Lecture 13 – Performance of Methods

I. Introduction: We've spent a fair amount of time describing the some of the various methods for phylogeny estimation. As we have experienced, there is a huge array of methods to choose from, and there are various criteria by which one might guide one's choice.

As the statistical view has become prevalent and has come to dominate phylogenetic research, performance has been acknowledged as relevant in choice of methods and this has led to a huge body of literature that assesses the performance of various methods.

The term "reliability" is often used, but without a very clear definition of what it is. Here, we'll address several criteria for evaluating the performance of methods and discuss three methods that have been used to assess performance using these criteria.

Criteria	Methods
Accuracy	Simulation
Consistency	Congruence Analysis
Efficiency	Experimental Phylogenies
Robustness	

II. Methods of assessing performance.

A. Simulation studies – One of the most widely used approaches to assessing phylogenetic performance is to simulate data and assess how well various methods estimate the true phylogeny (that was used to generate the data).

Simulations have the enormous advantage that a large number of replicates can be examined, and this allows us to account for stochasticity.

There are a couple different approaches to simulation studies. The first is **prospective simulations**, in which a set of conditions is specified *a priori*, and this defines the conditions under which data are simulated.

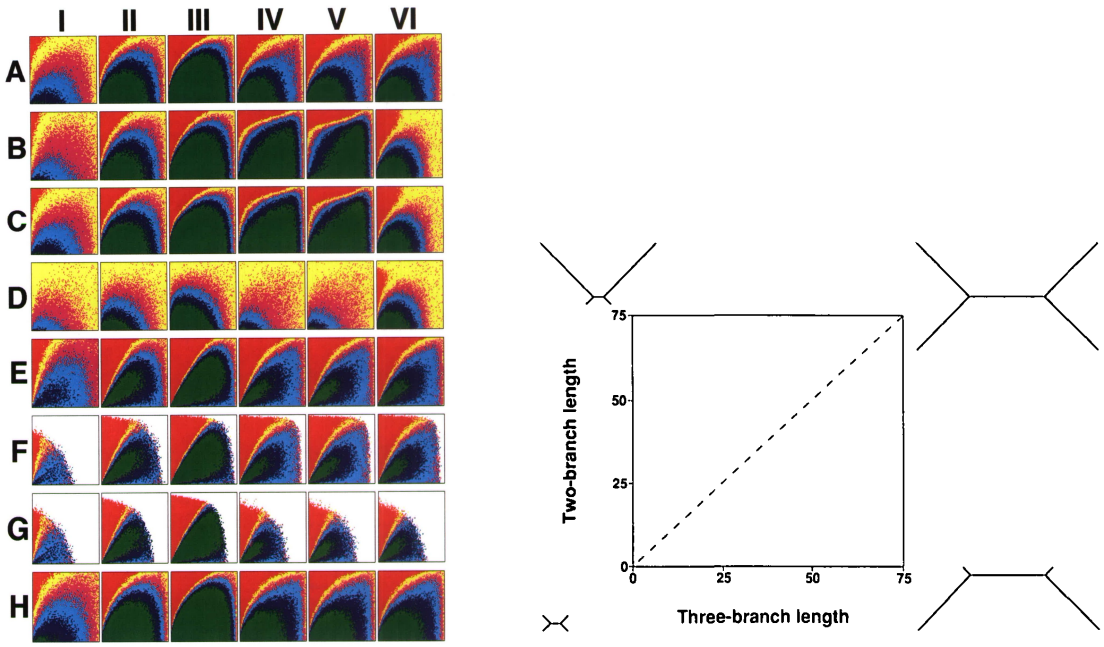
The second is **retrospective simulations**, in which simulation conditions are defined by analysis of a particular data set that's relevant to some question.

Both are very useful and have contributed enormously to our understanding of the performance of methods.

Prospective simulations have been particularly important, but they can certainly be abused.

