Bootstrapping and Tree reliability

Biol4230 Tues, March 13, 2018

Bill Pearson wrp@virginia.edu 4-2818 Pinn 6-057

- Rooting trees (outgroups)
- Bootstrapping
  - given a set of sequences
  - sample positions randomly, with replacement
  - build trees (using distance, ML, or parsimony)
  - compare trees with `consens`
- Tree reliability
  - pathological situations - the "Felsenstein zone"
  - performance with different methods (Distance, ML, MP)
  - performance with different rate models
  - performance on real data

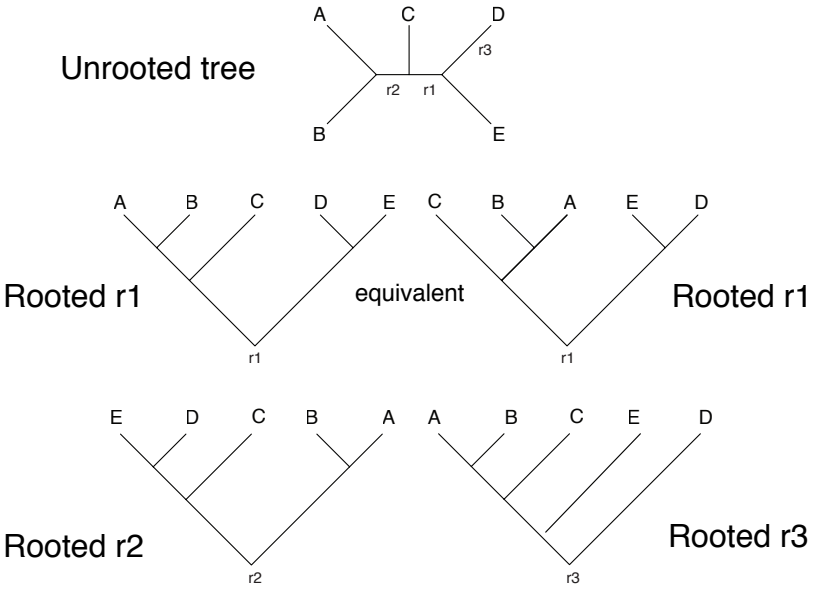
To learn more:

- Pevsner, Ch. 7, pp. 266-269
- Hillis, D. M., Allard, M. W., and Miyamoto, M. M. (1993) Analysis of DNA sequence data: phylogenetic inference. Meth. Enzymol. 224:456-487.
- Hillis, D. M., Heulsenbeck, J. P., and Cunningham, C. W. (1994) Application and accuracy of molecular phylogenies. Science 264:671-677.
- Hillis, D. M., Huelsenbeck, J. P., and Swofford, D. L. (1994) Hobgoblin of phylogenetics. Nature 369:363-364.
- Medina, M., Collins, A. G., Silberman, J. D., and Sogin, M. L. (2001) Evaluating hypotheses of basal animal phylogeny using complete sequences of large and small subunit rRNA.. Proc Natl Acad Sci U S A 98:9707-9712.

