Assessing significance [JF 20]

Fredrik Ronquist

November 30, 2005

1 Introduction

With the exception of Bayesian analysis, phylogenetic inference procedures typically identify a best estimate of phylogenetic relationships, a so called *point estimate* of the phylogeny. However, the point estimate is often relatively uninteresting in itself unless we have some measure of its reliability. This lecture will be about techniques for examining the robustness or significance of the results of phylogenetic analysis. The techniques can be divided into non-parametric and parametric approaches. The non-parametric methods are quite general in that they can be combined with any approach to phylogenetic inference; the parametric approaches, however, rely on stochastic models. Five different methods will be discussed here, three of which are non-parametric (permutation tests, jackknifing and bootstrapping) and two of which are parametric (parametric bootstrapping and posterior predictive distributions).

2 Permutation tests

Permutation tests are widely used non-parametric techniques. They are particularly easy to apply in hypothesis-testing situations. Assume, for instance, that we have two samples of 50 data points each and want to examine the hypothesis that they both come from the same distribution. To use a permutation test to accomplish this, first calculate the difference between the means of the two samples. Then randomly permute all of the 100 data points and calculate the difference between the mean of the 50 first and the 50 last data points. Because of the permutation, each of these sets will consist of a random mix of data points from the first and the second sample. Repeat the permutation a large number of times, say 999 times, and calculate the difference between the

means of the first 50 and last 50 points for each permutation. If we add the observed difference between the means we now have 1,000 values, which we order from lowest to highest. If the actual difference lies in the bottom 25 or top 25, corresponding to the 0.025 and 0.975 percentiles, we can reject the null hypothesis with $\alpha = 0.05$.

A simple application of this approach to test for the presence of phylogenetic signal is the *permutation tail probability* test. Assume we have an aligned data set $X = \{x_1, x_2, x_3, ..., x_n\}$ where each x_i is a column vector representing a single site (or character) in the alignment. The idea is to generate a large number of permuted data sets, in which the elements of each column vector x_i have been permuted independently, and calculate some test statistic, such as the parsimony or likelihood score, for each. The distribution of values from the permuted data sets are then compared to the value of the test statistic for the observed data. If the latter value is sufficiently far into the appropriate tail of the reference distribution, we would say there is significant phylogenetic signal.

Permutations can also be used to test for the conflict between the phylogenetic signal in two partitions of a data set. Known as the *incongruence length difference* or *partition homogeneity* test, it represents a straight-forward phylogenetic extension of the two-sample permutation test described above. Instead of using the difference between the sample means, the test statistic in this case is typically chosen to be $s(D|T_D) - s(D_1|T_1) - s(D_1|T_1)$, where D is the entire data set, D_1 and D_2 are the two data partitions, T_i is the tree estimate for data set i, and $s(D|T)$ is the score for data set D on tree T. This test statistic can be used both for parsimony and for likelihood, using the logarithm of the likelihood scores in the latter case. Other test statistics are possible. For instance, one could use a tree distance metric to find the difference between the best trees for the two data partitions. This would be a very close analogue to the difference between sample means.

The reference distribution generated by permutation reflects some qualities of the data, such as the frequency of different character states, but it is less clear that its properties in general are such that the resulting statistical tests are efficient. For this reason, reference distributions generated by resampling of the original data are often used instead. Two techniques are used for resampling: jackknifing and bootstrapping.

3 Jackknifing

Jackknifing is the older of the two resampling techniques. It is based on generating a series of pseudosamples by deleting one or more data points in the original sample. The test statistic is calculated for each of the pseudosamples and then the distribution of these values is compared to

the value obtained from the original data.

A possible application of the jackknife to phylogenetic analysis has one site (character) at a time deleted from the aligned data matrix. The results from analysis of each these data sets, however, will be very similar to the original results, so relatively little new information is gained through this procedure. More commonly, phylogenetic jackknifing involves deletion of 0.50 or e^{-1} of the characters, resulting in behavior similar to the bootstrap. It is still unclear whether there is any reason to prefer the jackknife over the bootstrap; there is currently a general preference for the bootstrap, to a large extent because its properties are better known.

