Estimating Reliability of Phylogenetic tree

Inferring the ancestral relationship among organism is a routine exercise among the evolutionary biologists. There are several ways, through which we can make phylogenetic tree based on either distance or character based criteria. We know that there are several tree possible (rooted and unrooted) for N taxon. The challenging part is the reliability of tree topology, branching i.e. estimating who is parents of whome, and who the siblings. To solve this problem we take the help of statistic. There are two sampling techniques, through which we can access the reliability of internal branching of a tree.

Resampling is a way to reuse data to generate new, hypothetical samples (called *resamples*) that are representative of an underlying population. It's used when:

- You don't know the underlying distribution for the population,
- Traditional formulas are difficult or impossible to apply,
- As a substitute for traditional methods

Two popular tools are the bootstrap and jackknife, which are used in estimating reliability of phylogenetic tree although they have many similarities (e.g. they both can estimate precision for an estimator θ), they do have a few notable differences.

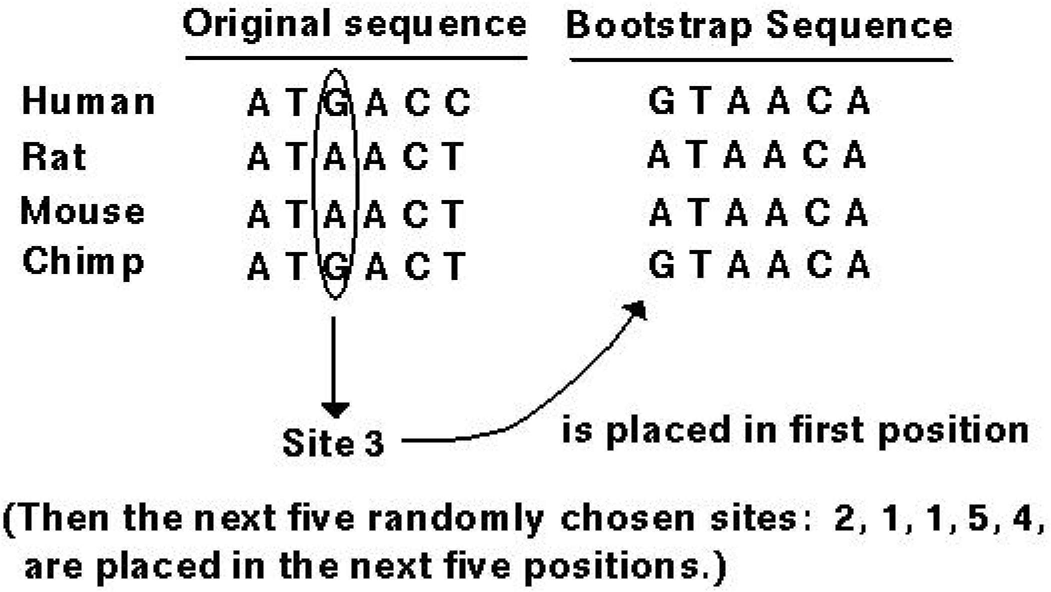
What is a Bootstrap?

Bootstrapping is the most popular resampling method today. It uses sampling with replacement to estimate the sampling distribution for a desired estimator. The main purpose for this particular method is to evaluate the variance of an estimator. It does have many other applications in bioinformatics and phylogeny.

In terms of your phylogenetic tree, the bootstrapping values indicate how many times out of 100 the same branch is observed when repeating the phylogenetic reconstruction on a re-sampled set of your data. If you get 100 out of 100 (and your data is sufficiently large to support this), we are quite sure that the observed branch is not due to a single extreme datapoint. if you recover the same node through 95 of 100 iterations of taking out one character and resampling your tree(which is done by the tree maker automatically), then you have a good idea that the node is well supported (your bootstrap value in that case would be 0.95 or 95%). Often if bootstrap values that are lower than (say 50%) are not included on a tree