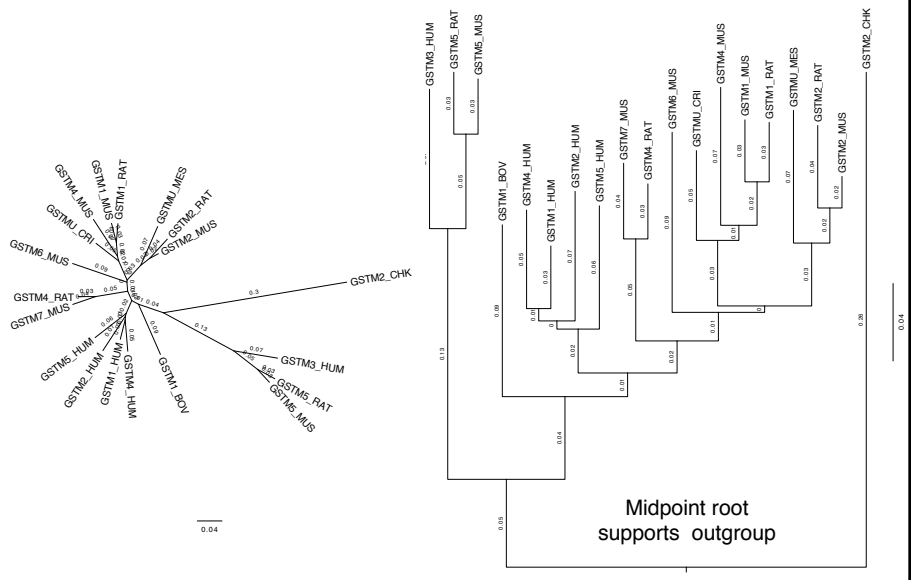
Rooting a tree: outgroups and midpoints

- Molecular sequence data is "time-reversible"
  - A->G or G->A, no way to tell
- Trees based on sequence data *only* are *unrooted*
- The root of the tree specifies a direction from past to present
  - Mid-point rooting: put the root between the most distant taxa
  - Outgroup rooting: use a known distant homolog to specify the root (chickens vs mammals; must be orthologs)

Trees – Rooted and UnRooted

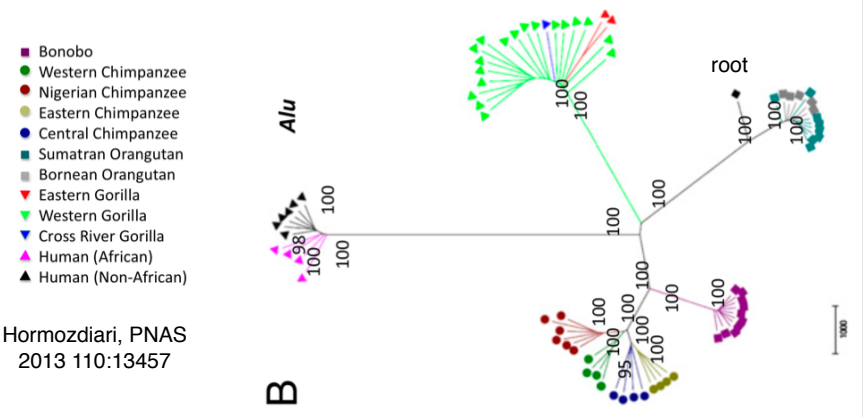


Trees – Rooted and Un-Rooted



Trees – Rooted and UnRooted

- Sequence data is reversible – additional data for root (mid-point implies clock-like tree)
- Some data (Alu repeat insertions) is less reversible – insertions go in and stay



Hormozdiari, PNAS
2013 110:13457

Bootstrapping and Tree reliability

Evolutionary tree reliability:

- Trees describe events in the past. They cannot be confirmed for real data
  - simulations guarantee "correct" answer, but do they simulate biology?
- Tree space is enormous, and tree finding methods tend to find similar trees
  - are there "almost as good" trees that are very different topologically
- Do some methods prefer certain kinds of trees?
  - long branch attraction

Building (and evaluating) evolutionary trees

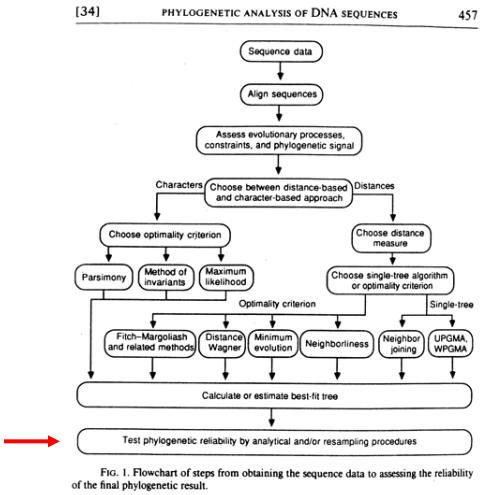


FIG. 1. Flowchart of steps from obtaining the sequence data to assessing the reliability of the final phylogenetic result.

 Hillis (1993) Meth. Enz. 224:456-487.

Estimating *true* phylogenies with Bootstraps

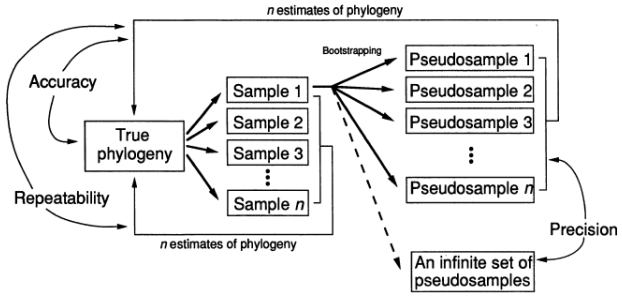


FIGURE 1. The relationships among a true phylogeny, samples of characters drawn from the taxa, bootstrap pseudosamples drawn from an initial sample, and the concepts of precision, accuracy, and repeatability.

