Chapter 13

Gene trees and species trees

After twelve chapters discussing trees, models, and methods of inference, its time (or past time) to bring up the issue of exactly what the trees we are inferring represent.

When DNA or other sequences are collected, they come from individuals of a taxon, and not the taxon as a whole. Moreover, the sequence is not the entire genome of the individual, but more typically the sequence of one gene. Thus the trees we find should represent the evolutionary relationships of these individual genes or loci, and are better called *gene trees.* For reasons we will discuss in more detail in this chapter, these trees need not represent the evolutionary relationships of the full taxa, or even of the full individual. It is possible that they represent the relationships of the taxa on a *species tree*, and under some circumstances it is even likely. However, it is also quite likely that the gene trees and species trees will be in conflict.

The chapter develops the framework for modeling a primary source of conflict between gene trees and species trees, *incomplete lineage sorting.* While the distinction between gene trees and species trees has been understood from the early days of phylogenetics, only recently have serious attempts been undertaken to develop statistical tools to deal with it directly. For many years, inferred gene trees were simply accepted as likely proxies for species trees, with more careful scientists acknowledging this distinction. As it has become cheaper and easier to sequence many genes in a collection of taxa, it has become less easy to ignore the discordance of gene trees among each other.

The model used to capture incomplete lineage sorting is the *multispecies coalescent.* It modifies Kingman's basic coalescent model of population genetics so that several populations are linked to form a tree. Although this modeling framework is now fairly well established, it is not yet clear what methods of species tree inference inspired by it will ultimately prove most useful. Thus any enthusiasm or criticism we offer of particular approaches should be taken lightly; more progress can be expected in the next few years.

13.1 Gene Lineages in Populations

The primary reason we should not expect gene trees to always match species trees is shown in Figure 13.1. Here a tree with wide pipe-like branches represents the history of the species in their descent from a common ancestor, where the width of a branch is meant to convey that a species is actually a population of individuals of some size. The thin trees within it represent a gene tree, of the sort we might construct by standard phylogenetic methods, that relates some particular sample of sequences collected from individuals in those populations. Because of the many individuals present in the species at any time (represented by the width of the species tree branches), it is possible that several gene lineages may exist within a branch, without merging together. This then makes it possible for the lineages to eventually merge in some earlier population on the species tree in a way that gives a gene tree topology that differs from the species tree topology. Multiple gene lineages persisting in this way, not merging within a single branch of the species tree, is referred to as *incomplete lineage sorting.*

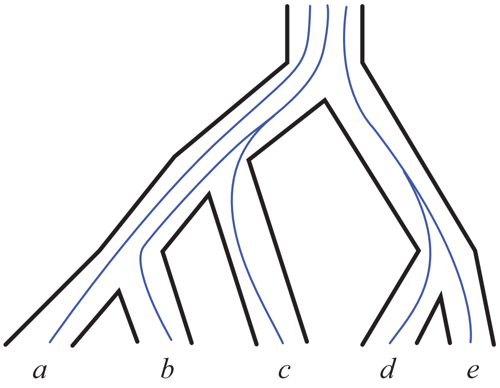

Figure 13.1: Gene trees may differ from species trees, since species trees are built from populations, and multiple gene lineages may persist through a species tree population (edge) and then merge with other lineages in a way that conflicts with the species tree topology. Here the species tree has topology $(((a, b), c), (d, e))$, while the gene tree might have topology $((A, (B, C)), (D, E))$, or several others, depending on how the lineages coalesce "above the root" of the species tree.

To model this phenomenon, we first step back from considering a full species tree, and instead consider the simpler situation of gene lineages in a single population.

The *Wright-Fisher* model imagines that there are some number N of individuals in the population at all times, with time tracked in discrete steps, corresponding to generations. In case of a haploid or diploid organism, there are thus either N or $2N$ copies of a specific gene present at each time step. We depict these as in Figure 13.2, with each row representing a generation, and each dot representing a gene.

To create gene lineages, we begin at generation 0 (the present) and imagine each gene picks a parent gene uniformly at random from the previous generation. Thus from a gene in the current population at the bottom of the figure we can draw a random lineage, back one generation at a time, through the ancestral generations. (There are some obvious idealizations in this model: By picking parents at random, we assume the population is panmictic, we ignore sex and the grouping of two genes in one diploid organism, and we assume neutrality under selection.) An example of a simulation from this process is shown in Figure 13.2.

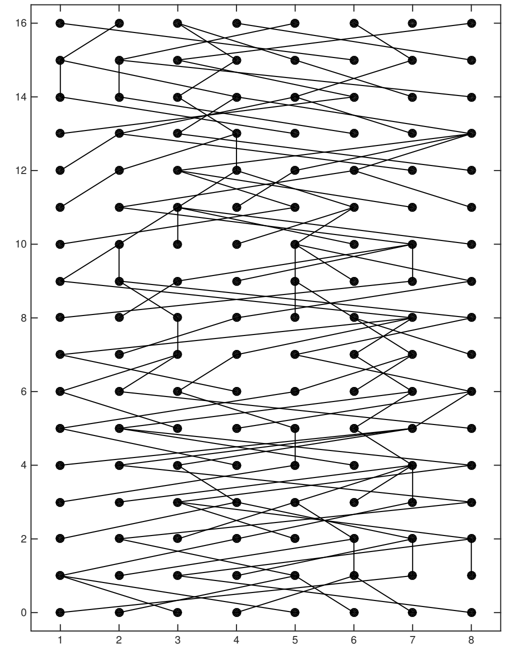

Figure 13.2: A simulation from the Wright-Fisher model, with 8 genes. A gene lineage is produced by a current gene (at bottom) choosing a random parent in the previous generation, which then chooses its parent, etc.

Figure 13.2 is too much of a tangle to interpret easily. However, by appropriate sorting of genes in each generation to prevent lineages from crossing, and suppressing lineages with no descendants in generation 0, we obtain the more understandable Figure 13.3. Working backwards in time, we see that gene lineages merge, or *coalesce* with some regularity. At each generation, there may be coalescent events which reduce the number of lineages coming from the extant genes. After moving enough generations into the past, all the lineages from the extant genes will have merged into a single one.

Viewing Figure 13.3 in the other direction, from the past to the present, we see many genes will have no progeny in the present. Indeed, if going backwards in time all lineages from the present have merged into a single one, then all genes not on that lineage will have no current progeny.

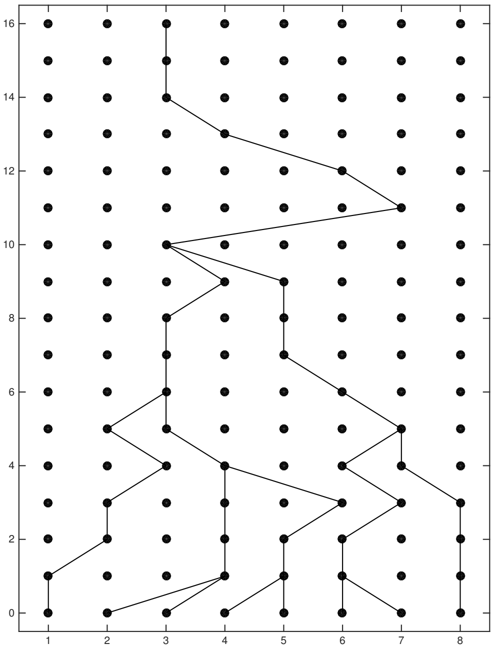

Figure 13.3: A Wright-Fisher simulation, after sorting of genes to untangle lineages, and retaining only lineages with extant descendants.