4 Bootstrapping

In bootstrapping (Fig. 4), the reference distribution is generated by drawing data points randomly from the original sample, with replacement, until a pseudoreplicate data set of the same size as the original data set is obtained. A large number of pseudoreplicate data sets are then used to derive the reference distribution for the test statistic. Each pseudoreplicate data set can be described as a weighted combination of the original data points. Some of the original data points are not represented at all, they have a weight of 0, while other data points are represented once or more than once, they have an integer weight of 1 or more.

In phylogenetic analysis, we would infer phylogeny based on each of the pseudoreplicate data sets resulting from bootstrapping (or jackknifing). The resulting set of trees would then be used to assess the uncertainty in the original point estimate of the phylogeny. The standard approach is to compute a majority rule consensus tree from all the trees resulting from the bootstrap analysis. The fraction of the bootstrap trees that contain a particular clade is a measure of the support for that clade; it is known as the *bootstrap proportion* (BP) of the clade.

Ideally, one would like to interpret the BP values as confidence intervals. That is, a BP value of 0.95 should indicate that the clade is correct 95 % of the time. However, this is not a valid interpretation. A better interpretation is that the probability is 0.05 $(1-0.95)$ that the clade would have this much bootstrap support given that it did not exist in the true tree. Even this interpretation is not correct, among other things because of the complexity of the problem of regions, of which the phylogeny bootstrap is an example.

To illustrate the problem of regions, consider a simple distribution and an associated true distribution of sample means for a particular sample size (Fig. 4, see JF Fig. 20:4). We are interested in

