2}(d(x, y) + d(x, r) - d(y, r)), \\ d_T(y, v) &= d((x, y) - d_T(x, v). \end{aligned} \tag{15.6.3}$$

If d is already a tree metric for some tree T, then the algorithm will find it, and the metric d_T constructed in Step 1.5 will be again d. Otherwise, we hope that d_T will be close to d.

Example. To illustrate the algorithm, let a dissimilarity map be given by the matrix

$$D = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} - & 10 & 11 & 14 \\ 10 & - & 3 & 12 \\ 11 & 3 & - & 13 \\ 14 & 12 & 13 & - \end{pmatrix} \end{array}. \tag{15.6.4}$$

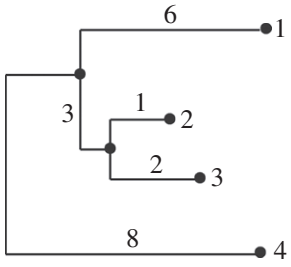


Fig. 15.6.3.

Since this derives from a tree metric (see Figure 15.6.3), the algorithm should regenerate the tree.

First, calculate the Q_d matrix using (15.6.2). For example,

$$Q_d(1,2) = (4-2)*10 - (10+11+14) + (10+3+12) = -40.$$

In full, the matrix is

$$Q_d = \begin{bmatrix} - & -40 & -40 & -46 \\ -40 & - & -46 & -40 \\ -40 & -46 & - & -40 \\ -46 & -40 & -40 & - \end{bmatrix}.$$

The minimal value is -46 in two places, $1,4$ and $2,3$. This means that we may choose either; the resulting graphs will be equivalent. Selecting 1 and 4, join them by an internal node, i_2 say, and remove them from the list, Step 2. In their place, add i_2 to the list with distances constructed using (15.6.1),

$$d(i_2,2) = \frac{1}{2}(d(2,1) + d(2,4) - d(1,4)) = 4,$$
$$d(i_2,3) = \frac{1}{2}(d(3,1) + d(3,4) - d(1,4)) = 5.$$

To figure the tree distances $d_T(1,i_2)$ and $d_T(4,i_2)$, select as "root" $r = 2$. Then from (15.6.3),

$$d_T(1,i_2) = \frac{1}{2}(d(1,4) + d(1,2) - d(2,4)) = 6,$$
$$d_T(4,i_2) = d(1,4) - d_T(1,i_2) = 8.$$

At this point the construction is as shown in Figure 15.6.4(a), and the new dissimilarity distances figured above are noted in D',

$$D' = \begin{array}{c} \\ 2 \\ 3 \\ i_2 \end{array} \begin{array}{c} \begin{array}{ccc} 2 & 3 & i_2 \end{array} \\ \begin{pmatrix} - & 3 & 4 \\ 3 & - & 5 \\ 4 & 5 & - \end{pmatrix} \end{array}.$$

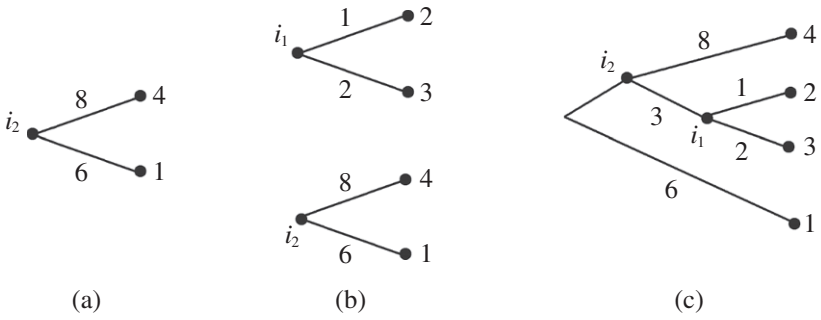


Fig. 15.6.4. (a)–(c) Stages of the tree construction.

The next round continues with $\{2, 3, i_2\}$. Invoking (15.6.2), now with $n = 3$, gives -12 for all terms; e.g.,

$$Q_d(2, 3) = 1 \cdot 3 - (3 + 4) - (3 + 5) = -12.$$

In fact, at this final stage, each term of Q_d equals the negative of the sum of the distances, $-(3 + 4 + 5) = -12$. Therefore, any pair may be selected, and each choice gives an equivalent graph. So we choose 2 and 3 and join them through a common vertex i_1. The tree distances are calculated via (15.6.3) using, say, $r = 1$ as root,

$$d_T(2, i_1) = \frac{1}{2}(d(2, 3) + d(1, 2) - d(3, 1)) = 1,$$
$$d_T(3, i_1) = d(2, 3) - d_T(2, i_1) = 2.$$

This gives the tree of Figure 15.6.4(b).