Bootstraps introduced by Felsenstein (1985) to estimate the "repeatability" (not "accuracy") of a tree.

- precision: do bootstrap proportions from N bootstraps represent *all* bootstraps (not correct phylogeny)
- repeatability: do bootstrap proportions represent what would happen with more (independent) data
- accuracy: does the data (or the bootstraps) induce the correct phylogeny

Hillis (1993) Syst. Biol. 42:182-192 fasta.bioch.virginia.edu/biol4230 9

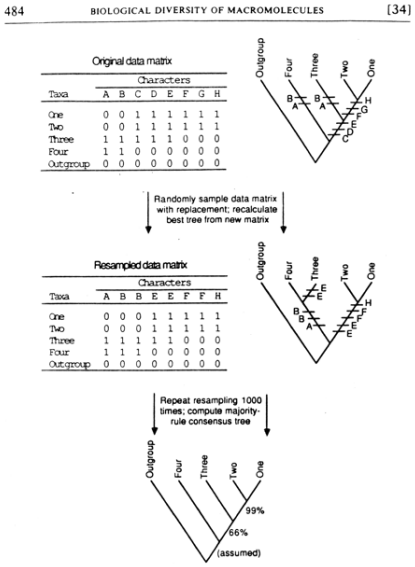


FIG. 7. Bootstrap analysis among characters in a parsimony analysis. The tree to the right of each matrix is the most parsimonious tree for that matrix. The final results of the bootstrap analysis are shown in the tree at the bottom. The number of times each branch was supported in the bootstrap replication is shown as a percentage. Outgroup rooting carries the assumption of ingroup monophyly, so no confidence interval can be assigned to the branch that unites the ingroup.

Hillis (1993) Meth. Enz. 224:456-487.

fasta.bioch.virginia.edu/biol4230 10

Bootstrapping with PHYLIP

```
wrpmpb 22% seqboot
seqboot: can't find input file "infile"
Please enter a new file name> gstm.phy_n

Bootstrapping algorithm, version 3.63

Settings for this run:
  D      Sequence, Morph, Rest., Gene Freqs?  Molecular sequences
  J  Bootstrap, Jackknife, Permute, Rewrite?  Bootstrap
  %    Regular or altered sampling fraction?  regular
  B      Block size for block-bootstrapping?  1 (regular bootstrap)
  R                    How many replicates?  100
  W              Read weights of char acters?  No
  C                Read categories of sites?  No
  S     Write out data sets or just weights?  Data sets
  I             Input sequences interleaved?  Yes
  0      Terminal type (IBM PC, ANSI, none)?  ANSI
  1       Print out the data at start of run  No
  2     Print indications of progress of run  Yes

  Y to accept these or type the letter for one to change
y
Random number seed (must be odd)?
12345
```

Produces 100 randomly sampled datasets for dnadist/fitch, dnapars, dnaml

Bootstrapping with PHYLIP

```
franklin: 1 $ fseqboot -help
   Standard (Mandatory) qualifiers:
  [-sequence]          seqset     (Aligned) sequence set filename and optional
                                  format, or reference (input USA)
  [-outfile]           outfile    [*.fseqboot] Phylip seqboot_seq program output file

Additional (Optional) qualifiers (* if not always prompted):
   -categories         properties File of input categories
   -test               menu       [b] Choose test (Values: b (Bootstrap); j
                                  (Jackknife); c (Permute species for each
                                  character); o (Permute character order); s
                                  (Permute within species); r (Rewrite data))
*  -seqtype            menu       [d] Output format (Values: d (dna); p
                                  (protein); r (rna))
*  -blocksize          integer    [1] Block size for bootstraping (Integer 1
                                  or more)
•   -reps               integer    [100] How many replicates (Integer 1 or more)

*  -seed               integer    [1] Random number seed between 1 and 32767
                                  (must be odd) (Integer from 1 to 32767)
```

Produces 100 randomly sampled datasets for dnadist/fitch, dnapars, dnaml

Bootstrapping with PHYLIP

```
franklin: 2 $ fseqboot -sequence gstm.n_phy -outfile gstm.n_boot_phy -seed 54321 -reps
100
Bootstrapped sequences algorithm
Warning: integer value out of range 54321 more than (reset to) 32767

 bootstrap: true
jackknife: false
 permute: false
 lockhart: false
 ild: false
 justwts: false

completed replicate number   10
completed replicate number   20
...
completed replicate number   90
completed replicate number  100

Output written to file "gstm.n_boot_phy"

Done.
```

