sequences led to more accurate inference, but that is not necessarily true for all methods.

In this chapter, we discuss one circumstance that is well-understood to cause difficulties in phylogenetic inference: the phenomenon of *long branch attraction*.

Figure 11.1: A metric quartet tree

Consider the metric quartet tree of Figure 11.1. Here two of the pendent edges, in different cherries, are much longer than the other edges. That such a tree can be difficult to correctly infer from data is not too surprising. The two taxa that are the most closely related metrically are not most closely related topologically. In fact, this was the problem that motivated our introduction of neighbor joining as an improvement over UPGMA when we discussed distance methods.

But the neighbor joining algorithm, or even a full maximum likelihood approach to inference, may still perform poorly for real data from such a tree. Our goal in this chapter is to be more precise about this issue.

11.1 Statistical Consistency

Suppose we consider some model of the evolution of sequences along a tree, and some method of inference. To be concrete, we might focus on the Jukes-Cantor DNA model, and Neighbor-Joining with the Jukes-Cantor distance as the method of inference. The most basic question we can ask about this pair is whether we would correctly infer the tree if we analyzed data generated by this model according to this method.

However, this is not a simple yes-or-no question. With a small amount of data, we will certainly sometimes infer the wrong tree. For an extreme thought experiment, imagine data sequences of only a few sites. Then just through randomness, we might find all the sequences were identical. All distances between sequences would then be 0, and the inferred tree would be just a single vertex, representing all taxa.

On the other hand, with much more data in the form of very long sequences, we should expect that on average the observed joint distribution of patterns will match the theoretical joint distribution predicted by the model fairly well. Then when we compute distances, we should obtain ones fairly close to the true

distances determined by the model parameters. Then Neighbor Joining should work well, and is likely to recover the true tree.

In order to formalize these ideas, we make the following definition.

Definition. A method of inference is said to be *statistically consistent* for a particular model if, for any $\epsilon > 0$, the probability that inferred model parameters are within ϵ of the true values approaches 1 as the amount of data approaches infinity. That is, if the process has parameters s, and $\hat{s}$ is the inferred value of the parameters from data produced in n independent trials, then

$$\lim_{n\to\infty} \mathbb{P}(||\hat{s} - s|| < \epsilon) = 1.$$

Consistency should seem like a very basic requirement for a method of inference to be a good one. If a method is *not* consistent, then even if you had access to unlimited amounts of data, you could not reduce your doubts that your inference was wrong to arbitrarily small values. Note also that consistency of a method of inference is always in reference to a specific model. Thus it does not address the very real problem of model misspecification leading to erroneous inference. Consistency deals with an idealized problem, where we know the correct model, and have as much data as we like. If a method does not work well under such circumstances, then we should have serious doubts about its performance in the real world. If it does work well under these idealized circumstances, then we should next consider the impact of both model misspecification and limited data.

11.2 Parsimony and Consistency

Suppose we choose to use parsimony on a 4-taxon dataset to infer a tree. For the 4 taxa a, b, c, d, we have a sequence of characters. In this setting, parsimony reduces to a simple scheme.

First, there are only a few types of characters that are informative. An informative character must have for the 4 taxa a, b, c, d, in order, states

$$xxyy, \quad xyyx, \text{ or } xyxy,$$

where x, y denote two distinct states. So letting n_1, n_2, n_3 denote the count of such characters in a data set.

We wish to compute the parsimony score for this collection of charcters on the 3 quartet trees

$$T_1 = ab|cd, \quad T_2 = ad|bc, \quad T_3 = ac|bd, .$$

For T_1, each character $xxyy$ will give a parsimony score of 1, while characters $xyxy$ and $xyyx$ will produce scores of 2. Taking into account the number of these characters, we find

$$ps(T_1) = n_1 + 2n_2 + 2n_3 = (2n_1 + 2n_2 + 2n_3) - n_1.$$

Similarly,

$$ps(T_2) = 2n_1 + n_2 + 2n_3 = (2n_1 + 2n_2 + 2n_3) - n_2, \tag{11.1}$$
$$ps(T_3) = 2n_1 + 2n_2 + n_3 = (2n_1 + 2n_2 + 2n_3) - n_3.$$

To choose the most parsimonious tree(s) T_i, we therefore simply pick the value(s) of i maximizing n_i.