The power of prospective simulations is that a variety of conditions may be examined, and the performance of methods can be compared across these conditions. The first paper to do really extensive prospective simulations of phylogenetic performance was Huelsenbeck &

Hillis (1993. Syst. Biol., 42:247). This paper led to a host of prospective simulation studies that have tremendously advanced our understanding of the conditions across which phylogenetic estimation methods perform well. This paper also defined the **Felsenstein Zone**.



The danger of prospective simulations, however, is that it's very easy to stack the deck. There are lots of examples of this in the phylogenetics literature. One of the first I became aware of these was Tateno et al. (1994. Mol. Biol. Evol. 11:261-277).

They simulated data under an F84+Γ model of sequence evolution. They then compared how well NJ on Γ-corrected distances did at estimating the tree with ML under an equal-rates model. Of course, they concluded that the NJ method with gamma-corrected distances outperforms ML under these conditions, but that was simply because they made inappropriate comparisons. They matched the model perfectly in NJ, but not in ML.

One major weakness of simulation studies is that the models that are used to simulate data are clearly overly simplistic. Thus, simulation studies, while very important and informative, have limitations and are subject to investigator bias.

B. Congruence Analysis – Use of well-corroborated phylogenies.

For real data, there's no such thing as a "known phylogeny." However, there are a few groups of organisms for which congruence among a large array of diverse data sets has resulted in the next best thing (Miyamoto and Fitch, 1995. Syst. Biol. 44:64-76). "Trees of

natural taxa, well supported by many independent lines of evidence, should be used in the same way as the known phylogenies of simulations and of certain laboratory and domesticated groups, i.e., as standards for evaluating the accuracy of different phylogenetic methods." These usually include model organisms, such as particular groups of deer mice, *Drosophila*, or something like that.

Sequence data are then collected and phylogeny estimated with a variety of methods and those that yield the well-corroborated relationships with greatest certainty are deemed to be best performing.

This has the advantage that the data have been produced by the actual complex evolutionary process that has led to the diversity of the group being used, circumventing the weakness of simulation.

There are several weaknesses, though.

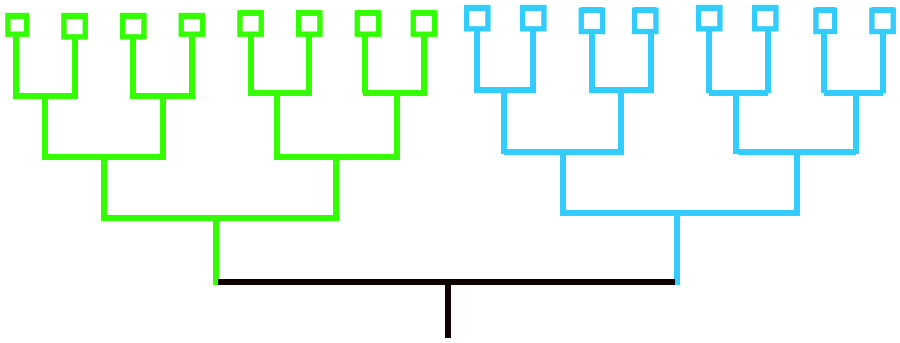
The history of the group can't be manipulated to explore different combinations of branch lengths and properties of the data.

Replication is non-existent.

Assumes gene tree equals species tree (coalescent stochasticity is ignored as is HGT/hybridization).

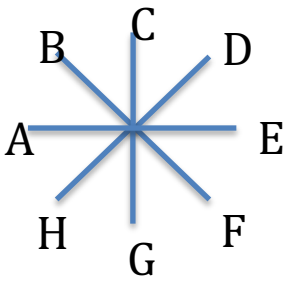
C. Experimental Phylogenetics – Building known phylogenies in the lab.

This approach combines the advantage of congruence analysis that DNA sequences evolve via more or less natural processes with the advantage of simulation studies that the tree topology can be anything the investigator chooses. In addition, the ancestors can be archived and used to assess other issues like accuracy of reconstructing ancestral states.

