

Evaluating trees, pt.2

MEP: 11 Nov 2004

For phylogenetic trees:

- ◆ Accuracy: proximity to the true tree (P_T)
 - ◆ the true tree may have 395 steps, but there may be 15 MP trees with this value
- ◆ Precision: how many alternative trees are excluded
 - ◆ $\ln L = -6850.52$ is very precise, but the ML tree may not be the true tree

Types of error in phylogenetic reconstruction:

- ◆ sampling error, associated with estimates based on the given data set (v. sensitive to amount of data); variance associated with parameter estimates; due to stochastic processes
- ◆ systematic error, where assumptions of model are inaccurate, misleading the interpretation (e.g., parsimony in the Felsenstein zone; not using a gamma)

Interior branch test

- ◆ When the length of an interior branch is greater than 0, we regard the reality of the branch as established
- ◆ A unique property of trees made by distance methods
- ◆ We can also plot log likelihood (for the tree) against a range of values for an interior branch - this will give us an approximate standard deviation for the estimate of the branch length

Interior branch tests, cont.

- ◆ For the true tree, we expect all interior branches to be > 0
- ◆ For incorrect topologies, at least one interior branch must be ≤ 0 ; see Figure 6.5
- ◆ This seems to be true when an unbiased distance estimate is used and branch lengths are estimated by LS method
- ◆ The test for all interior branches for $a \leq 0$ is conducted with a one-tailed test of the normal deviate $Z=a/s(a)$, where $s(a)$ is the standard error of the estimate of branch length a

Interior branch tests, cont.

To test for monophyly of taxa 1 & 2 in Figure 9.2A:

- ◆ Test for branch $b \leq 0$; $Z=b/s(b)$

Analytical interior branch test

- ◆ Only for distance approaches, calculates a confidence probability (P_C) based on the standard error of the branch length estimate, computed by the Z test
- ◆ Only reliable for closely related sequences - very complicated evolutionary processes increase the standard error estimation

Bootstrap interior branch test

- ◆ Data are bootstrapped with replacement
- ◆ For the original topology, branch lengths are estimated for each bootstrap pseudoreplicate
- ◆ The length estimate for each branch will vary among reps and may be negative
- ◆ The mean and standard error of each branch length estimate can be computed and a Z test used to test if ≤ 0
- ◆ Nei implies that this can be used with non distance methods

Ordinary bootstrap

- ◆ Reiterative resampling with replacement
- ◆ Technically, a “pseudoreplication” of the original sample; usually, hundreds or thousands are made
- ◆ Each bootstrap pseudoreplicate of the data is analyzed with the same method
- ◆ The percentage of times a branch on the original tree occurs in the bootstrap pool is recorded; alternatively, a consensus bootstrap tree can be computed
- ◆ The estimates of branch lengths from each pseudoreplicate is used to compute error values around branch length estimates on the original tree

Bootstrap in PAUP

- ◆ `Bootstrap [options] [/search options]`
- ◆ `NReps=integer value`
- ◆ `Search=Heuristic | BandB | FastStep | NJ | UPGMA`
- ◆ `ConLevel=integer value`
- ◆ `Treefile=filename`

Statistical properties of the bootstrap test for a NJ tree

IF:

1. each site evolved independently,
 2. the distance measure is an unbiased estimator,
 3. n and m are sufficiently large...
- ... then the null hypothesis of the BS test is that the length of each interior branch is 0 and P_B for a branch is supposed to measure the probability (P_C) of the branch length being greater than 0

Statistical properties of the bootstrap test for MP & ML trees

- ◆ Since branch lengths are never 0, interpreting the statistical significance of the BS test is more difficult
- ◆ Simulation studies suggest that P_B for MP trees is the same as that for NJ trees

Bootstrap considerations

- ◆ BS values tend to be conservative in all uses
- ◆ Typically, researchers accept the validity of branches that have 70-90+ bootstrap support
- ◆ It is very important to remember that the BS value reflects whether the data support a given branch, not whether the supported branch is true.

Parametric bootstrap

- ◆ A cross between bootstrapping and data simulation
- ◆ A tree is constructed from the original data set and parameters are estimated for those data (π 's, r 's, Γ , I , etc) - we assume this tree is somewhat near the true tree and the model is somewhat near the true model
- ◆ Some number, R , data sets are created
- ◆ The data sets are treated as bootstrap reps
- ◆ More useful for small data sets (when ordinary bootstrapping may be insufficient), but is entirely dependent on the quality of the model chosen - because of this, it is infrequently used

Permutation tests

- ◆ Technically, a reordering of the data
- ◆ For phylogenetics, the permutation tail probability test (PTP) shuffles the data matrix in each column (sites), with each site being shuffled independently of other sites
- ◆ Resulting data sets should have no phylogenetic data, but should have characteristics typical of the original data set
- ◆ The tree score (likelihood, parsimony, ME, etc) for the original tree is compared to the distribution of scores of the permutation data sets
- ◆ Test: does a given topology have support greater than random?

Other tests: ME trees

- ◆ For ME trees, we can compute $D = S_B - S_A$ between two topologies A & B
- ◆ this is equivalent to the test of the lengths of the interior branches at which the two topologies are different
- ◆ the test is two-tailed and computed by the Z test using a bootstrap method to determine the standard error of D

Other tests: ML trees

- ◆ We can't conduct a LRT between trees because there are 0 degrees of freedom for the X^2 test
- ◆ Other likelihood-based tests have been proposed, such as the paired-sites tests, where 2 trees can be compared by examining the difference in likelihood per site
 - ◆ the appropriate comparison stat is not known
 - ◆ the relationship to BS & LRT is not known

Other tests: MP

- ◆ Bremer support values, "consistency"
- ◆ difference in TL between two topologies when one branch is changed