Now suppose our data are generated by a probabilistic model of the sort in Chapter 6. Then we can compute expected values of the numbers n_i, and see whether parsimony will infer the correct tree. Of course, using expected values of the n_i is essentially the same as imagining we have infinitely long data sequences that are produced exactly in accord with our model. Although we are omitting some details (involving limits, and formal treatments of probabilities of correct inference) needed for a rigorous mathematical treatment, we are dealing with the issue of consistency.

More precisely, we ask the question: If parsimony is applied to longer and longer sequences produced exactly according to the model, do the inferred trees eventually stabilize on the correct one?

The first examples of parameter choices for a Markov model leading to inconsistency of parsimony are due to Felsenstein, for a 2-state model. Although the result has been generalized, for simplicity we follow the original argument.

Theorem 18. For a 2-state Markov model on quartet trees, there are parameters for which parsimony is an inconsistent inference method.

Proof. For the tree in Figure 11.1, place the root ρ at the internal vertex joined to a, b, and denote the other internal vertex by v. Consider the two-state Markov model with parameters

$$\mathbf{p}_\rho = \begin{pmatrix} 1/2 & 1/2 \end{pmatrix},$$
$$M_{(\rho,a)} = M_{(v,d)} = \begin{pmatrix} 1-q & q \\ q & 1-q \end{pmatrix},$$
$$M_{(\rho,b)} = M_{(\rho,v)} = M_{(v,c)} = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}.$$

Then the probabilities of the 3 patterns that are informative for parsimony are

$$p_1 = p_{xxyy} = (1-q)^2 p(1-p)^2 + 2q(1-q)p(1-p)^2 + q^2 p^3,$$
$$p_2 = p_{xyyx} = (1-q)^2 p^2(1-p) + 2q(1-q)p^2(1-p) + q^2(1-p)^3, \tag{11.2}$$
$$p_3 = p_{xyxy} = (1-q)^2 p^3 + 2q(1-q)p(1-p)^2 + q^2 p(1-p)^2.$$

But as the sequence length, N, goes to infinity, we have that the proportion of time we see pattern i is $n_i/N \to p_i$. Since parsimony will be consistent only when $n_1 > n_2, n_3$, we have consistency only when $p_1 > p_2, p_3$.

Now, by straightforward algebra, equations (11.2) imply

$$\begin{aligned} p_1 - p_2 &= (1-2p)(p(1-p) - q^2), \\ p_1 - p_3 &= p(1-2p)(1-2q). \end{aligned} \tag{11.3}$$

In these formulas we should only consider values of $p, q \in (0, 1/2)$ as biologically plausible. (See also Exercise 4.) For parameters in this range,

$$1 - 2p > 0, \quad 1 - 2q > 0,$$

so $p_1 > p_3$ always holds. In addition, $p_1 > p_2$ holds precisely when

$$p(1-p) > q^2.$$

However, there are values of p, q in the allowed range where this last inequality does not hold, and so parsimony is inconsistent for such choices. □

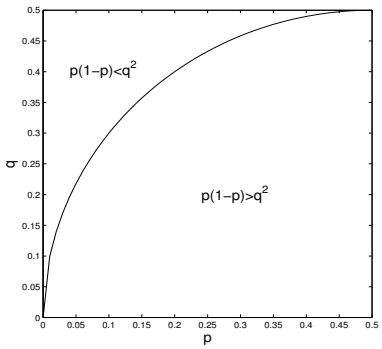

Figure 11.2: Parsimony is consistent only for parameters below the curve.

A graph of the (p, q)-parameter space for the model in the proof is given in Figure 11.2, indicating the regions for which parsimony is a consistent method of inference. We see that if p is sufficiently small in comparison to q, so the internal branch of the tree is very short, we have inconsistency of parsimony.

Note that when the parameters fall in the region of inconsistency (sometimes called the *Felsenstein zone*), the proof has shown $p_2 > p_1$, so the tree parsimony will infer from perfect data is T_2. In other words, the two long branches are erroneously joined, and we have an instance of *long branch attraction*.

An extreme interpretation of this theorem and figure is that since parsimony is sometimes inconsistent, it is not a good method to use. A more moderate view is that parsimony is in fact consistent over quite a wide range of parameters, and we now have some idea of what might lead it to behave poorly. Of course for real data we might not have much idea in advance of how the tree should look, so that whether the tendency for long branches to attract one another is

a problem can be difficult to determine. However, if we infer a tree with several long branches joined to one another, we should consider the possibility.

