

Exploring, comparing and evaluating trees, pt.1

MEP: 9 Nov 2004

Distance based methods

- ◆ Calculate pairwise distances according to some metric that corrects for multiple substitutions and various types of biases
- ◆ Find a tree:
 - ◆ construct a tree using NJ or other method
 - ◆ search tree space and select the tree according to LS or ME criteria

Distance: objections

- ◆ If distance calculations are poor (biased), the resulting tree and parameter estimates may be adversely affected
- ◆ For construction methods, it's hard to know if there are more than one optimal trees - and what nearly optimal trees look like
- ◆ Summarizing sequences into a distance matrix loses information - e.g., we cannot trace the evolution of specific sites
- ◆ Branch lengths may not be interpretable - e.g., what is 0.5 substitutions? What does it mean when the distance branch length is smaller than the number of observed sub's

Parsimony methods

- ◆ For a given tree, count the number of substitutions observed
- ◆ Find the tree(s) with the fewest number of observed substitutions
- ◆ Make no explicit assumptions about the evolutionary process for the data set

Parsimony: objections

- ◆ The model is implicit - change is rare, all changes are equally probable
- ◆ It is not consistent: in certain circumstances (long branches) it will find the wrong tree even when more data are added
- ◆ Branch length estimates are unreliable

Likelihood methods

- ◆ For a given tree, we calculate the likelihood of a model give the observed data
- ◆ Find the tree(s) with the highest likelihood value
- ◆ Incorporates a specific model of evolution
- ◆ Permits robust statistical tests of evolutionary hypotheses

Likelihood: objections

- ◆ Takes too long to search all of tree space; as a result, good trees are often undiscovered
- ◆ Requires a model to find the best tree: to find a good model, we need a tree...
- ◆ We can search tree & model space for the combination that maximizes likelihood, but this is far too computational intensive
- ◆ More than one likelihood value can exist for a given tree (when different computational approaches are used) - is a given tree really maximal?
- ◆ When a tree is selected as ML, what do we know about the "nearly ML" trees? (For MP, we know they are slightly longer.)

Bayesian methods

- ◆ For a given tree + parameter set, calculate the probability of observing the sequence data
- ◆ Search tree + parameter space for the combination that maximizes the probability
- ◆ A prior distribution for the model parameters is required
- ◆ Instead of selecting the tree & model parameter values based on ML, the one with the largest area under the curve is selected

Bayesian: objections

- ◆ Felsenstein: “A Bayesian is defined, not by using a prior, but by being willing to use a controversial prior.”
 - ◆ the author and reader of a paper may have different ideas about what the prior should be - leading to different results: the likelihoodist lets the reader supply their own prior

Bayesian: objections, cont.

- ◆ Using a flat (“uninformative”) prior may not really be flat:
 - ◆ using a JC model, the probability of change (p-distance) can range from 0-3/4 - we can make this flat w/ respect to frequency
 - ◆ the distribution of branch lengths resulting from these distance is not flat!
- ◆ (The likelihood value calculated for the same 2 parameters is the same.)

Bayesian: objections, cont.

- ◆ We can't put a flat prior on d (the branch length) because it is unbounded: ranges from 0 to ∞ (the density would be 0.0 everywhere)
- ◆ To account for this, we can truncate the d value, but doing so affects the estimate (and prior) of p :
 - ◆ the higher the truncation value, the more p is affected
 - ◆ very high truncations will eliminate the true value from the Bayesian 95% credible interval

Criteria used to compare tree reconstructed by different methods:

- ◆ Computational speed
- ◆ Consistency as an estimator
- ◆ Statistical tests of the trees reconstructed
- ◆ Probability of obtaining the true tree
- ◆ Reliability of branch length estimates

Tests need to be done head-to-head...

Computational speed:

- ◆ UPGMA and NJ are fast even to $m=500$ as an approximation of the ME tree
- ◆ Tree search methods are slow when all trees are examined

Consistent estimation

- ◆ A consistent estimator is one that gives the correct tree as n approaches ∞
- ◆ MP is sometimes inconsistent; others are consistent when unbiased estimators (models) are used
- ◆ In practice, only hundreds or thousands of sites are used - all methods may fail

Statistical tests

- ◆ [We'll talk about this on Thursday.]

Reliability of branch length estimates

- ◆ [We'll talk about this on Thursday.]