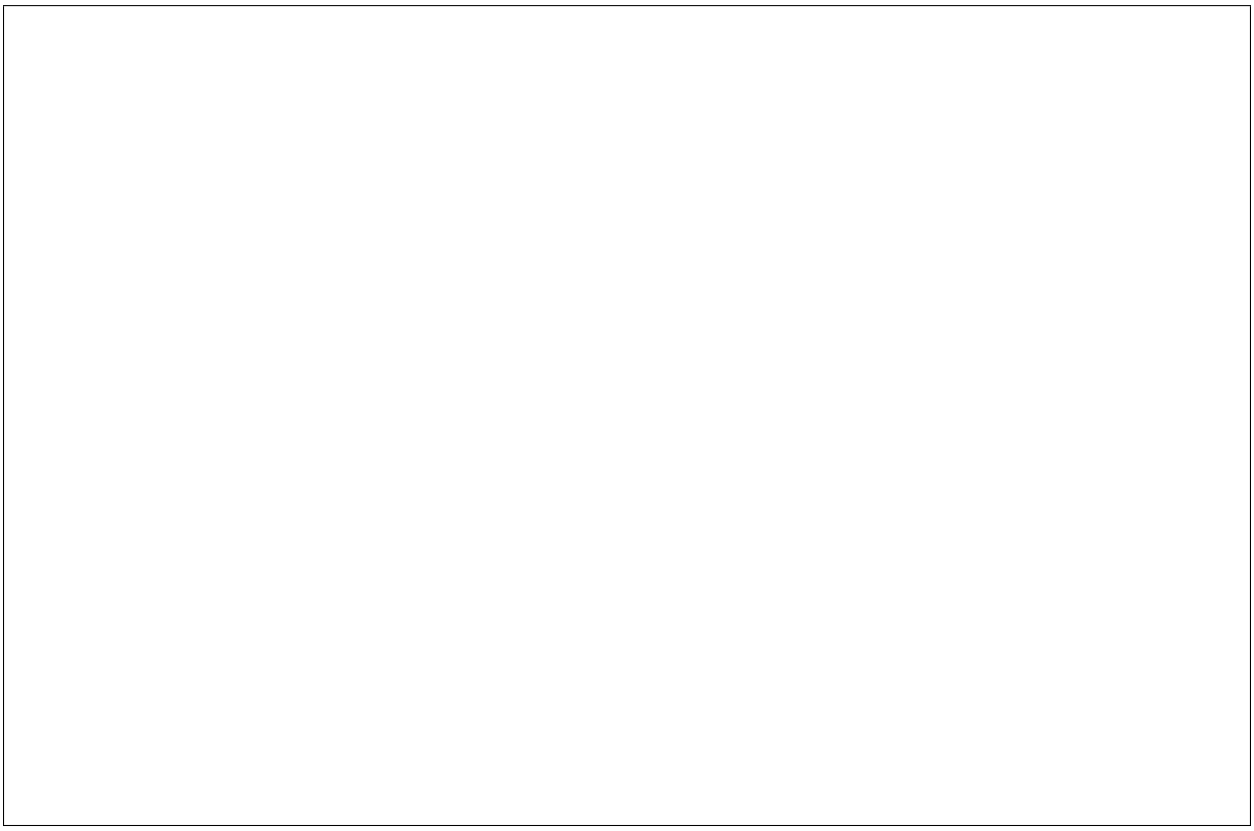

Figure 1: The bootstrapping procedure.

the hypothesis that the true value of the mean is larger than 0, that is, that the true mean is in the region above 0. From the true distribution of sample means, we can easily calculate the true probability, P, with which the sample mean would be in the right region. In reality, we don't have the true distribution of sample means but this distribution can be estimated using bootstrapping. Each estimated distribution of sample means will be centered around the observed mean of that particular sample. About half of the estimated distributions will overestimate P, the other half will underestimate P. However, because the first half will still be close to P whereas the second half will be much farther removed from the true P, the average estimated P value is going to be an underestimate of the true P value.

In phylogenetic inference, the problem of regions is even more complex because of the high dimensionality of tree space, apparently leading in general to underestimates of the true bootstrap proportions. Several approaches have been proposed to correct for this bias. Perhaps the most sophisticated one was described by Efron, Halloran and Holmes (1996). It is based on doing a second bootstrap analysis for each of the initial bootstrap samples. The procedure is fairly complex and computationally demanding and an exact description is outside of the scope of this chapter.