Of course this theorem does not apply to a 4-state model more appropriate to DNA. However, similar calculations for a Jukes-Cantor model show the same phenomenon; for some edge lengths, parsimony will be a consistent method of inference of a 4-taxon tree topology, but for others it will not.

But the situation can actually be worse than this last statement seems to indicate. In fact, for a 4-taxon tree it is possible to show that under a κ-state generalization of the Jukes-Cantor model, for any $\kappa \geq 2$, the set of distributions of parsimony-informative sites that arises on one tree topology is *exactly* the same as on any other tree topology (Allman, Holder, and Rhodes 2010). In other words, with only 4 taxa the parsimony informative sites alone give no information whatsoever about the tree topology. Thus when parsimony does pick the correct tree, it's essentially just getting lucky. (Note however that when more taxa are used, the parsimony-informative sites do carry information on the tree topology.)

But all these results must be interpreted very carefully. For instance, as Figure 11.2 shows in the 2-state case, parsimony is consistent if all branch lengths are of equal size, and in many other circumstances as well. Thus we have identified a *possible* source of error in inference by parsimony, not a fatal flaw. The problem, of course, is that it may be hard to know ahead of time whether we should expect parsimony to be perform well on any given data set.

Also note that the metric tree used in the proof above is not ultrametric; unless $p = q$ there is no place a root can be located so that all leaves will be equidistant from it. In fact, for a 4-leaf molecular clock tree, it's possible to show parsimony is consistent for simple models. Unfortunately, for larger sets of taxa parsimony can again be inconsistent even assuming the tree is ultrametric.

11.3 Consistency of Distance Methods

The other methods of inferring trees that we've discussed — distance methods and maximum likelihood — are statistically consistent when paired with certain models, though some mild restrictions may be needed on parameter values. We won't give formal proofs of these facts, but will instead sketch the ideas.

Consider first the computation of distances from data sequences produced according to a specific model. Then as sequence length grows, it is easy to show that provided we use a distance formula that is associated with the model generating the data, we will infer distances that approach the true ones for the parameter choice. (Thus if our data are generated by a Kimura 2-parameter model, we may use Kimura 2-, or 3-parameter distances, or the log-det distance, but not the Jukes-Cantor distance.) This basic fact follows from the continuity of the distance formulas.

Once we have distances, suppose we use the Neighbor Joining algorithm to infer a tree. We've already claimed in Theorem 16 (and shown in Exercise 24

of Chapter 5) that Neighbor Joining recovers the correct tree provided we have *exact* distances from a binary tree with all positive edge lengths. It's necessary, then, to show that from dissimilarities sufficiently close to the exact distances, we recover the same tree. So what is needed is in essence a statement that the output of the algorithm is a continuous function of the input dissimilarities, at least in the neighborhood of exact distances for binary trees with all positive edge lengths. This is certainly plausible, and in fact can be rigorously proved. For other distance methods, such as the least-squares approach, consistency can be established similarly

Note that the requirements that the tree be binary and edge lengths be positive are the mild restrictions that we indicated were needed beforehand.

An important point in this, however, is that we need to be able infer distances correctly. If data are generated by a mixture model, but distances are computed by a simpler model, then there are no guarantees that further processing of these incorrect distances will lead to a correct tree. Thus consistency of any distance method is likely to depend on the consistency of the inference of pairwise distances between taxa. Since distance formulas exist only for the most basic models, we will be limited to these if we insist on consistency.

11.4 Consistency of Maximum Likelihood

There are rather general proofs that maximum likelihood is a consistent method of statistical inference in many settings, and these can be modified for phylogenetic models. However, these proofs require that one first establish that the model has *identifiable* parameters. In the phylogenetic setting, this means that any joint distribution of patterns at the leaves of a tree arises from a unique tree and choice of numerical parameters. If different choices of trees and numerical parameters lead to the same joint distribution of patterns, then even with large amounts of data they would be indistinguishable, and no method would be able to consistently infer parameters. Thus a lack of identifiability for a model would cause *any* method to be inconsistent. Maximum likelihood has the fortunate feature that it is generally consistent without any additional substantive requirements on the model beyond identifiability.