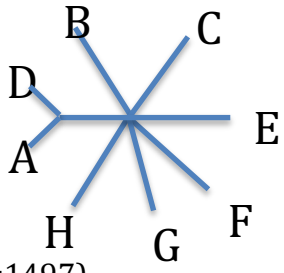


A True Experimental Phylogeny

We can also do things like expose experimental populations to various selective regimes to assess the effects of all kinds of evolutionary paradigms on phylogeny reconstruction.



Subject A & D to similar selection



Bull et al. (1997. Genetics. 147:1497)

So, we can adjust the true tree in any way we wish, and to some extent, we can alter the evolutionary process.

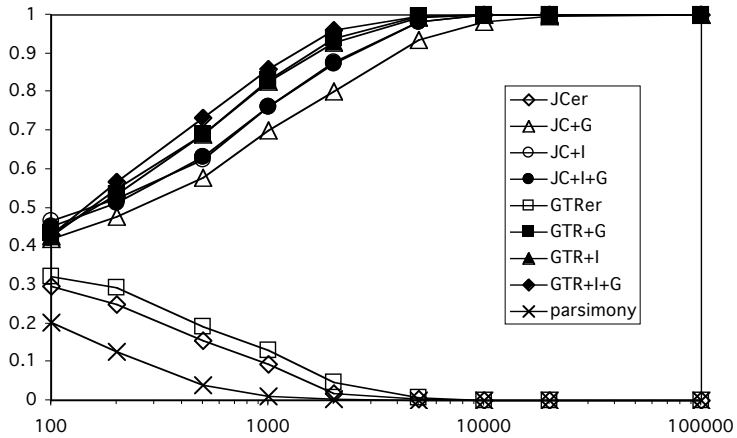
However, ability to replicate is quite low, usually limited to a couple or a few replicates.

II. Criteria.

There are a number of properties that are desirable in statistical estimation, and these can be used as criteria in assessing performance. Of interest may be the performance of any one method across a variety of conditions, the performance of a variety of methods under a particularly relevant set of condition, or, perhaps most importantly, the performance of several methods across a range of conditions.

A. Consistency – This term has a very explicit statistical definition. A statistically consistent estimator is one that converges to the true value of the parameter being estimated as the amount of data increases. With sufficient data, an estimate that is consistent will be equal to the true value with certainty.

In terms of phylogeny, this means that the true tree will be found with increasing probability as more data are acquired (i.e., longer sequences are analyzed). This is exemplified above.



Sequence Length

So, *under the conditions simulated here* (FZ tree, GTR+I+Γ, moderate ASRV), six of the methods examined are consistent and three are inconsistent.

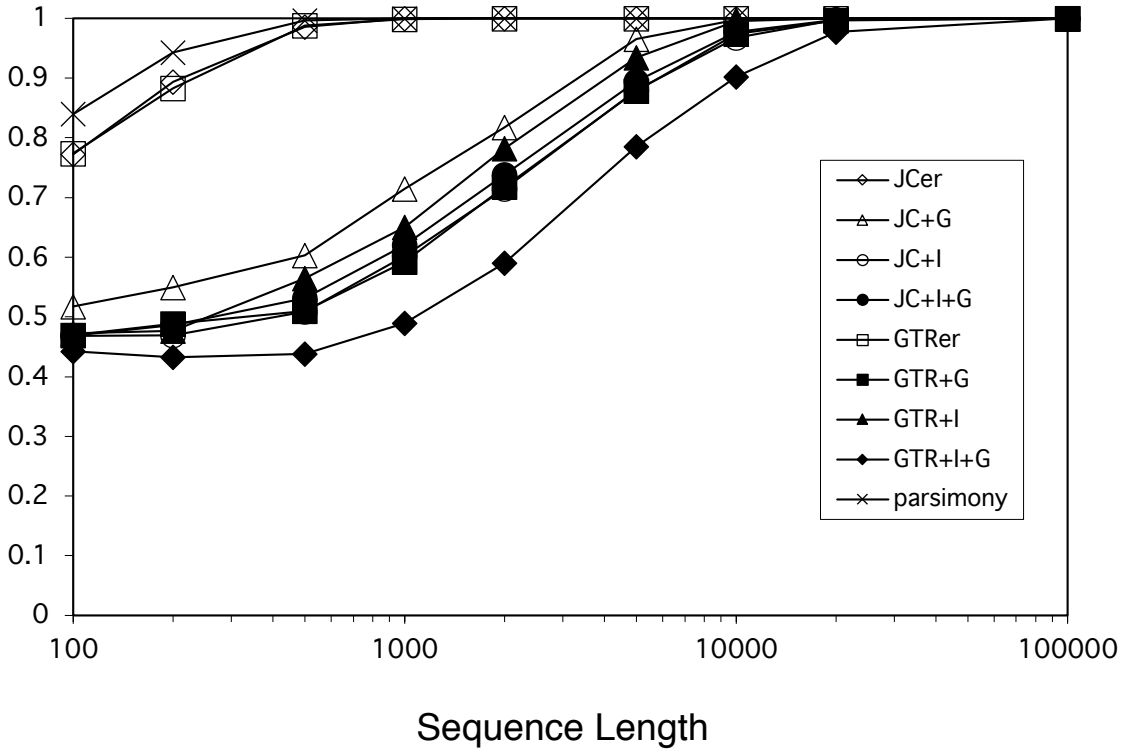
This certainly is an important and desirable property, and it has been used heavily as an indicator of performance. However, given that consistency is an infinite data property, it has been criticized as a criterion for choosing a method to analyze a finite data set.