Produces 100 randomly sampled datasets for dnadist/fitch, dnapars, dnaml

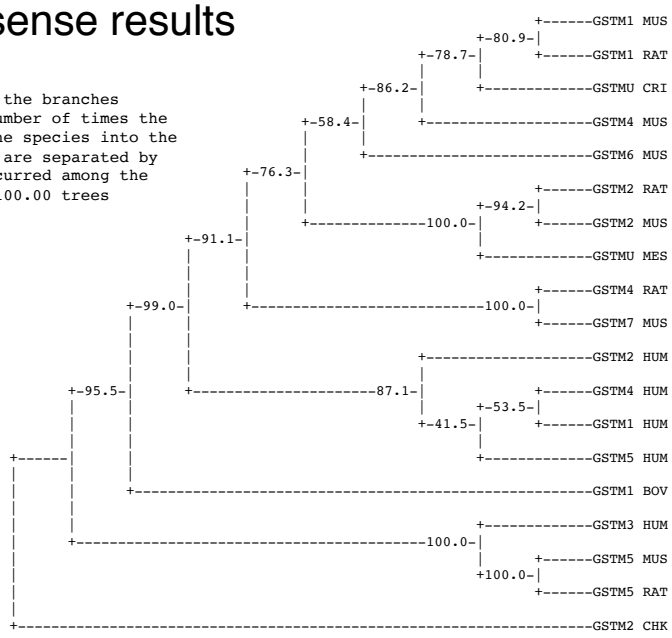
Consense results

```
Consensus tree program, version 3.68

Species in order:
                        Sets included in the consensus tree
  1. GSTM2 CHK
  2. GSTM3 HUM          Set (species in order)     How many times out of  100.00
  3. GSTM5 RAT
  4. GSTM5 MUS          .....**... .........        100.00
  5. GSTM1 BOV          .......... ..***....        100.00
  6. GSTM7 MUS          .***...... .........        100.00
  7. GSTM4 RAT          ..**...... .........        100.00
  8. GSTM6 MUS          .....***** *********        99.00
  9. GSTM4 MUS          ....****** *********        95.50
  10. GSTMU CRI         .......... ...**....        94.18
                        .....***** *****....        91.10
  11. GSTM1 MUS         .......... .....****        87.05
  12. GSTM1 RAT         ........** **.......        86.17
  13. GSTMU MES         .......... **.......        80.88
  14. GSTM2 RAT         .........* **.......        78.67
  15. GSTM2 MUS         .......*** *****....        76.31
  16. GSTM2 HUM         .......*** **.......        58.37
  17. GSTM5 HUM         .......... .......**        53.48
  18. GSTM4 HUM         .......... ......***        41.47
  19. GSTM1 HUM
```

Consense results

CONSENSUS TREE:
the numbers on the branches
indicate the number of times the
partition of the species into the
two sets which are separated by
that branch occurred among the
trees, out of 100.00 trees



Estimating *true* phylogenies with Bootstraps

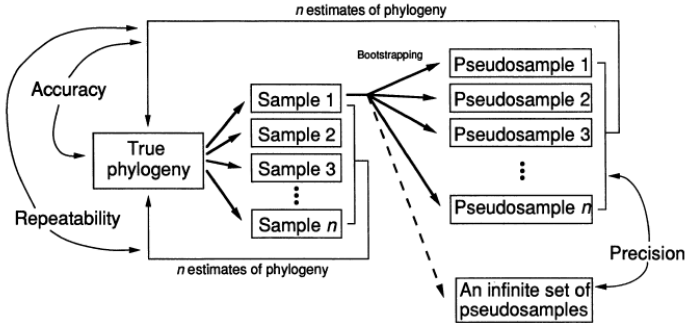


FIGURE 1. The relationships among a true phylogeny, samples of characters drawn from the taxa, bootstrap pseudosamples drawn from an initial sample, and the concepts of precision, accuracy, and repeatability.