In fact, parameters for the general Markov model are not strictly identifiable. For instance, suppose two n-taxon tree topologies are chosen that are one NNI move apart. For numerical parameters on each, choose the same root distribution, and the same Markov matrices on all edges not affected by the NNI move. On the edge that the NNI move collapsed on the first tree, and the new edge it introduced in the second, let the Markov matrix be the identity. Then these two trees and parameter choices will produce exactly the same joint distribution of patterns at the leaves. The reason for this, of course, is that the choice of the identity as a Markov matrix means no substitutions occur on the edges affected by the NNI, so both trees produce sequences as if there is an unresolved polytomy, with 4 edges emerging from a single vertex. The simple way to rule this problem out is of course to not allow Markov matrices to be

the identity.

By imposing restrictions of this sort, under the general Markov model the tree topology is identifiable, by means of the log-det distance and 4-point condition, for instance. For this, it is sufficient to require that the tree be binary and all Markov matrices have determinant $\neq \pm 1, 0$. This rules out not only the identity matrix for Markov matrices, but also some other possibilities that are fortunately not relevant biologically. For instance, in a GTR submodel, the only way a Markov matrix could have determinant 0 is if the edge length were infinite.

There remain other non-identifiability issues, though. For instance, one can permute the bases at an internal node of the tree, adjusting the Markov parameters on incident edges accordingly by permuting rows or columns, and again not change the joint distribution (Exercise 7). This gives only finitely many parameter choices for each joint distribution, though. Moreover we can make a unique choice from these by imposing biologically reasonable assumptions that the diagonal entries of all matrices be the largest in their rows.

Even after eliminating these issues, it is not easy to show the parameters are identifiable. The essential difficulty is that phylogenetic models have hidden variables, representing the states at internal nodes of the tree, which cannot be observed. These greatly complicate the form of the model parameterization, leading to the many-term sum of products we saw in Chapter 6. Moreover, there are similar models with hidden variables outside of phylogenetics that are *not* identifiable.

Currently, maximum likelihood has been proven to be consistent for the following models, by first showing the models are identifiable. For some of these models, there are minor technical conditions that must be placed on parameters, about which we will not be explicit.

- the GM model, and its submodels such as GTR, K2P, JC
- GTR+Γ, and its submodels such as K2P+Γ, JC+Γ
- GTR+Γ+I, except when the GTR component is JC or F81, when two trees differing by an NNI may give the same distribution
- GM+I, and submodels such as GTR+I, K2P+I, JC+I
- rate-variation and covarion models with c classes, provided c is smaller than the number of observable states (i.e., $c < 4$ for DNA models, $c < 19$ for proteins, etc.)

Much more general mixture models on a single tree topology are also known to be identifiable for *generic* choices of parameters, provided the number of components is less than a bound depending exponentially on the number of taxa. 'Generic' here means that if the numerical parameters are chosen at random, then almost certainly they will be identifiable. Mixtures in which components have different tree topologies are similarly identifiable, provided the trees all have two 'deep splits' in common. (Rhodes and Sullivant, 2012)

These theoretical results safely cover all models used routinely for data analysis, and many that are being explored in less routine investigations. And while it is likely that even more complex models than these, that have yet to be proposed, have identifiable parameters, one should not forget that there are real limits. For instance, under the no-common-mechanism model mentioned in Chapter 10, the tree topology is not identifiable. Thus while we might find such a model appealing for its biological realism, it is not useful for data analysis.

11.5 Performance with Misspecified models

While statistical consistency is certainly desirable for an inference method, establishing it does not lay to rest all concerns we should have. First, a claim of consistency is a statement about behavior under idealized conditions: If we use the correct model, and have access to as much data as we like, then we are likely to draw the right inferences. If, say, we attempt to use maximum likelihood with a Jukes-Cantor model for inference, and the data actually is not fit well by that model, the consistency results above give us no guarantees. Since biological data are unlikely to be described perfectly by any simple model, we certainly have a violation of assumptions in any real-world application. How that violation of assumptions effects inference results is a question of *robustness*. While experience with data analysis generally indicates that the inference of the tree topology may be fairly robust to variations in choice of models, numerical parameters inferred under different models often vary more widely.