Probability of obtaining the true topology

- ◆ Requires knowledge of the true tree, which is rarely available, but it can also be done for simulated data and in theoretical studies
- ◆ These provide us with knowledge about when reconstruction models fail and provide an estimate of P_T for various methods and conditions

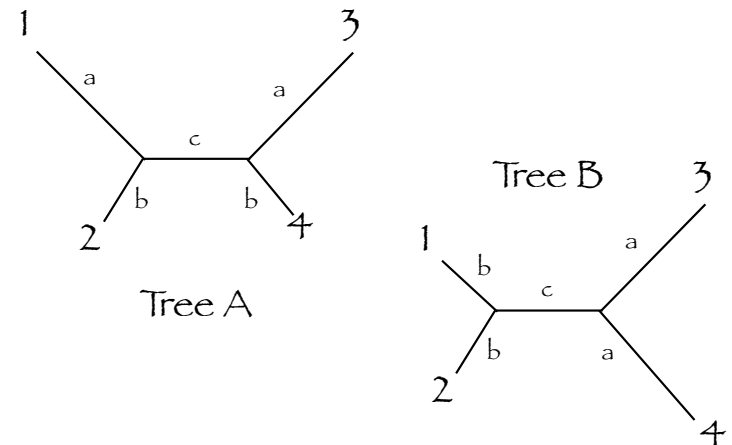
Four ways to compare trees created by different methods

- ◆ Analytical, theoretical studies: establish mathematical conditions under which a given approach will fail (e.g.: UPGMA when there is no molecular clock; MP when some branch lengths are especially long)
- ◆ For $m=4$, P_T for $NJ > MP, ME > LS$

Four ways to compare...

- ◆ Simulation: from a starting sequence, we can evolve a set of sequences computationally, using a given model
- ◆ We know the edge taxa to be compared and also the ancestors at each node
- ◆ We can explore the effects of a wide variety of parameters on the performance of various methods of reconstruction
- ◆ The chosen model may be a bit artificial - models for reconstructions are assumed to be mere shadows of actual evolutionary processes. In fact, the model chosen may bias results towards methods using that model implicitly

Simulation: 4 taxa



$$P_T \times 100$$

Tree A

Tree B

n	Tree A					Tree B				
	NJ:p	NJ:k2p	MP:UW	MP:W	ML	NJ:p	NJ:k2p	MP:UW	MP:W	ML
100	44	72	47	64	76	98	74	88	96	64
200	41	81	52	80	84	100	82	97	99	76
300	43	88	59	80	92	100	86	98	100	82
500	35	95	62	89	97	100	94	100	100	90
800	29	96	63	94	98	100	96	100	100	94
1000	35	99	66	98	100	100	99	100	100	96

Simulation: 6+ taxa

- ◆ P_T becomes very small for any method unless a large number of nucleotides are used.
- ◆ Typically, $ML \geq NJ > MP$, although differences are quite small when biologically reasonable models trees are considered; most variation found when rates vary drastically among lineages
- ◆ Values of n is the most important factor - when it's small, reliable trees are hard to find
- ◆ Strangely, the best method found varies depending on the bias of the person conducting the study

Four ways to compare...

- ◆ Known phylogenies: ability to reconstruct a "real" tree of known origin, such as from a bacteriophage T7 mutated in the lab
- ◆ We have the actual ancestors, so we can evaluate how well a given method finds the tree and infers the ancestral sequence
- ◆ Generally, creates only trees w/ little evolution - all methods work well...
- ◆ Also, note the lack of replication...

Four ways to compare...

- ◆ Congruence: compare estimates of phylogeny based on different characters. If the data sets are independent, the chance of obtaining the same or similar tree by chance is small, since the number of trees is huge.
- ◆ Assumes the same evolutionary history, of course
- ◆ Can be used to validate a method of inference: a method that consistently finds the same tree using different data sets is preferred over a method that finds different trees
- ◆ Can be used to validate a new source of data: come trees from that source with trees from other sources

true tree & congruence example

- ◆ when 13 mtDNA genes are used, every method finds the known phylogeny of vertebrate species (Nei Figure 9.4)
- ◆ Some genes alone always found the true (or nearly true) topology irrespective of method
- ◆ Some genes alone always found an incorrect tree, regardless of the method used
- ◆ Sophisticated methods (ML w/ complex model) were no better than simpler methods (NJ w/ ρ)
- ◆ The addition of a couple more taxa made all methods with all (or any) genes find the wrong tree - unless a gamma parameter is used...