- precision: do bootstrap proportions from N bootstraps represent *all* bootstraps (not correct phylogeny)
- repeatability: do bootstrap proportions represent what would happen with more (independent) data
- accuracy: does the data (or the bootstraps) induce the correct phylogeny

Hillis (1993) Syst. Biol. 42:182-192

Bootstraps estimate repeatability

50 character bootstrap

50 char. sample

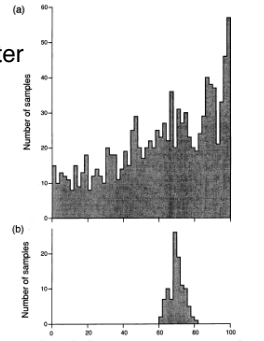
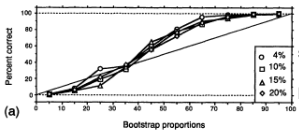


FIGURE 3. Bootstrap proportions as estimates of repeatability in phylogenetic analysis. (a) Results of 1,089 bootstrap analyses (100 pseudoreplicates each) on 1,089 actual replicates from simulation 21 (Table 2). Results shown are the proportions of the initial tree (tree A, which is the correct tree) found in each of the 1,089 analyses. (b) Proportions of solution A (the correct tree) in 100 samples of 100 actual replicates of simulation 21. The probability of estimating tree A from any given replicate is approximately 70%; samples of 100 actual replicates produce estimates of this value of 60–80%. In contrast, estimates of repeatability based on 100 bootstrap pseudoreplicates range from 0% to 100%, depending on the initial sample examined.

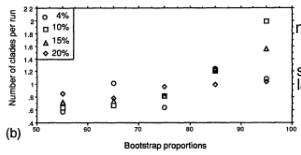
50 - 1000 characters

more correct than bootstrap predicts



small change

large change



medium change

small change

large change

FIGURE 4. (a) Relationship between bootstrap proportions and the probability of the corresponding clade being correct at various rates of internodal change (shown in inset) in nine-taxon simulations (1–4). The diagonal line indicates direct correspondence between x and y axes. (b) The average number of clades found within given bootstrap proportions in simulations 1–4 (Table 1).

fasta.bioch.virginia.edu/biol4230

Hillis (1993) Syst. Biol. 42:182-192

17

Bootstrap accuracy on “balanced” (left) and “asymmetric” (right) trees

more correct than bootstrap predicts

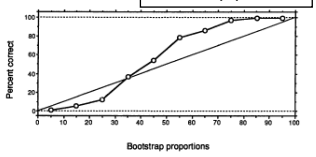


FIGURE 5. Relationship between bootstrap proportions and the probability of the corresponding clade being correct for the laboratory-generated phylogeny of nine taxa derived from bacteriophage T7. The diagonal line indicates direct correspondence between x and y axes.

similar accuracy

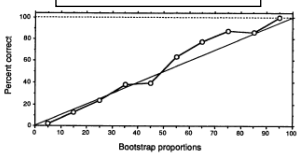


FIGURE 11. Relationship between bootstrap results and the probability of the corresponding clade being correct for a completely asymmetrical topology (simulation 10, with nine taxa). For such a topology, the bootstrap proportions are still conservative measures of reliability but less so than for symmetrical topologies (contrast with Fig. 4).

fasta.bioch.virginia.edu/biol4230

Hillis (1993) Syst. Biol. 42:182-192

18

Application and Accuracy of Molecular Phylogenies
Hillis, Huelsenbeck, and Cunningham (1994) Science 264-671

- Accuracy of phylogenetic methods can be assessed with numerical simulations or "observed evolution"
- Approaches are complementary - simulations more general, but include simplifications (independence, substitutions equally probable)
- Measures of accuracy - consistency (converge with more data), accuracy
- What is the appropriate level of complexity? 4 taxa? (exhaustive seach possible) 100 taxa? (heuristic search only)

Long Branch attraction: The "Felsenstein zone"

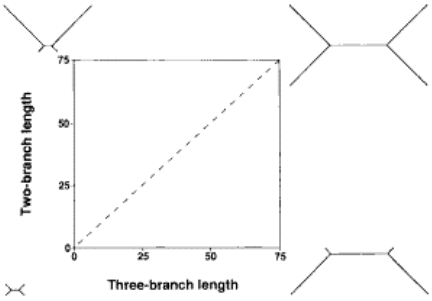


FIGURE 4. The results of the simulations were plotted with the three-branch length on the abscissa and the two-branch length on the ordinate. Different areas of the graph space represent trees with different branch lengths. Change along branches was varied from 1% internodal difference in 1% increments to the maximum length possible (=75% for four-character states). These axes apply to Figures 5–8.