A typical gene length is 500 – 5000 nucleotides.

B. Efficiency – This also has a specific statistical interpretation. A consistent estimator is an efficient one if it converges to the true value quickly, that is, without a lot of data.

This certainly is a relevant criterion for evaluating performance, but it's been pointed out, for example by Sanderson and Kim (2000), that it's rarely addressed. While they're correct that a large amount of work has been done on consistency, the efficiency of various methods can be examined at the same time as consistency. So, look at the consistency figure above, although ML estimation with both GTR+I+Γ and JC+Γ are consistent (for these conditions), estimation with GTR+I+Γ is more efficient.

Conversely, if we look at different conditions (i.e., a different tree), we see a different situation.

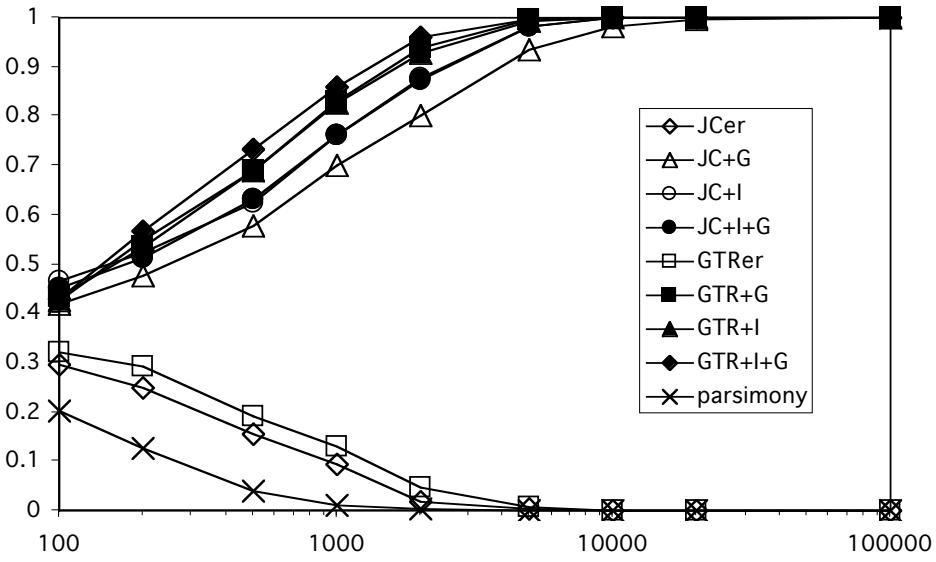


Parsimony and the equal-rates methods are much more efficient than are any of the other methods under these conditions

C. Robustness – A third criterion for evaluating the performance of a method is robustness. This assesses how well a method performs in the face of violations of its assumptions.