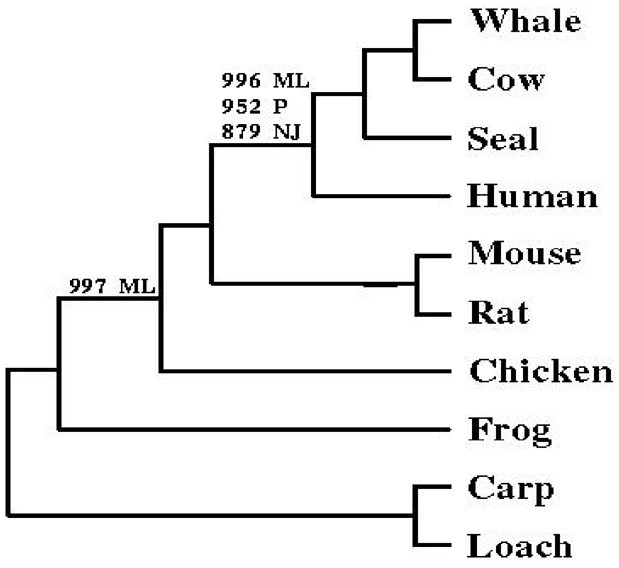
The bootstrap technique involves generating artificial sequences by randomly sampling sites from the original sequences with replacement. This randomly generated data set has the same sequence length but a slightly different composition (i.e some sites will be oversampled and others not).

For example, consider a simple sequence with 6 sites. Say that the first site chosen randomly is 3. For each species, site 3 is placed in the first position of the bootstrap sequence. This is repeated until the bootstrap sequence is also 6 bp long.



The "best" tree is then determined from the bootstrap sequences, using the same method as used with the original data set. This whole process is repeated at least 100 times. The number of times that a clade is seen among the bootstrap trees is reported. The more often a clade is present among the bootstrap trees, the more strongly the data support that clade, because the result is insensitive to which basepairs happen to be sampled.

Cummings et al (1995) used the entire mitochondrial genomes of ten vertebrates and obtained the following tree using parsimony (P), neighbor joining (NJ), and maximum likelihood (ML)



methods:

All clades were supported in 1000/1000 bootstrap data sets, with the exception of the two clades shown, which still had strong support

Jackknife:

The Jackknife works by sequentially deleting one observation in the data set, then recomputing the desired statistic. It is computationally simpler than bootstrapping, and more orderly (i.e. the procedural steps are the same repeatedly). This means that, unlike bootstrapping, it can theoretically be performed by hand. However, it's still fairly computationally intensive so although in the past it was common to use by-hand calculations, computers are normally used today. One area where it doesn't perform well for non-smooth statistics (like the median) and nonlinear (e.g. the correlation coefficient).

The main application for the Jackknife is to reduce bias and evaluate variance for an estimator. It can also be used to:

- Find the standard error of a statistic,
- Estimate precision for an estimator θ.

Main Differences between Jackknife and Bootstrap

To sum up the differences, Brian Caffo offers this great analogy: "*As its name suggests, the jackknife is a small, handy tool; in contrast to the bootstrap, which is then the moral equivalent of a giant workshop full of tools."*

Some specific differences:

- The bootstrap requires a computer and is about ten times more computationally intensive. The Jackknife can (at least, theoretically) be performed by hand.
- The bootstrap is conceptually simpler than the Jackknife. The Jackknife requires *n* repetitions for a sample of n (for example, if you have 10,000 items then you'll have 10,000 repetitions), while the bootstrap requires "B" repetitions. This leads to a choice of B, which isn't always an easy task. A general rule of thumb is that B = 1,000 unless you have access to a large amount of computing power.
- In most cases, the Jackknife doesn't perform as well the Bootstrap.
- The Jackknife is more conservative than bootstrapping, producing slightly larger estimated standard errors.
- The Jackknife gives the same results every time, because of the small differences between replications. The bootstrap gives different results each time that it's run.
- The Jackknife is more suitable for small original data samples**.**