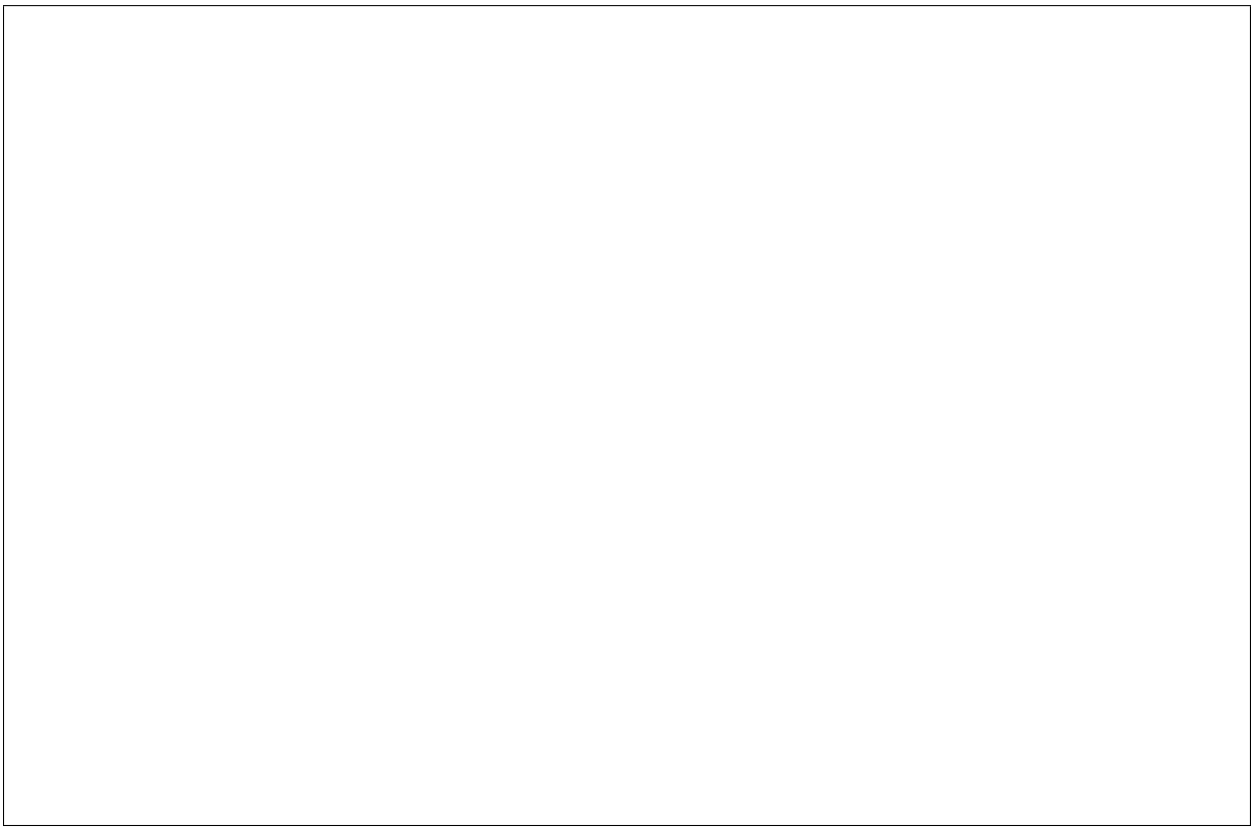

Figure 2: The problem of regions and the bias in bootstrap values.

5 Parametric bootstrapping

Parametric bootstrapping is completely different from standard bootstrapping, which is sometimes called non-parametric bootstrapping to make the distinction clear. As the name implies, parametric bootstrapping is a parametric technique for generating a reference distribution, which can be used to assess the significance or robustness of statistical results.

In phylogenetic analysis, parametric bootstrapping is used together with maximum likelihood methods. First, a typical maximum likelihood analysis is performed to find the maximum likelihood (profile likelihood) values of the parameters in the phylogenetic model (Fig. 5). We can now generate new data sets by simply simulating data on the maximum likelihood tree using the maximum likelihood estimates of the substitution model parameters until we have a new data set of the same size as the original data set. The procedure can be repeated a large number of times and the new data sets can be used pretty much in the same way as the pseudoreplicate data sets generated by resampling procedures.

Perhaps the most common application of parametric bootstrapping is in testing two hypotheses