Long branch attraction

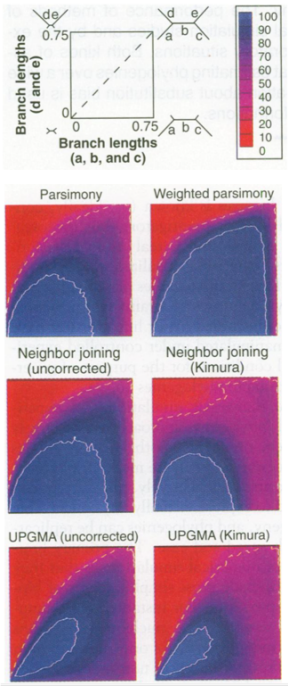


Fig. 1. Performance of three methods of phylogenetic analysis on the basis of simulation of four-taxon trees under the Kimura model of evolution (*18*). Two rates of evolution were simulated: one rate for branches a, b, and c (horizontal axis of each graph) and a second rate for branches d and e (vertical axis). The diagonal (dashed line, top left) represents equal rates of evolution along all lineages. Branch lengths are shown in expected frequency of divergent nucleotides at the two ends of the respective branches. At infinite rates of change, DNA sequences with equal base compositions are expected to differ at 75% of their positions. Blue indicates that the method estimates the correct tree a high percentage of the time under the simulated conditions; red indicates poor performance of the method (see color bar, top right). The solid white lines circumscribe the regions in which each method estimates the correct tree over 95% of the time. In the regions above the dashed white lines, the methods estimate the correct tree less than one-third of the time (a rate worse than that obtained by choosing a tree at random). The three colored graphs on the left were based on nontransformed data; the three graphs on the right show the effects of character-state weighting (for parsimony, top) and distance correction (for neighbor joining and UPGMA, middle and bottom).

Hillis (1994) Science 264:671

Tree accuracy: history, algorithm, and data

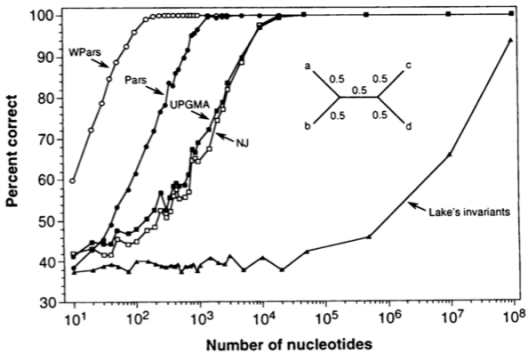
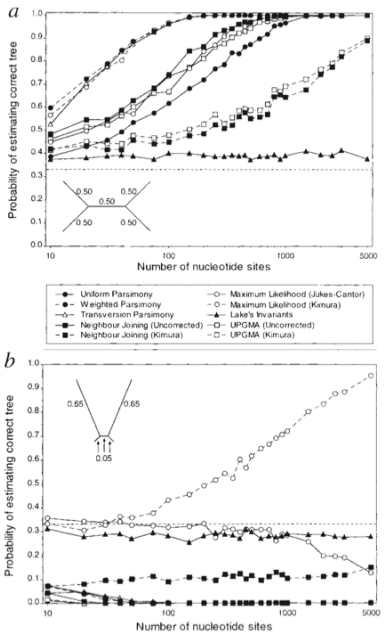


Fig. 2. Efficiency of five methods of phylogenetic analysis for a four-taxon tree with equal rates of evolution, evolving under a Kimura model of evolution and a 10:1 transition:transversion ratio. The branch lengths shown on the tree indicate that 50% of the nucleotide sites are expected to change along each branch. Although all five methods are consistent under these conditions (they all eventually converge on the correct solution), the methods differ markedly in the number of nucleotides needed to find the correct solution. All points are based on 1000 simulated trees. WPars is weighted parsimony (*45*) (any weighting of transversions over transitions from 5:1 to infinity produces results indistinguishable from those shown); Pars is uniformly weighted parsimony (*45*); NJ is neighbor joining with Kimura distances (*38*); UPGMA is the unweighted pair-group method of averages with Kimura distances (*40*); Lake's invariants is the method also known as evolutionary parsimony (*22*).