It is relatively easy to compute a few probabilities associated to this model. In the haploid case with N genes per generation, the probability that any two will choose the same parent, and thus have their lineages coalesce immediately, is $1/N$. This is because no matter what parent the first gene chooses, the second must choose the same one of the N possibilities to produce an immediate coalescence.

The probability that two lineages do not coalesce in the parental generation is therefore $1 - 1/N$. The same reasoning as before then gives the probability they will coalesce in the previous generation is $1/N$. Extending this reasoning shows that, using $C_2 = n$ to mean the event that two specific extant lineages coalesce n generations before the present,

$$\begin{aligned}
\mathcal{P}(C_2 = 1) &= \frac{1}{N} \\
\mathcal{P}(C_2 = 2) &= \left(1 - \frac{1}{N}\right)\frac{1}{N} \\
\mathcal{P}(C_2 = 3) &= \left(1 - \frac{1}{N}\right)^2 \frac{1}{N} \\
&\vdots \\
\mathcal{P}(C_2 = n) &= \left(1 - \frac{1}{N}\right)^{n-1} \frac{1}{N}
\end{aligned}$$

The probability of coalescence by generation n is thus

$$\mathcal{P}(C_2 \leq n) = \sum_{i=1}^{n} \mathcal{P}(C_2 = i) = \frac{1}{N}\sum_{i=1}^{n}\left(1 - \frac{1}{N}\right)^{i-1} = 1 - \left(1 - \frac{1}{N}\right)^{n} \tag{13.1}$$

As the number of generations n grows to infinity, we see the probability of coalescence approaches 1.

We can also compute the expected number of generations to coalescence of two specific lineages:

$$\sum_{n=1}^{\infty} n\mathcal{P}(C_2 = n) = \frac{1}{N}\sum_{n=1}^{\infty} n\left(1 - \frac{1}{N}\right)^{n-1} = N \tag{13.2}$$

This is plausible, as the larger the population size, the longer it should be before coalescence occurs, on average.

If we are interested in more than 2 lineages coalescing, things become more complicated. For instance for 3 lineages, there are likely to be two coalescent events needed for 3 lineages to merge down to 1 (though with low probability all three merge at once). We would have to consider the two generations in which these events occurred, and the expected final coalescence time would involve a double summation. Although quantities such as this can be worked out exactly, the formulas become rather complicated.

13.2 The Coalescent Model

Kingman's coalescent model can be viewed as a continuous-time approximation of the Wright-Fisher model (and also of other discrete models of population genetics, including the Moran model.) Rather than fully develop it from the Wright-Fisher model, we will instead simply define it. At an informal level the connection between them should seem reasonable. (See [Wak09] for an excellent full treatment.)

The model describes the coalescence of lineages as we move backwards in time within a single population. Time will be denoted by u, from the present with $u = 0$ into the past with $u > 0$, and is measured in *coalescent units*. We will relate coalescent units to more familiar quantities later, but for now they are simply some measure of time. The entire model is this:

> Given pairs of lineages at any fixed time, the rate at which pairs coalesce into single lineages is constant, and equal to 1. Simultaneous coalescence of more than two lineages does not occur. Coalescence of different pairs is independent, and identically distributed.

Two small points are in order here. First, by *rate* we mean something very similar to what was meant in the discussion of the GTR model of base substitution. A rate of a continuous-time probabilistic process determines, through some calculation, a probability at any time, and so we will be able to compute probabilities of coalescences from it. Second, to make this rate 1 we simply rescale time units in a way that is convenient, just as we were able to freely rescale time with the GTR. As we will see, the resulting rescaled time measured in coalescent units need not be proportional to real time, except perhaps for very short periods.

Consider two lineages which are distinct at time 0, and $u > 0$, let $h(u)$ denote the probability that the two lineages are distinct at time u (that is, the two did not coalesce between time 0 and time u). Then the model tells us

$$\frac{d}{du}h(u) = -1 \cdot h(u),$$

where the negative sign is due to the fact that $h(u)$ should decrease.[1] Since we additionally know $h(0) = 1$, we find

$$h(u) = e^{-u}.$$

Thus if $\mathcal{P}(u)$ denotes the probability the two lineages did coalesce between time u_0 and u, we have

$$\mathcal{P}(u) = 1 - e^{-u}.$$

This is the analogue of equation (13.1) of the Wright-Fisher model, and graphing the two formulas shows they display quite similar behavior.

We can also compute the expected time to coalescence of two lineages, to obtain an analog of equation (13.2). Since the probability of coalescence in a short time interval is $\mathcal{P}(u + \Delta u) - \mathcal{P}(u) \approx \mathcal{P}'(u)\Delta u$, the expected time is

$$\int_0^\infty u\mathcal{P}'(u)du = \int_0^\infty ue^{-u}\,du = 1. \tag{13.3}$$

That this expected time to coalescence does not depend on the size of the population should seem surprising. However, as defined here the coalescent

[1] More formally, we are assuming coalescent events occur as a Poisson process, a standard probabalistic model for events that occur rarely, but with equal chance in any small interval of a fixed size.

model ignores the population size — it is never even referred to in the definition of the model. We have simply *defined* our time scale so the rate of coalescence is one

In fact the population size does matter, but it is taken care of by the definition of coalescent units. To see why this is reasonable, return to the Wright-Fisher model. Imagine that the population size changed as we move backwards in time, forming a *bottleneck*, as in Figure 13.4. If the large population below the bottleneck is N_1, coalescence of pairs of lineages occurs with probability $1/N_1$ in each generation. When lineages enter the bottleneck, the same formula applies, but with a smaller population size N_2, the probability is now larger, $1/N_2$. Thus coalescence becomes more likely to occur. This means that if we had no access to the generational time scale, and could only query whether coalescence had occurred, the bottleneck of a relatively small number of generations of size N_2 would be indistinguishable from a larger number of generations where the population had remained constant at size N_1.

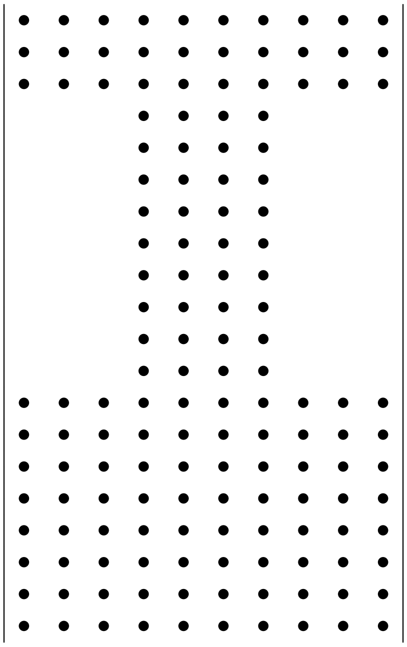

Figure 13.4: A bottleneck in the Wright-Fisher model, with population size $N_1 = 10$ except for $m_2 = 9$ generations with population $N_1 = 4$. The bottleneck causes faster coalescence of lineages, producing similar behavior to a longer time span with no bottleneck.