This is particularly important because all methods make assumptions, either explicitly or implicitly.

This really can be assessed via any of the three methods but is perhaps best addressed via simulation because we can control the exact nature of assumption violation. Again, we can use the same approach as before.



In this case, the sequences were simulated with the GTR+I+Γ and when we violate various assumptions in the model that we use to analyze the data consistency is not compromised, at least for the non-equal rates models.

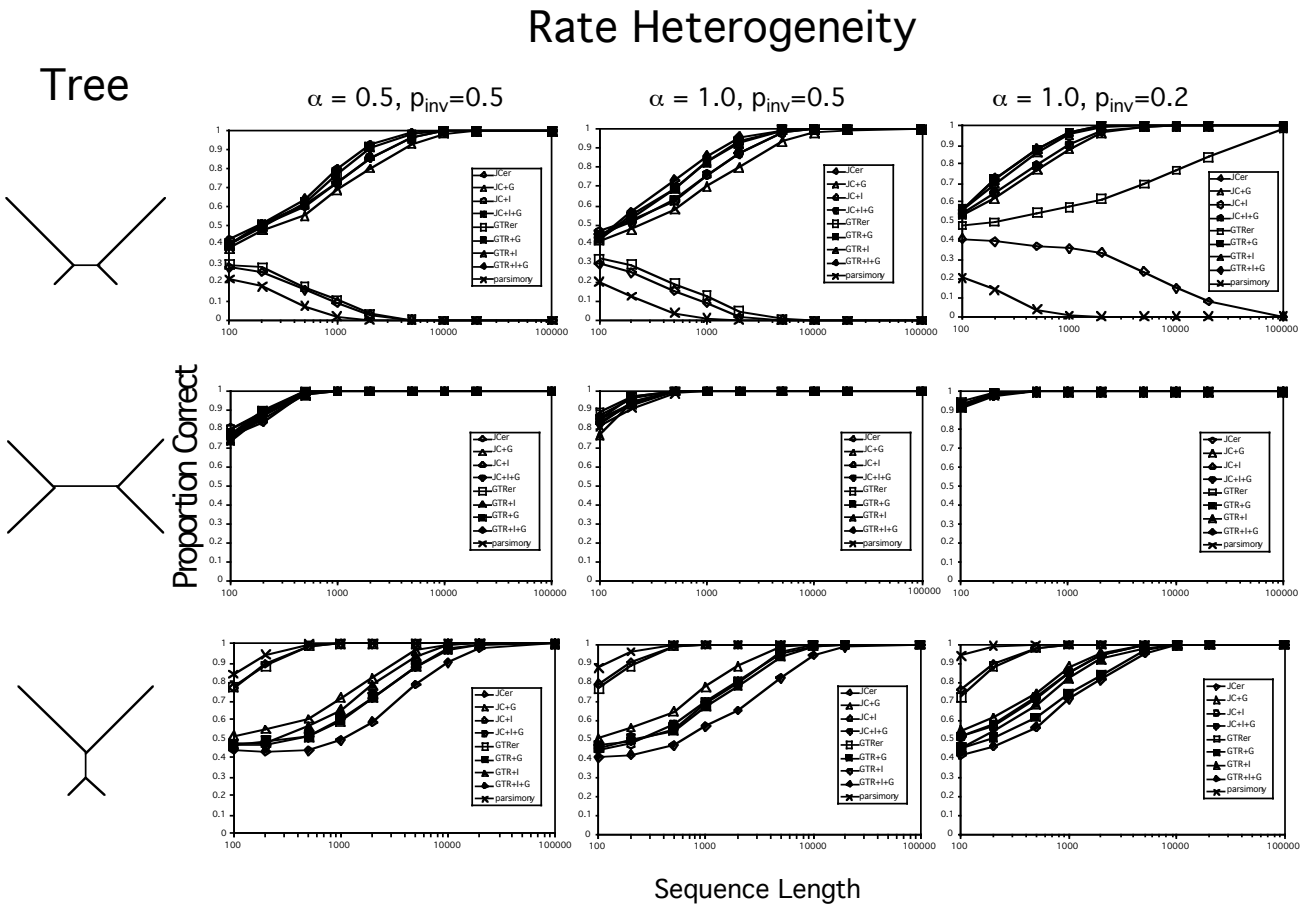
So, in this case, ML estimation is robust to violation of some of its assumptions – as long as we do something to account for ASRV, we don't need to model it precisely. However, violating these assumptions does influence efficiency.