Hillis (1994) Science 264:671



Hillis (1994) Nature 369:363

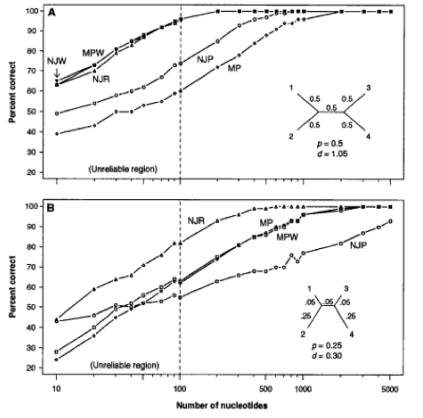
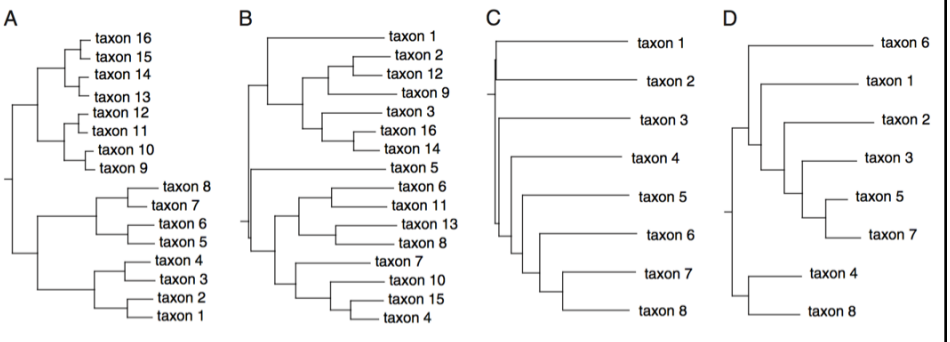


Fig. 2. Relationships between the percentage ($100 \times P_3$) of correct topologies obtained and n in computer simulations. The number of replications used is 1000. Phylogenetic trees are rarely constructed for $n < 100$, as the trees obtained for this case are unreliable. NJW, NJ with weighted p distance with the α/β ratio. (**A**) Case considered in (1). $\alpha/\beta = 10$. The results for NJR and NJW are essentially the same as those for MPW. (**B**) A case which is biologically more meaningful than case (A). $\alpha/\beta = 2$. NJR is better than MP, MPW, and NJP for all values of a = c and d = e whenever a ≠ d for the fixed value of the d distance for a + d equal to 0.3. NJ with original Kimura distance gives essentially the same results as those for NJR (data not shown).

Nei (1995) Science 267:253

Performance with "BLAST" homologs

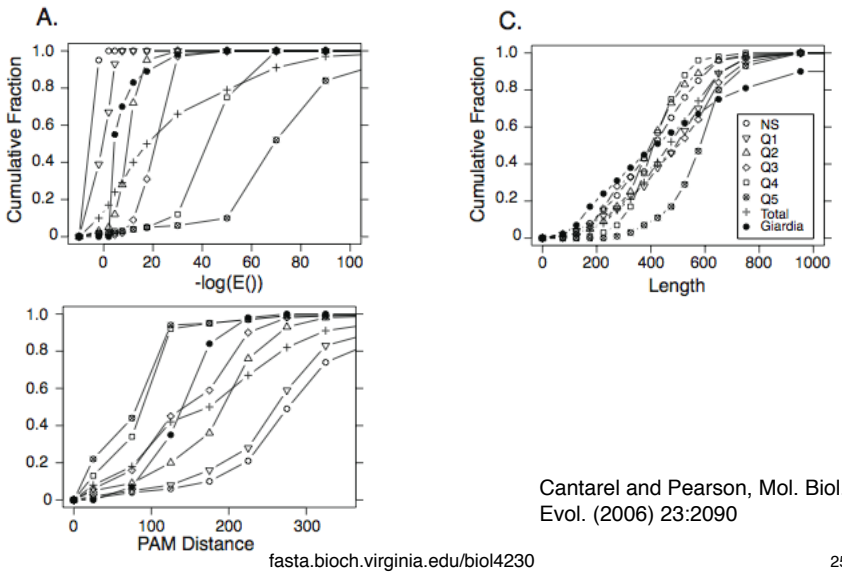
balanced topology | asymmetric topology



recent divergence | uniform divergence | ancient divergence | uniform divergence

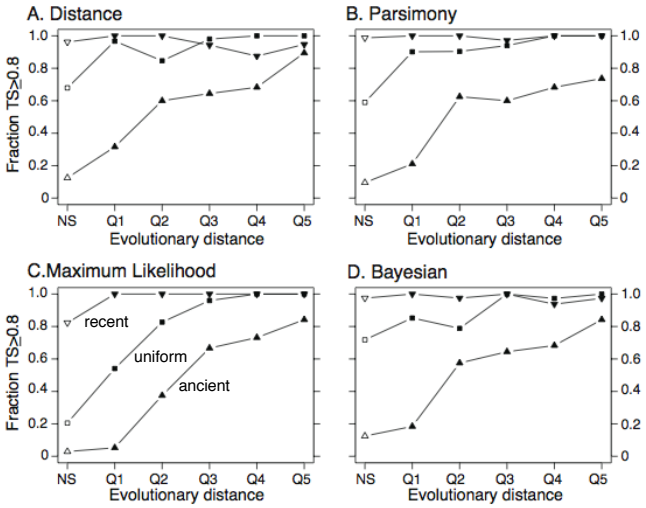
Cantarel and Pearson, Mol. Biol. Evol. (2006) 23:2090