Reasoning roughly with the Wright-Fisher model, since the expected number of generations until two lineages coalesce is equal to the population size, N_2 generations in the small population has the same impact on coalescence as N_1 generations in the large one. If we introduce a new time scale for each population, where we use

$$\Delta u = \frac{\Delta t}{N_i}$$

in a population of size N_i, where Δt is a number of generations, then N_1 generations in the large population and N_2 generations in the small population both yield $\Delta u = 1$ Thus scaling time inversely by population size enables us to treat the coalescence of lineages as proceeding at a constant rate.

For the discrete Wright-Fisher model, this change of time scales only approximately creates a uniform rate; after all, we do not have fractional generation time. However, in the coalescent model this becomes the definition of a coalescent unit. Since the coalescent is a continuous-time model, we can define units in term of 'infinitesimal' increments:

$$du = \frac{1}{N(t)} dt. \tag{13.4}$$

If the population size $N(t) = N$ is constant, then equation (13.4) integrates to give

$$u = \frac{t}{N}.$$

With this assumption, we can thus convert the expected coalescent time of 1 coalescent unit in equation (13.3) to N generations, which is exactly in accord with the Wright-Fisher result. However, equation (13.4) is more general, and allows changing population sizes to lead to non-linear relationships between u and t.

Since coalescent units are used as the time scale in formulating the model, the population size is doesn't explicitly appear in any calculations. However, the population does have an effect whenever we relate coalescent models to true time. A large population at some time means the coalescent clock 'runs slow' with respect to true time, so it takes more true time for coalescent events to occur. A small population means the coalescent clock 'runs fast' with respect to true time, so coalescent events occur more rapidly. More generally, as population size changes, the coalescent clock may be constantly changing its speed with respect to true time.

There is, of course, a price to pay for this relationship. It will be impossible to separate out the individual contributions of time and population size from their combined effect, unless we are willing to make some strong assumptions. While we might wish this was just an artifact of this model that we could do away with in some way, by thinking about the Wright-Fisher model it should become clear it would be a feature of any reasonable model we might formulate.

To demonstrate another calculation with the standard coalescent model, consider the expected time to coalescence of n lineages down to one. For all n lineages to coalesce, first 2 must coalesce so only $n-1$ lineages remain. Then 2 of these must coalesce so only $n-2$ remain, and so on, until the last 2 coalesce.

When all n lineages are present at time $u_0 = 0$, there are many pairs that might coalesce. It is thus reasonable that the first coalescent event will occur sooner than if only two lineages were present. For a more precise calculation of the expected time of coalescence from n to $n-1$ lineages, recall that the coalescence of different pairs is i.i.d., so the overall rate of coalescence is increased by a factor of the number of pairs, $\binom{n}{2} = n(n-1)/2$. Thus with $k(u)$ being the probability that k lineages remain distinct at time $u > 0$, we have

$$\frac{d}{du}k(u) = -\frac{n(n-1)}{2}k(u),$$

with $k(0) = 1$. Thus

$$k(u) = e^{-\left(\frac{n(n-1)}{2}\right)u}.$$

Proceeding similarly to the calculation of the expected coalescent time for 2 lineages, we find that the expected time for n lineages to coalesce to $n-1$ is (see Exercise 6)

$$\frac{n(n-1)}{2}\int_0^\infty ue^{-\left(\frac{n(n-1)}{2}\right)u}\,du = \frac{2}{n(n-1)}. \tag{13.5}$$

Thus while the expected time for 2 lineages to coalesce to 1 is 1 unit, the time for 3 to coalesce to 2 is only 1/3 unit, the time for 4 to coalesce to 3 is 1/6 unit, etc. Adding these, we obtain the expected time for n lineages to coalesce to 1 is

$$\begin{aligned}\sum_{i=2}^{n}\frac{2}{i(i-1)} &= 2\sum_{i=2}^{n}\left(\frac{1}{i-1}-\frac{1}{i}\right)\\ &= 2\left(\left(1-\frac{1}{2}\right)+\left(\frac{1}{2}-\frac{1}{3}\right)+\left(\frac{1}{3}-\frac{1}{4}\right)+\cdots+\left(\frac{1}{n-1}-\frac{1}{n}\right)\right)\\ &= 2\left(1-\frac{1}{n}\right)\end{aligned}$$

As n approaches ∞, this grows, but the limit of the expected time until all lineages coalesce is 2, which is only twice that of when $n = 2$.

These calculations indicate that in a typical coalescent tree formed by a large number of lineages coalescing we should expect to see a lot of coalescence near the leaves of the tree, and longer edge lengths near the root. Roughly half the tree will have only two lineages which coalesce at the root, one third will have 3 lineages, etc. These characteristics are depicted in the tree shown in Figure 13.5.

Note that for a sexually reproducing diploid organism, each individual has two copies of most genes. Both the Wright-Fisher model, and the standard coalescent model ignore the fact that 2 gene lineages reside in each individual, and that individuals have sexes. However, since these copies come from different parents, their lineages are distinct, and in a panmictic population should have histories that are independent of one another.

If two different unlinked genes are considered in such an organism, even if they are sampled from the same individuals, there should also be little relationship between the gene trees for the two. Since the number of ancestors

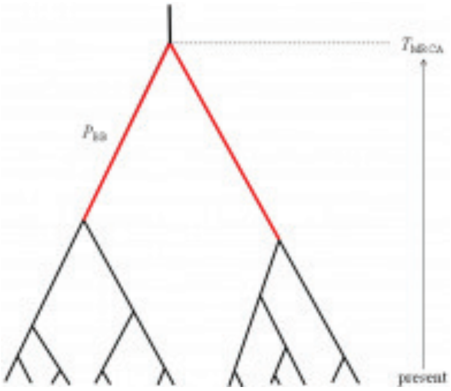

Figure 13.5: A typical tree produced by the coalescent process in a single population. Edge lengths are measured in coalescent units with the time during which there are k lineages being, on average, $2/((k)(k-1))$.

n generations in the past grows exponentially by the formula 2^n (provided, of course, the population is sufficiently large), going back even a few generations, the lineages of different genes are likely to pass through different individuals. Once this happens, under a panmictic assumption coalescence with different lineages should then be independent between the two genes. If genes are linked, then there lineages will not be independent and a more complicated model is needed to capture how lineages coalesce.

Finally the version of the coalescent presented here is not appropriate for all organisms. In some species of fish, for instance, the number of offspring of a successful breeder can be quite large. In that case, then a model must allow simultaneous coalescence of more than 2 lineages at a time. Such modifications result in a model called the Λ-coalescent.

13.3 Coalescent Gene Tree Probabilities

Suppose we sample the same gene in several individuals within a population and assume the coalescent model describes the probabilistic way their lineages coalesce. If the coalescent process is continued until all the lineages have become one, then a rooted gene tree is formed. Each possible gene tree can arise with some probability, which can be computed. In this section, we demonstrate how this can be done in the simplest situation — the gene samples are taken from a single population, which persists back in time 'forever'. (In essence, we assume a species tree with only a single taxon.) Gene tree probabilities can be computed for either topological gene trees or metric gene trees, so we give examples of both.