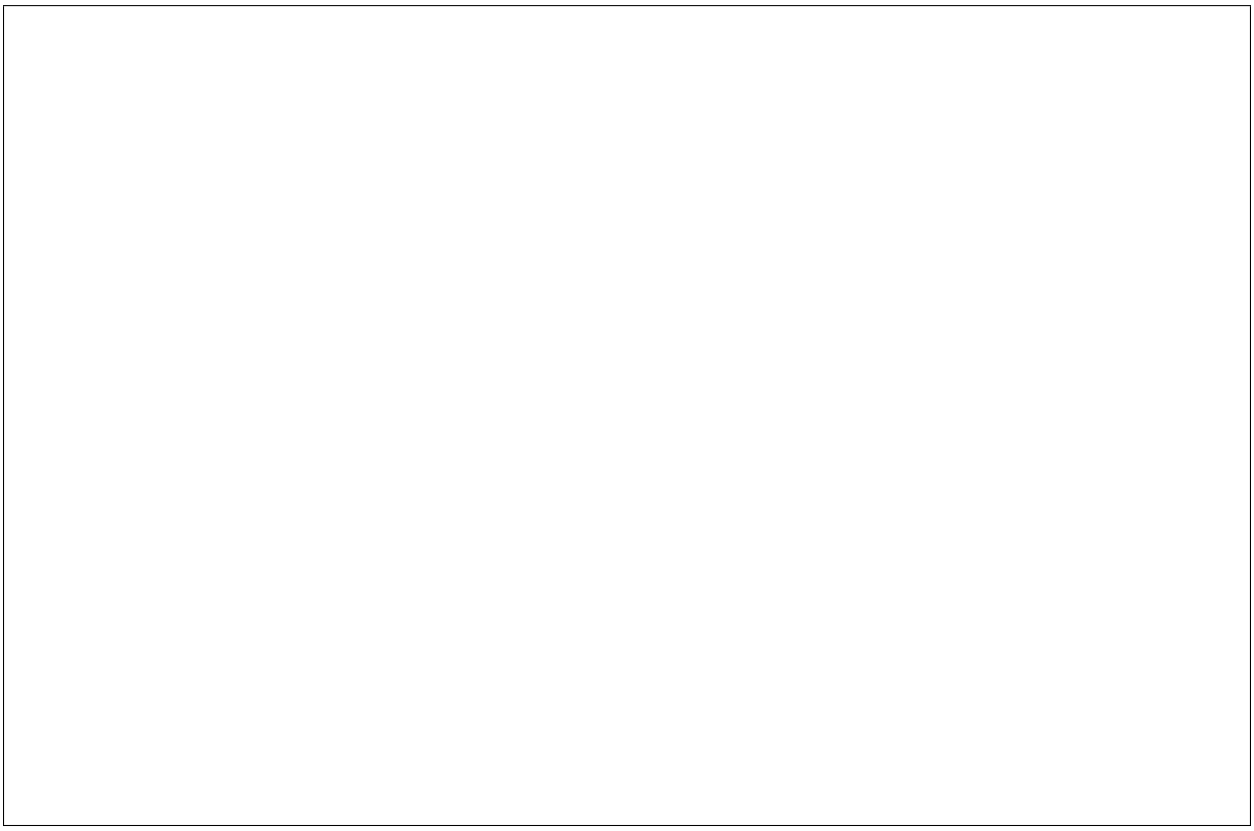

Figure 3: The parametric bootstrap.

against each other. Say, for instance, that we obtain tree T_1 in a phylogenetic analysis but were expecting T_2. Then we can test the null hypothesis that the data were actually generated on tree T_2 but that stochasticity resulted in T_1 being preferred in the particular data set we analyzed. We would then estimate all model parameters on T_2 and then generate a set of reference data sets using those values and parametric bootstrapping. On each of the generated data sets, we would measure the difference in likelihood score of the two trees, and finally we would use this reference distribution in evaluating whether the preference for T_1 in the original data set could be due to chance alone causing a deviation in data actually generated on tree T_2.

An intriguing advantage of parametric bootstrapping over its non-parametric namesake is that it assesses uncertainty based on information about the evolutionary process itself rather than an ad-hoc distribution of site patterns. This may be advantageous in many cases, for instance when the data set is small and just sampling columns may leave many kinds of variation in the data unrepresented. There are two major drawbacks to parametric bootstrapping, however. First, it relies on the model being accurate. Second, it ignores any uncertainty concerning the true values of the estimated parameters. The second problem can be easily addressed using posterior predictive distributions generated in a Bayesian context.

6 Posterior predictive distributions

Recall that the goal of Bayesian phylogenetic inference is to estimate the posterior probability distribution of the parameter values (θ) given the data (X), or $p(\theta|X)$. Assume we have used some estimation procedure to obtain a sample of N data points from this distribution, $\theta_1, \theta_2, ...\theta_N$. We can use this sample to generate new data sets from the posterior distribution, taking the uncertainty concerning the true values of θ into account. We simply pick a value at random from the sample and generate one data column from it. We then pick another value to generate the next data column, and continue in this fashion until a new data set of the same size as the original has been generated. The procedure is repeated until a sufficiently large set of new data matrices is generated.

In principle, these data sets can be used to examine the uncertainty concerning the parameter estimates in the analysis but this has already been obtained as part of the standard Bayesian analysis as the marginal distributions of each parameter. However, the posterior predictive distributions can be used to examine aspects of the model that were not sampled during analysis. For instance, posterior predictive distributions have been used to generate samples of character evolution histories (the ancestral states and change points of individual characters). Posterior predictive distributions can also be used to check that the model is reasonable. If so, the generated data sets should have similar properties to the original data set.

7 Study Questions

1. What is a permutation test?
2. Describe the partition homogeneity test.
3. What is the difference between jackknifing and bootstrapping?
4. Why is the bootstrap biased?
5. Compare non-parametric bootstrapping with parametric bootstrapping.
6. What is a posterior predictive distribution? What can it be used for?