There are also specific circumstances that have been found where an analysis with a misspecified model leads to errors in tree inference. For instance, Shavit Grievink, Penny, Hendy, and Holland (2009) simulated data using a covarion version of the Jukes-Cantor model, in which the proportion of invariable sites changed over the tree. Though plausible as capturing reasonable biological behavior, this is not a model that has been implemented for data anlaysis in any software, nor studied theoretically. When a Bayesian analysis of the simulated data was performed under the model's closest implemented covarion cousin, which did *not* allow for changing proportions in the covarion classes, serious errors were made in recovering the tree topology. The extent to which such processes may be misleading us with analysis of real data is simply not known.

While it can be reassuring to analyze a data set under several models and see that the inferred parameters of interest are similar, one should never forget that the models currently available in software are limited. If they do not capture some key process in the data's production, they might all give similar results yet still be misleading.

11.6 Performance on Finite-length Sequences

Even with a correctly specified model used to analyze data, in the real world sequences are always of finite length, and therefore short in comparison to those

with which statistical consistency is concerned.

How well various methods perform on sequences of the lengths similar to experimental data sets has been investigated through simulation. We can choose a model, tree, and numerical parameters, and simulate data according to these. Then we can apply the inference method to the simulated data, and see if it recovers the original tree. (See Exercise 8.) If we use the model which generated the simulated data to analyze it, then *only* the effect of sequence length will be investigated.

One important finding of such simulation is that the long branch attraction phenomenon seems quite universal, regardless of the method of inference. For instance, for any fixed sequence length, there is a region of parameter space much like the upper left corner of Figure 11.2 in which we often infer the wrong tree. The precise shape and size depends on the method of inference and the sequence length, but the region is there, even for maximum likelihood. Because maximum likelihood is consistent, the size of the region must become smaller if the sequences are made longer. However, in practice we may not be able to obtain long enough data sequences whose evolution can be modeled well, so the assurance of statistical consistency may not be helpful.

Several long edges leading from the ends of a short edge, then, are quite generally problematic. It is wise to keep this is mind when inference produces a tree with several long edges leading from a common vertex. When such a tree is inferred, it might be possible to include additional taxa in the data set in order to create additional bifurcations breaking up the long edges, and with an expanded data set, we may be able to better infer a tree. However, there may be no taxa available to break up the long edges, so this problem cannot always be overcome.

11.7 Exercises

1. Explain the computation of the parsimony scores in equations (11.1) for the quartet trees.

2. Check the formulas for p_1, p_2, p_3 in equations (11.2) in the proof of Theorem 18.

3. Check the formulas for $p_1 - p_2$ and $p_1 - p_3$ in equations (11.3) in the proof of Theorem 18 (using software, or by hand).

4. Show that if the 2-state Markov matrix

$$M(p) = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$$

is of the form $M = e^{Qt}$ for some non-zero rate matrix Q, $t > 0$, then $p \in (0, 1/2)$. Why is this biologically plausible?

5. For the Markov model in the proof of Theorem 18, explain how Figure 11.2 fits with the intuition that as long as sufficiently few mutations occur,

parsimony should infer the correct tree. Is this intuition strictly correct? Show the curve $p(1-p) = q^2$ has a vertical tangent at (0,0), and explain why this is relevant.

6. Assuming we use a valid distance formula for our model, explain informally why UPGMA will be statistically consistent if the underlying tree has all leaves equidistant from the root, but will not be without this assumption on the tree.

7. Consider a 3-taxon tree, with root ρ at the central vertex. Suppose for the general Markov model on this tree we have parameters $\mathbf{p}_\rho$, M_1, M_2, M_3, producing a certain joint distribution of states at the leaves. If P is a permutation matrix and $\mathbf{p}'_\rho = \mathbf{p}_\rho P$, what Markov matrices M'_1, M'_2, M'_3 will, along with $\mathbf{p}'_\rho$, produce the same joint distribution as the original parameters?

8. Using software, simulate some sequences of length 300 bases according to a Jukes-Cantor model on the tree in Figure 11.1 using a variety of parameter choices (p, q) as in the proof of Theorem 18. Then use Neighbor Joining with the Jukes-Cantor distance to infer a tree. Find parameter choices for which the method seems to usually give the correct tree, and other parameter choices for which you usually see long branch attraction.

   If you are ambitious, perform many simulations for enough values of (p, q) to produce an empirical diagram like Figure 11.2, showing when this approach to this inference problem usually produces the correct tree.