3-sample trees

First we consider sampling 3 extant lineages, which we will denote by A_1, A_2, A_3, at $u = 0$.

If we are only concerned with gene tree topologies, then we observe that the rooted gene tree which relates them is determined by the first pair of lineages to coalesce. Since the coalescence of pairs is i.i.d., this first coalescence involves A_1, A_2 or A_1, A_3, or A_2, A_3 with equal probability 1/3. Thus each of the 3 gene trees $((A_1, A_2), A_3)$, $((A_1, A_3), A_2)$, and $((A_2, A_3), A_1)$ arises with probability 1/3.

For a metric gene tree probability, note that distances will be given in coalescent units, and the tree must be ultrametric. Consider the gene tree $((A_1{:}u_1, A_2{:}u_1){:}u_2, A_3{:}u_3)$, where $u_3 = u_1 + u_2$. This is formed by

(a) 3 lineages and no coalescent events for a time interval of length u_1,

(b) a coalescent event of 3 lineages to 2 at time u_1, with the specific lineages A_1, A_2 coalescing,

(c) 2 lineages and no coalescent event for a time interval of length u_2,

(d) a coalescent event of 2 lineages to 1 at time $u_3 = u_2 + u_1$.

Now from previous calculations (a) has probability e^{-3u_1}, and (c) has probability e^{-u_2}. Both (b) and (d) require probabilities of coalescence at an instant, which is simply the rate of coalescence times du. Thus (d) has probability $du_3 = du_2$, while (b) involves both the probability of a coalescence, $3du_1$, and an additional factor of 1/3 that the lineages involved in the coalescence are A_1, A_2. The total probability is thus

$$\mathcal{P}(u_1, u_2) = e^{-(3u_1+u_2)} du_1 du_2,$$

so the probability density function is

$$f(u_1, u_2) = e^{-(3u_1+u_2)}. \tag{13.6}$$

When this is integrated over $0 \le u_1, u_2 \le \infty$, it gives a different computation of the probability 1/3 of the topological tree $((A_1, A_2), A_3)$. (See Exercise 7.)

Larger trees

If 4 extant genes A_1, A_2, A_3, A_4 are sampled at $u = 0$, the computing probabilities of topological trees is a little more complicated. A caterpillar gene tree like $(((A_1, A_2), A_3), A_4)$ can only be formed by a specific sequence of coalescent events. That the first lineages to merge are A_1, A_2 has probability $1/\binom{4}{2} = 1/6$. Once there are only 3 lineages, that the A_1A_2 lineage merges with the A_3 lineage next has probability $1/\binom{3}{2} = 1/3$. Then that the $A_1A_1A_3$ lineage merges with A_4 has probability 1. Thus the probability of this, or any of the other caterpillar gene trees is $(1/6)(1/3) = 1/18$.

A balanced tree like $((A_1, A_2), (A_3, A_4))$ can be formed by either of two sequences of coalescent events: A_1A_2 may merge first, followed by A_3A_4, or *vice versa*. Each of these has probability 1/18 since they are determined by a specific sequence of coalescent events. Thus the total probability of this, or any other balanced gene tree is $2(1/18) = 1/9$.

This last calculation shows a significant feature of the coalescent model in a single population. Topological gene trees that show more 'balance' tend to have higher probability than those that are less balanced, because they can be achieved by more distinct orderings of coalescent events. In fact, it is not hard to generalize the calculation of topological gene tree probabilities from 4-samples to more. Define a *ranked gene tree* as a rooted binary leaf-labelled topological tree with an ordering to the internal nodes (from the leaves to the root) such that the ranking of any node is greater than all its descendants. Then under the coalescent all ranked gene trees are equally probable. Since there are

$$R(n) = \prod_{k=2}^{n} \binom{k}{2} = \frac{n!(n-1)!}{2^{n-1}} \tag{13.7}$$

ranked gene trees (see Exercise 10), the probability of any gene tree is simply the number of rankings it may be given divided by $R(n)$.[2]

For a metric gene tree, the edge lengths determine an ordering to the coalescent events. For instance, the gene tree $((A_1{:}u_1, A_2{:}u_1){:}u_3, (A_3{:}u_2, A_4{:}u_2){:}u_4)$ is formed by coalescent events occurring at times u_1, u_2, and $u_1 + u_3 = u_2 + u_4$. If $u_1 < u_2$, then the first cherry formed was (A_1, A_2), while if $u_2 < u_1$ it was (A_3, A_4). For either case, computing the probability is very similar to the 3-sample case above (see Exercise 12). Unlike the topological tree case, there is no extra care that needs to be taken, since only oner ranking can arise for a metric tree.

13.4 The Multispecies Coalescent Model

The multispecies coalescent extends the basic coalescent model of the last section to a species tree of populations.

In terms of true time, we can picture the species tree as an ultrametric tree, with each branch represented by a pipe as in Figure 13.1. However, since we will be measuring time in coalescent units, which are only related to true time through inverse scaling by the population size, the species tree need not be ultrametric in these units. Moreover, using coalescent units means we have in some sense standardized the widths of the pipes to all be the same, so that standard Newick notation can be used to specify a species tree.

[2] The probabilities obtained for rooted topological trees here is the same as is produced by the Yule model, a model of branching that proceeds from the root toward the leaves, and is often taken as the simplest probabilistic speciation model. One could argue that it is the most natural distribution of rooted trees in biological contexts, and is perhaps a better choice of a "non-informative" prior for a Bayesian tree inference.

The multispecies coalescent model is simply the standard coalescent model on each edge of the species tree 'glued' together. But now, since species tree edges have finite length, we may have several gene lineages in a population an edge represents that fail to coalesce within that edge. When they reach the ancestral end of the edge, one or more additional lineages will enter from another branch of the species tree. If we are computing probabilities of metric gene trees, then we specify precisely where each coalescent event occurs, and though the bookkeeping can be rather cumbersome, we merely have to combine various probabilities of lineages coalescing, or not coalescing, in a single edge of a specific length, in ways very similar to the single population coalescent. For topological gene trees, the work is similar, but we first need to compute probabilities than if k lineages 'enter' an edge of length x, that $0 \leq \ell < k$ coalescent events will occur on that edge. (A future version of these notes will derive that....)

3-sample trees

We begin with a simple, yet very important example, of a 3-taxon species tree $((a{:}y, b{:}z){:}x, c{:}w)$, where we sample one gene from each taxon. We will denote the sampled genes by A, B, C, using uppercase letters corresponding to the taxon names. Before we consider a specific gene tree, note that y, z, w will have no effect on the probability of observing any *topological* gene tree. This is because under the one-sample-per-taxon scheme, as shown in Figure 13.6, there will be only one lineage in each of the pendant species tree branches, and so no opportunity for coalescence in these populations.

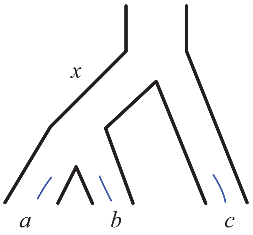

Figure 13.6: With one gene sampled per taxon, no coalescence can occur in pendant populations on the species tree, so the lengths of those edges are irrelevant to the probability of observing any topological gene tree. For the species tree $((a{:}y, b{:}z){:}x, c{:}w)$, only x will appear in probability formulas.