What remains is i_1 and i_2; their distance is figured using (15.6.1),

$$d(i_1, i_2) = \frac{1}{2}(d(2, i_2) + d(3, i_2) - d(2, 3)) = 3.$$

These final two elements are simply joined, finishing the tree, Figure 15.6.4(c).

Exercises/Experiments

1. What is the branch length between the HSPs (high scoring pairs) of the *Latimeria chalumnae* Hoxa-11 gene (AF287139) and the *Polyodon spathula* Hoxa-11 gene (AY661748.1) as calculated by (15.2.8)? Use blastn to get the "identities" (match percent) between these DNA segments.
2. What is the substitution matrix $a = e^{Q_{\mathrm{K80}}}$ for the Kimora-80 rate matrix? Notice that there are three distinct terms in a. Labeling these a_0, a_1, and a_2, work out the probabilities p_{same} and p_{dif} for the two-claw problem.
3. Under the Jukes–Cantor model, work out the probabilities p_{same} and p_{dif} for the two-claw problem with an interior node on one edge; see Figure 15.6.5(a).

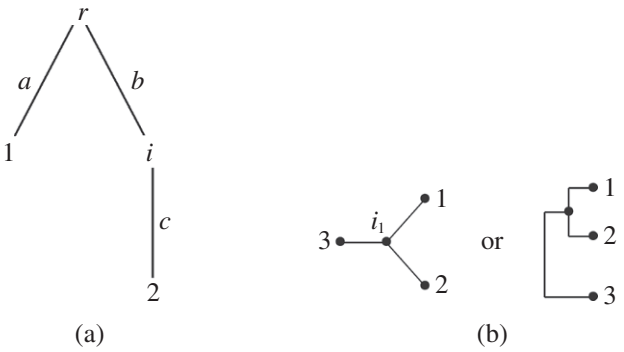


Fig. 15.6.5.

4. In the text, the maximum likelihood solution for the three-leaved tree with internal node (Figure 15.4.1 or, equivalently, Figure 15.6.5(b)) was worked out. The solution for the observed data, $u = [31\ \ 5\ \ 7\ \ 11\ \ 13]$, can be computed from the MAPLE program on p. 526 or the SINGULAR program on p. 527. Find the solution to obtain $\theta_1 = e_1$, $\theta_2 = c_1$, and $\theta_3 = d_1$. From these find the branch length from the internal node i to 2; recall from (15.2.6) that

$$\text{branchlen}_2 = 3\gamma t, \qquad \text{and} \quad c_1 = \frac{1}{4}(1 - e^{-4\gamma t}),$$

the branch length from i to 3; similarly,

$$\text{branchlen}_3 = 3\delta t, \qquad \text{and} \quad d_1 = \frac{1}{4}(1 - e^{-4\delta t}),$$

and the branch length from 1 to i,

$$\text{branchlen}_1 = 3(\alpha + \beta)t, \quad \text{and} \quad e_0 = \frac{1}{4}(1 + 3e^{-4(\alpha+\beta)t}).$$

5. Assume that the following matrix defines a dissimilarity function among the proteins: 1 = At1g20880, 2 = Hs20556011, 3 = CE13934, 4 = Hs14192947, and 5 = At5g53680. Construct the phylogenetic tree of these proteins based on this dissimilarity function. Compare with KOG0149 of the Clusters of Orthogolous Groups at NCBI,

$$\begin{bmatrix} - & .96 & .46 & .54 & .38 \\ .96 & - & .64 & .55 & .43 \\ .46 & .64 & - & .33 & .33 \\ .54 & .55 & .33 & - & .39 \\ .38 & .43 & .33 & .39 & \end{bmatrix}.$$

Questions for Thought and Discussion

1. Discuss the problem of speciation: How do new species arise? How does a new organism find a mate (speciation is defined in terms of mating)? How are new species confirmed (how can you say that a species is new)? What problems are there in verification?
2. What evidence do fossils and the fossil record provide in helping to fix phylogenetic trees?

References and Suggested Further Reading

[1] H. B. Stenzel, Successional speciation in paleontology: The case of the oysters of the sellaeformis stock, *Evolution*, **3** (1949), 33–50.
[2] J. Felsenstein, *Inferring Phylogenies*, Sinauer, Sunderland, MA, 2004.
[3] N. Eldredge and J. Cracraft, *Phylogenetic Patterns and the Evolutionary Process*, Columbia University Press, New York, 1980.
[4] R. D. M. Page and E. C. Holmes, *Molecular Evolution: A Phylogenetic Approach*, Blackwell Science, Oxford, UK, 1998.
[5] L. Pachter and B. Sturmfels, *Algebraic Statistics for Computational Biology*, Cambridge University Press, Cambridge, UK, 2005.
[6] C. Zimmer, What is a species?, *Sci. Amer.*, **298**-6 (2008), 72–79.