One of the real advantages of simulations is that we can look for interactions between tree topology and each of these.

The figure above represents data simulated on a Felsenstein Zone tree. This, of course, stacks the deck against parsimony and the conclusions about robustness are restricted to trees with similar properties.

To be fair, we really need to assess the issues of consistency, efficiency and robustness across a variety of tree shapes.

From Sullivan & Swofford (2001. Syst. Biol. 50:723-729). 50:723-729



The effect of tree shape needs to be considered, when assessing these measures of performance.

The top row represents Felsenstein-zone trees, with long branches separated by a short internal branch; the middle row represents equal rates trees (what I like to call the Goldman Zone); and the bottom row represents inverse Felsenstein trees, with long branches on the same side of a short internal branch.

The columns represent different rate heterogeneity conditions. In all cases the data were simulated with a GTR+I+Γ model.

So, the effect of violating model assumptions varies across the shapes of the underlying true trees.

All methods appear to be robust to violations of assumptions when the underlying tree has equal branch lengths. They're all consistent and they all are quite efficient.

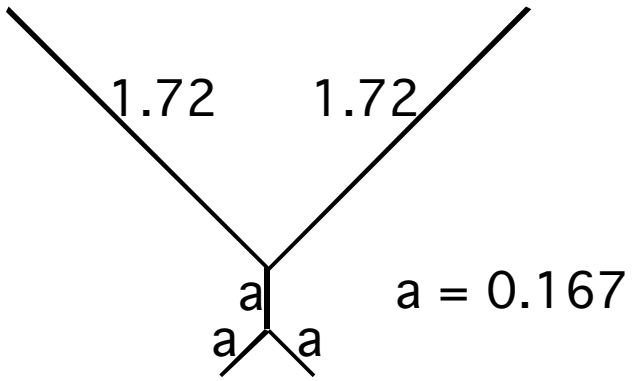
If the true tree is a Felsenstein-zone tree, the match between model and data becomes more important.

If the true tree is in the inverse Felsenstein zone, the most strongly violated models appear to be behaving the best.

Let's look at this situation more closely, because some have suggested that we may be able to use this phenomenon to our advantage.

Swofford et al. (2001. Syst. Biol., 50:525-539) examined the situation in detail.

So, for the following tree, we can calculate the probability that a site shared by the long branch taxa actually evolved on the internal branch and changed nowhere else.



This essentially represents the probability that any site pattern of the for *xxyy* (across the four taxa starting in the top right and moving clockwise) is the result of a true synapomorphy.

This is calculated under a JC model, which is what Sowfford et al. used.

Pr [True Synapormorphy] = Pr [No chane on long branches]

X Pr [No change on short terminals]

X Pr [Single change on internal branch]

≈ 0.0032

Conversely, the probability of the site pattern *xxyy* being seen in the data under any scenario is 0.1172 (This is the sum of the single-site likelihoods for all possible site patterns of the form *xxyy*).

Thus, ca. 97% of all sites that exhibit the pattern *xxyy* will be the result of multiple hits, not true synapomorphy. Analyses involving strong model violations will incorrectly interpret these (to a greater or lesser extent) as evidence favoring grouping the long-branch taxa.

It's probably better to use methods that fairly assess the support for a group than hoping that the bias inherent in your method that favors the true tree and not some alternative.

So, we can think about choice of methods in the same sense as we think about the importance of models. There are tree shapes that are easy to estimate (those with long internal branches) and tree shapes that are difficult to estimate. Thus, the choice of methods can be of trivial or critical importance, depending on the underlying shape of the true tree.