There are 3 possible topological gene trees: $((A, B), C)$, $((A, C), B)$, and $((B, C), A)$. To compute the probabilities of observing them, it is easiest to begin with $((A, C), B)$. The only way this gene tree can be formed is if the A and B lineages enter the population with length x and then reach the root of the tree before merging. That results in 3 lineages being present at the root, and then the A and C lineages must coalesce in the population ancestral to the root.

For the previous section we know the probability of two lineages not coalescing in x coalescent units is e^{-x}. The probability that the correct 2 lineages of the 3 then coalesce first is 1/3. Thus

$$\mathcal{P}(\,((A,C),B)) = \frac{1}{3}e^{-x}.$$

The same reasoning shows

$$\mathcal{P}(\,((B,C),A)) = \frac{1}{3}e^{-x}.$$

Since the probabilities of the 3 topological gene trees must add to 1, this also means

$$\mathcal{P}(\,((A,B),C)) = 1 - \frac{2}{3}e^{-x}.$$

Notice that if $x = 0$, so the species tree has a polytomy at the root, all of these become 1/3, as is reasonable. If $x = \infty$, on the other hand, the A and B lineages must coalesce in the infinitely long branch, and so we find $\mathcal{P}(\,((A,B),C)) = 1$, with the other gene trees having probability 0. But for $0 < x < \infty$, we have

$$0 < \mathcal{P}(\,((A,C),B)) = \mathcal{P}(\,((B,C),A)) < \mathcal{P}(\,((A,B),C)).$$

Thus in this case the most probable gene tree has the same topology as the species tree, while the two discordant gene trees have smaller probabilities.

This leads to a simple method of species tree inference in the 3-taxon, 1-sample-per-taxon case: after inferring many gene trees by standard phylogenetic methods, simply tally the number of gene trees with each topology. Then accept the most frequent one as the species tree topology. But we can actually infer even more. After estimating $\frac{2}{3}e^{-x}$ as the proportion of gene trees that are discordant with the inferred species tree, we can solve for an estimate of x. Thus the gene tree distribution not only contains information on the topology of the true species tree, but also on the lengths of its edges.

The simplicity of the 3-taxon case, though, is misleading. For n-taxa, the most probable gene tree need *not* match the topology of the species tree. This fact means the natural intuition that whatever gene tree is inferred the most often should be reflect the specie tree can be misleading.

If we instead needed to compute the probability of a metric gene tree (with edge lengths in coalescent units) we proceed more similarly to the example discussed in the previous section. But now the edge lengths in the gene tree determine in which populations on the species tree the various coalescent events occurred. For example, for the species tree $((a{:}y, b{:}z){:}x, c{:}w)$to calculate the gene tree probability $\mathcal{P}(((A{:}u_1, B{:}u_2){:}u_3, C{:}u_4))$, we must separate the cases where A and B coalesce in the species tree branch of length x, and when they coalesce in the population ancestral to the root of the species tree. Moreover, for the gene tree to have non-zero probability, we must have

$$u_1 > y,\ u_2 > z,\ u_4 > w,\ u_1 - y = u_2 - z,\ u_1 + u_3 > y + x,\ u_1 + u_3 - y - x = u_4 - w.$$

Assuming these conditions are met

$$\begin{aligned}\mathcal{P}(((A{:}u_1, B{:}u_2){:}u_3, C, {:}u_4)) &= \mathcal{P}(u_1, u_3)\\ &= \begin{cases} e^{-(u_1-y)-(u_1+u_3-x-y)}, & \text{if } u_1 < y+x,\\ e^{-x-3(u_1-x-y)-u_3}, & \text{if } u_1 > y+x. \end{cases}\end{aligned} \tag{13.8}$$

The formulas depend only on u_1 and u_3 since those values determine u_2 and u_4 by the restrictions above.

Larger trees

To illustrate how probabilities are computed for larger trees, we give only one (partial) example. For the species tree is $(((a{:}u_1, b{:}u_2){:}u_3, c{:}u_4){:}u_5, d{:}u_6)$, consider the topological gene tree $(((A, B), C), D)$ which has the same topology as the species tree. Since there is only one ranking of the coalescent events forming this gene tree, there are relatively few ways it could form. They are:

1. (A, B) coalesces in the branch of length u_3, $((A, B), C)$ coalesce in the branch of length u_5, and $(((A, B), C), D)$ coalesces in the population ancestral to the root of the species tree.
2. (A, B) coalesces in the branch of length u_3, and then both $((A, B), C)$ and $(((A, B), C), D)$ coalesce in the population ancestral to the root of the species tree.
3. (A, B) and $((A, B), C)$ coalesce in the branch of length u_5, and $(((A, B), C), D)$ coalesce in the population ancestral to the root of the species tree.
4. (A, B) coalesces in the branch of length u_5, and then both $((A, B), C)$ and $(((A, B), C), D)$ coalesce in the population ancestral to the root of the species tree.
5. All coalescent events occur in the population ancestral to the root of the species tree.

Since for scenario 1 there are never more than 2 lineages in any population, and we have seen that the probability that 2 lineages coalesce within an edge of length x is $1 - e^{-x}$, the probability is easy to compute as

$$(1 - e^{-u_3})(1 - e^{-u_5}).$$

Note that the coalescent event in the population ancestral to the species tree root is sure to occur, and so contributes a factor of 1 to the probability.

The probability for scenario 2 is not much harder. It is

$$(1 - e^{-u_3})(e^{-u_5})(1/3).$$

the first two factors are the probability there is a coalescence in the edge of length u_3, and that there is not one in the edge of length u_5. Since 3 lineages are present at the species tree root, the factor of 1/3 is the probability that the correct pair of the three possible ones coalesces next.

The remaining scenario probabilities can be worked out similarly (see Exercise 14), and the sum of the five of them gives the probability of the topological gene tree.

While not all scenarios leading to a topological gene tree are equally probable, the more scenarios there are, the larger the total probability can be. In particular, if a gene tree has several rankings, then that produces more scenarios, and tends to result in a higher probability for such gene trees than one might expect. In fact it is possible that the gene tree with the greatest probability has a *different* topology than the species tree, or that several gene tree topologies are more probable than the one matching the species tree. Although this phenomenon of *anomalous gene trees* has been studied extensively by Degnan and Rosenberg, here we give only a simple argument to illustrate it.

First consider a 'star' species tree relating 4 taxa a, b, c, d, with 4 pendant edges emerging from the root. Probabilities of topological gene trees are then easy to see, since all coalescent events occur ancestral to the species tree root. That means our earlier analysis of four lineages samples from a single population applies, so that each of the 12 possible caterpillar trees has probability 1/18, and the 3 balanced trees have probability 2/18 due to their two rankings. Now if we instead consider any binary 4-taxon species tree, and make all internal edges very short, the probabilities of the gene trees will not be very different (since the probabilities are continuous functions of the internal edge lengths, and as these approach 0 we move to the star tree). Since there is such a gap between the probabilities of the caterpillar and balanced gene trees in the star species tree case, we will still have a balanced tree as the most probable for the binary species tree. In particular, for a 4-taxon caterpillar species tree with sufficiently short internal edges, the most probable topological gene tree will still be balanced, and thus not match the species tree.

A more detailed analysis, in which exact probabilities of topological gene trees as functions of species tree edge lengths are computed, can show exactly what edge lengths allow for anomalous gene trees, and exactly which gene tree is most probable.

The examples above indicate how the multispecies coalescent model can be used to understand gene tree distributions. But as with most phylogenetic computations, when there are more than a few taxa calculations are best handled by software. But before turning to the use we might make of these probabilities, we summarize by listing some key points concerning the model:

- Given a metric rooted species tree, with edge lengths in coalescent units, it is possible (but painful) to compute the probabilities of either metric or topological gene trees.

- The distribution of gene trees, whether metric or topological, that arises from the multispecies coalescent model carries information about the species tree topology and edge lengths. Thus using many gene trees to estimate this distribution should allow us to infer the species tree. The discordance of gene trees that might otherwise be viewed as problematic is actually a source of meaningful information.

- The probabilities of metric gene trees are for gene trees with edge lengths *in coalescent units.* If we wish to relate these to gene trees inferred by standard phylogenetic methods, whose edge lengths are typically measured in amount of substitutions, we must make further assumptions.

- Species trees need not be ultrametric in coalescent units. Indeed, if the population sizes vary over the species tree, then typically it will not be ultrametric in these units. By working in these units we can allow very general changes in population sizes over time. If we wish to relate coalescent units to true time, we must make further assumptions, such as that the population size is constant over the entire tree, or that we have a specific description of the way the population size changes.

- The most probable topological gene tree need *not* match the species tree topology. If there are more than 3 taxa, and internal edges of the species tree might be short, picking the most frequent gene tree topology inferred from many genes is not a reliable way of inferring the species tree. Hoping that the species tree topology will be ‘obvious’ to the naked eye from considering many genes is simply naive.

13.5 Inferring Species Trees

A variety of ways of inferring species trees, either from a collection of gene trees or from a collection of sequence alignments, have been proposed and implemented in software in the past few years. However, these data analysis methods are still undergoing development, and using them is not yet routine. The scientific community does not yet have enough experience with them to understand fully their strengths and weaknesses, or under what circumstances they are likely to perform well or poorly. The collection [KK10] offers a number of articles on both theory and practice with some of these methods.

There are several different ways one could classify currently proposed methods. Some are based explicitly on the coalescent model, and have been proved to be statistically consistent. Others ignore the coalescent, and use some sort of heuristic approach to ‘averaging’ over different genes. One could also group them by whether they use metric gene trees, or only their topologies. Alternately, they could be classified by whether they attempt maximum likelihood, Bayesian analyses, parsimony, or other combinatorial approach to determining a species tree. Some use a combined model of sequence evolution and the multispecies coalescent, as opposed to performing inference as a two-step process by first inferring gene trees and then using these as ‘data’ to infer a species tree. Since any such classification scheme would emphasize one feature as being more important than another, we will instead simply list methods in no particular order.

We describe some of these methods only in the case where we sample one gene lineage from each taxon, so that if we are studying n taxa, each gene tree will have n leaves. There are extensions of most of these methods to cases of multiple samples of each gene lineage per taxon, including allowing different number of samples for each gene. Provided the sampling in each taxon is done well, data collected with multiple samples has potential for improved inference (at least in the case where pendant species tree branches are relatively short).

Concatenation of sequences.

Given sequences for a number of different genes, an early approach that remains in widespread use is to concatenate sequences from all the genes, and treat the combined sequence as if it was one giant gene in order to perform a standard ML or Bayesian phylogenetic analysis. Software returns an inferred tree (or distribution of trees), which is then proclaimed to be the inferred species tree. Bootstrap support is often high, since the sequences are quite long, so this resampling technique showed little variability.

Though this method still has its advocates, there is no theoretical reason one should expect it to be reliable if incomplete lineage sorting is an important factor in why gene trees vary. By concatenating genes and analyzing them as one long sequence, even if a partitioned analysis allows for different numerical parameters for each gene, one is forcing the analysis to use a single tree topology. But the fact that there are likely to be different gene tree topologies is exactly what we are hoping to overcome! In other words, we are using a model that we know is seriously incorrect as the basis of our analysis. Moreover, we have in no way used our understanding of how incomplete lineage sorting occurs to inform the way we analyze the data. Finally, it has been proved by Steel and Roch that concatenation is not a statistically consistent method of inference of a species tree under the multispecies coalescent. However, if one is convinced ahead of time that all edges in the (unknown) species tree are long, then incomplete lineage sorting may be negligible, and this approach is better justified.

Maximum Likelihood and Bayesian analyses

At the other extreme from concatenation is to use both a model of gene tree formation by the multispecies coalescent together with a standard phylogenetic model of sequence evolution of the sort discussed earlier in these notes. Then either Maximum Likelihood or a Bayesian framework can be adopted for a well-justified statistical framework, using the models in succession , or combined.

We first sketch the ML approach used in the software STEM (Kubatko, *et al.* (2009)). Given many gene sequences, one first infers metric gene trees by using ML and standard phylogenetic models. One then must convert the branch lengths on the gene trees from units of total substitutions to coalescent units. Assuming a common mutation rate of μ on all edges of all gene trees, one can divide by μ to obtain true times. Then assuming a constant population size N for all populations on the species tree, one can divide by $2N$ (for diploid organisms) to convert to coalescent units. Note however that these assumptions imply all gene trees should be ultrametric, so they must either be 'adjusted' somehow, or inferred with that as a constraint. With this done, it is now straightforward (though a substantial amount of work) to implement a standard ML analysis under the coalescent model, with a new tree search for the species tree. There are some generalizations one can make, for instance allowing independent rescalings of the various gene trees to account for the fact that they may evolve at different rates.

Bayesian approaches are similar, and have been implemented in software by two different groups, in Mr. Bayes/BEST (Hulsenbeck, *et al.* Liu, *et al.*), and in *BEAST (Heled, *et al.*) It is a bit easier to program an analysis using a combined model of both the coalsecent for formation of gene trees and a substitution model for evolution of sequences. Letting D_i denote DNA sequence data for the ith of k genes, G a metric rooted gene tree, and S the species tree, the posterior is computed in the usual way from a prior on metric species trees and the likelihood function

$$\mathcal{P}(D|S) = \prod_{i=1}^{k} \sum_{G} \mathcal{P}(D_i|G)\mathcal{P}(G|S). \tag{13.9}$$

Just as for ML, the issue of converting time scales on gene trees from substitutions to coalescent units arises, and is dealt with similarly.

Note that there is, as yet, no ML software that implements a combined model of the coalescent and the substitution process, in which a likelihood function for the species tree given the sequence data are used. STEM uses a two-step inference procedure instead, which means that while there is some statistical error in the inferred gene trees, they are treated as if they were 'data' in the coalescent model analysis. The Bayesian analyses, though, do not have this issue.

Though in principal using Maximum Likelihood and Bayesian approaches to infer species trees is attractive, current implementations have their shortcomings. A worrisome issue is the conversion of time scales on gene trees, which is done by making assumptions over all gene trees of a fixed mutation rate (i.e., a molecular clock) and over the species tree of a fixed population size. If these are approximately valid they may not cause problems, but if they are violated strongly it is unclear what the impact is. There have been reports of Bayesian analyses not converging to a stable posterior for some data sets, though the reasons for this are not clear. Liu now warns that if the gene trees appear to be strongly non-ultrametric, BEST may not perform well.

In addition, the amount of computation for the Bayesian approaches is large enough, that there are practical limits on sizes of data sets, both in the number of taxa, and in the number of genes. While future generations of programs are likely to be more efficient, it is unlikely that will be sufficient to really increase limits substantially.

Pseudolikelihood

A pseudolikelihood method follows the same approach as ML, but replaces the true Likelihood function by something simpler. Liu *et al.* (2010) defined a pseudolikelihood function for the species tree given a collection of gene trees by considering all the topological rooted triples displayed on the gene trees. For instance, a gene tree $((a, b),(c, d))$ displays the four rooted triples $((a, b), c)$,$((a, b), d)$,$((c, d), a)$,$((c, d), b)$. After counting how often each rooted triple occurs on the gene trees, one computes the probability of each rooted triple given a species tree. Treating the rooted triples as independent, the pseudolikelihood is simply the product of these probabilities, each raised to the count. This is a considerably simpler function than the true likelihood function, so that software can run much faster. Of course, the rooted triples are *not* independent, so the standard guarantees that ML behaves well do not apply. However, simulation evidence is that this works well, and it can deal with much larger datasets than the true Likelihood approach. The software implementing this is MP-EST.

Minimizing deep coalescence.

This method is in spirit very similar to the use of parsimony for inferring a gene tree. It is based in a reasonable assumption that while incomplete lineage sorting can occur, we are unlikely to see extreme examples of it. Thus if we have a collection of gene trees (previously inferred from sequence data, and assumed correct), we should choose as the 'best' species tree the one which, if all the gene trees had arisen on it, would have coalescences of lineages as close to the most recent common ancestral species as possible.

Conceptually, the method can be performed as follows. We consider each possible species tree T, and for each of the given gene trees g we compute a score $s_g(T)$ that measures the minimal number of 'deep coalescences' that are necessary for g to have arisen on T. We add these scores for all gene trees, to obtain a score for T:

$$s(T) = \sum_g s_g(T).$$

We then choose the tree(s) T that has the minimal value $s(T)$.

The score $s_g(T)$ is defined as shown in Figure 13.7. First, consider any internal node v in the gene tree, and let X_v denote its leaf descendants. Then v represents a coalescent event that could only have occurred on the species tree above (temporally before) the most recent common ancestor of X_v on the species tree. We therefore assume it occurred in the population just above the MRCA, and locate it along that edge of the species tree. Doing this for every node of g gives us a map of the gene tree to the species tree, and allows us to talk about the number of gene tree lineages entering (at the child end) and leaving (at the parent end) every edge. If the gene tree had the same topology as the species tree, there would be two lineages entering each species tree edge, and 1 lineage leaving. For a highly discordant gene tree, these counts will typically be larger. We define the number of *excess* lineages for an edge of the species tree as 1 less than the number leaving the edge. Then $s_g(T)$ is the sum over all edges of T of the number of excess lineages. Thus for a gene tree matching the species tree we have $s_g(T) = 0$, and discordant gene trees will have higher scores.

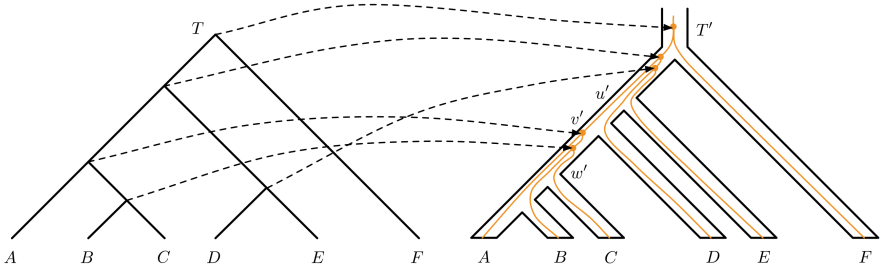

Figure 13.7: To compute the score $s_g(T)$ for a gene tree g on left and a species tree T on right, the nodes of g are first mapped to the lowest population edges on T on which they could have arisen. Then $s_g(T)$ is the sum over all edges of T of the number of 'excess' lineages in g exiting them at the ancestral end. (FIGURE from Than-Nakleh 2009)

Notice parsimony ideas have appeared twice in this method, once for assuming coalescences occur 'as soon as possible' on the species tree, and once for choosing the species tree with the lowest overall number of excess lineages.

Like parsimony for phylogenetic inference, this method is both reasonable and likely to work well under some circumstances. Moreover, Than and Nakleh have reformulated the optimization problem in ways that can be solved by techniques of either integer or dynamic programming, giving fast performance in practice. Unfortunately, and also like parsimony, it is known that minimizing deep coalescence is not a statistically consistent inference method for some species trees.

Finally, we note that this method uses only topological information on gene trees, and thus avoids attempting to relate time scales on the gene trees to those on the species tree. It also does not make any direct use of the coalescent model.

Consensus methods

If we have already inferred a collection of gene trees from sequence data, we can also simply combine them with a standard consensus method. While at first this might seem like just a combinatorial way to overcome the gene tree differences, and not one that has much to do with the coalescent, in fact this approach has some provably good properties under the coalescent model.

For a theoretical distribution of rooted gene trees, define the probability of a clade to be the sum of the probabilities of all gene trees displaying that clade.

The following theorem of Allman, Degnan, and Rhodes (2011) is the key to understanding how consensus methods behave for inferring species trees.

Theorem 19. Under the coalescent model on a binary species tree, any clade on a gene tree with probability greater than 1/3 is also a clade on the species tree.

The 1/3 cut-off in this theorem can not be reduced; for any lower number, examples can be constructed of clades with greater probabilities that are not on the species tree.

Now from a collection of gene trees one can estimate the probability of a clade by the proportion of the gene trees displaying it. If one builds a consensus tree from the clades with estimated probability greater than 1/3, then provided one has a large enough sample of gene trees, the only clades used will be ones on the species tree. As a result, the inferred tree may not be fully resolved, but any clades it shows will reflect species tree clades. Since the strict and majority-rule consensus trees will only use a subset of these clades, they may be less well resolved, but will also only reflect true species tree clades. While for an arbitrary collection of clades one cannot be sure those with frequency $\leq 1/2$ will be compatible, this result implies that if the trees come from the coalescent, all clades with frequency $> 1/3$ will be compatible, provided enough gene trees are given.

A more recent theorem of Allman, Ané, Rhodes and Warnow (2011, unpublished) establishes a similar result for splits on unrooted gene trees: Any gene tree split with probability greater than 1/3 is a split on the unrooted species tree. Thus constructing a consensus tree using gene tree splits of frequency $> 1/3$ will, for a large enough sample of gene trees, infer a tree that may not be fully resolved but whose splits will reflect ones on the species tree.

Greedy consensus, using either gene tree splits or clades, is however not statistically consistent, as has been shown by Degnan, Degiorgio, Bryant, and Rosenberg (2009). Accepting any splits or clades with frequency below 1/3 will not necessarily give the correct species tree, even with an infinite sample of gene trees.

Another approach, called R^* consensus (Degnan, *et al.* (2009), is justified by the result proved in section 13.4 that for any 3 taxa, the most probable gene tree matches the species tree. Given a collection of gene trees, one could consider every subset of 3 taxa, and the 3-taxon gene trees they induce. By counting the ocurances of each of the 3 possibilities, one can infer the 3-taxon tree induced on the species tree as the most frequent one. For 3-taxon trees, as we have seen, this is a consistent way of picking the species tree. Once one has obtained all these 3-taxon induced trees, one can then build the species tree displaying them all. In practice, one may find some of these are incompatible, so decisions must be made as to how that is to be handled. (We will not give details here.) Unlike the previous consensus methods mentioned, this method will, provided we have enough gene trees, result in a fully-resolved species tree. Moreover, it is statistically consistent.

A similar approach replaces rooted triples from rooted gene trees with quartets on unrooted gene trees. The most frequent gene tree quartet topology does consistently infer the species tree quartet topology, so one can then search for a species tree displaying most such quartets. With some elaborations, this is the approach taken in the ASTRAL software of the Warnow group.

STAR and NJ$_{st}$

There are also methods proposed by Liu and collaborators, that infer species trees not by the standard consensus approach, but by a process that in spirit extends them. These are both very fast, since they are built on distance methods of tree inference, and have been shown to have good performance in simulations,

The first of these, called STAR (Liu *et al.*, 2009) proceeds as follows to obtain a species tree from a collection of rooted topological gene trees. First, assign a length of 1 to all internal branches of all the gene trees. Then make pendant edges have whatever length is necessary so all leaves are distance n from the root, where n is the number of taxa. Now that the gene trees have all been made into ultrametric trees, compute a distance matrix for each, giving distances between the leaves. Next, average these gene tree distance matrices over all the gene trees (i.e., average each of the corresponding entries). From this average distance matrix, build a tree using your favorite distance method, such as Neighbor Joining. Finally discard the metric information on this tree and report it as the topology of the species tree.

Though at first this seems like a rather Rube Goldbergian way to infer a species tree, it is not as ridiculous as it may sound. Though its introduction was accompanied only by a proof of a 4-taxon special case, it turns out that this is a statistically consistent way to infer species trees assuming the multispecies coalescent. Moreover, simulations showed it can work reasonably well even when provided with only a small number of gene trees. A rigorous proof of consistency given by Allman, Degnan, and Rhodes (2013).

A similar proposal of Liu, *et al.* (2011) uses unrooted topological gene trees to infer an unrooted topological species tree. For this method, originally called NJ$_{st}$, all gene tree branches, internal and pendant, are assigned a length of 1, distance matrices for each are then calculated and averaged, and an unrooted tree is chosen to fit this average, for instance by Neighbor Joining. A proof of the statistical consistency of this has not yet been published is in the works. It has also been implemented efficiently in software called ASTRID (Vachaspati and Warnow, 2015), and shown in simulations to work well for very large data sets, with thousands of taxa and genes.

GLASS, STEAC

TO BE WRITTEN

Concordance factors

Although it is not based on the coalescent model, or any model of why gene trees might differ, a useful software tool for summarizing and understanding gene tree discordance in a Bayesian setting is BUCKy (Larget, *et al.* (2010)). It performs a technique called Bayesian concordance analysis (Ané, *et al.* (2007)) to take posterior distributions for many gene trees, and combine them into what can be interpreted as support for various clades. The lack of a model underlying this tool means that even if the coalescent process is not the source of the gene tree discordance, for instance, if horizontal gene transfer or hybridization occurs, one may still use it to gain some insights.

As a final comment, we note that most of these methods implicitly assume the genes being analyzed are unlinked. This is necessary so that each gene tree can be viewed as an independent trial of the coalescent process. If for instance, one wished to use mitochondrial genes in mammals, this assumption would be strongly violated due to maternal inheritance. For nuclear autosomal genes, one should also take some care not to use genes too closely located on a chromosome.

13.6 Exercises

1. Show the finite series in equation (13.1) has the stated value.

2. Show the series in equation (13.2) has the stated value. You will need to use the formula for the sum of the geometric series $\sum_{n=0}^{\infty} r^n$, and differentiate with respect to r.

3. The Wright-Fisher model behind the simulation in Figure 13.3 leads to an expected time to coalescence of any two lineages of 8 generations. Compute the average of the coalescent times for the $\binom{8}{2} = 28$ pairs of lineages in the simulation, and compare.

4. Show the integral in equation (13.3) has the give value.

5. Suppose a population is exponentially growing, so $N(t) = N_0 e^{-\alpha t}$ with $\alpha > 0$ where t is measured backwards from the present. Use equation (13.4) to give a formula relating coalescent units and true time.

6. Complete the derivation of the formula in equation (13.5).

7. By computing a double integral of the probability density in equation (13.6), recover that the probability of the topological gene tree $((A_1, A_2), A_3)$ is $1/3$.

8. Under the single population coalescent model, the probabilities of all 4-sample gene trees are computed in the text. How many such trees are there? How many of these are caterpillars? How many are balanced? Show the computed probabilities add up to 1.

9. Describe all possible orderings of coalescent events that could have led to each of the gene trees

   a) $((((A, B), C), D), E)$

   b) $(((A, B), C), (D, E))$

10. Explain the formula for the number of ranked gene trees in equation (13.7) by considering the formation of a tree by starting with n lineages and choosing pairs to coalesce according to the ranking.

11. If n genes $A_1, \dots, A_n$ are sampled from a single population, under the coalescent model compute

    a) the probability that the gene tree relating them is the specific caterpillar $((\dots((A_1, A_2), A_3), \dots, A_{n-1}), A_n)$.

    b) the probability that the gene tree relating them is *any* caterpillar tree.

12. Suppose lineages A_1, A_2, A_3, A_4 are sampled in a single population. Under the coalescent model, the probability density for a metric gene tree relating them will depend only on the 3 times at which the coalescent events occurred.

    (a) Find the probability density for $((A_1{:}u_1, A_2{:}u_1){:}u_3, (A_3{:}u_2, A_4{:}u_2){:}u_4)$ when $u_1 < u_2$.

    (b) Find the probability density for the same tree when $u_2 < u_1$.

13. Check that the formulae in equation (13.8) are correct. If not, give correct ones, and inform the authors.

14. In Section 13.4, five scenarios were considered in the text in which the topological gene tree $(((A, B), C), D)$ is formed under the multispecies coalescent model on $(((a{:}u_1, b{:}u_2){:}u_3, c{:}u_4){:}u_5, d{:}u_6)$, and the probabilities were computed for two of them. Compute the probabilities of the three remaining ones, and add them to obtain the probability of $(((A, B), C), D)$.

    Note: Scenarios 3 and 4, which involves coalescence of 3 lineages to either 1 or 2 in a single branch of length u_5, are the most involved to compute. For instance, for scenario 3, let x be the time of the coalescent event, You will need to explain why the integral

    $$\int_0^{u_5} e^{-3x}(1 - e^{-(u_5 - x)})dx$$

    arises, and evaluate it.