Protein families at different distances



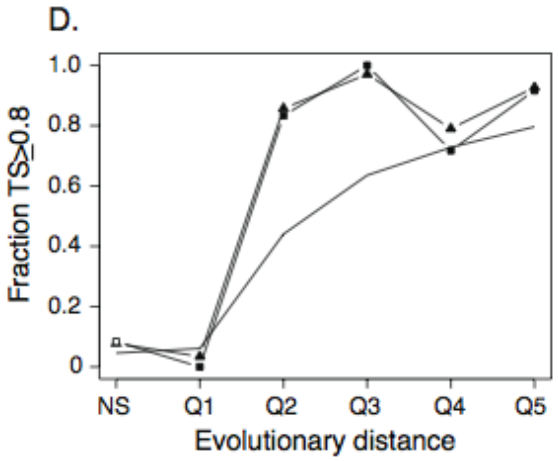
Cantarel and Pearson, Mol. Biol. Evol. (2006) 23:2090

Tree accuracy with different methods



Cantarel and Pearson, Mol. Biol. Evol. (2006) 23:2090

Robust (Bootstrap, Posterior-probability) trees are more accurate



Cantarel and Pearson, Mol. Biol. Evol. (2006) 23:2090

Methods for assessing confidence limits

TABLE V
METHODS FOR ASSESSING CONFIDENCE IN RESULTS

Method	Comments	Refs.[a]
Analytical techniques		
For parsimony procedures		
Wilcoxon rank-sum test, sign test, winning sites method	Determines whether significant character support exists for one tree relative to a second. Wilcoxon rank-sum test allows one to assign mutations different weights (i.e., transversions favoring one tree are given greater importance than transitions). For six or fewer taxa and no ordering as above, Wilcoxon rank-sum test reduces to simpler sign test. In winning sites method, binomial test is used to determine whether a greater number of phylogenetically informative positions (*sensu* parsimony) supports one tree versus a second	*1–3*
Confidence limits without clock	Assumes worst-case scenario for four taxa (two unrelated taxa with fast rates of evolution, with other two and common stem experiencing virtually no change). Under these conditions, two unrelated taxa are expected to share 3/16 of their positions by chance alone. Thus, to be statistically significant, a tree must be supported by more than 3/16 of its characters	*4*
Confidence limits with clock	Here, polytomy (star phylogeny) for four taxa is taken as worst-case situation. Thus, probability that a phylogenetically informative site supports a tree is same for all three resolutions of polytomy, 1/3	*5*
Williams/Goodman confidence limits	Similar to approach just described, except that a clock is not assumed. Method is based on a worst-case situation whereby support for correct tree is ⩾ 1/3 and ⩽ 2/3 for the two incorrect topologies combined	*6*
For evolutionary parsimony	A chi-square or binomial test is used to determine which phylogenetic invariants deviate significantly from zero and which do not	*7*
For maximum likelihood		
Likelihood ratio test	Ratio of likelihood scores for selected tree and star phylogeny is treated as a chi-square statistic with one degree of freedom. Alternatively, standard normal test of the mean and variance of the difference of their likelihood scores can be used to compare one tree to another	*2, 8, 9*
For distance approaches		
Branch length variances	An internal branch length is considered significant only if its length plus or minus two standard errors exceeds zero	*10–12*
Resampling techniques	Characters of original data set are randomly sampled and a tree is produced from new matrix. Many resampled matrices are analyzed (usually ⩾ 100). Frequency of replication of a group is taken as measure of its statistical reliability or, at least, its stability	
Bootstrapping	Characters are randomly sampled with replacement, leading to new data set of same size as original	*13, 14*
Jackknifing	Characters are randomly sampled without replacement, leading to new data set smaller than original one. Jackknifing of taxa is sometimes done instead of characters	*15, 16*

Hillis (1993) Meth. Enz. 224:456-487.

Tests for differences between alternate trees

Table 3. Total, variable, and parsimony informative characters

SSU	1,595	705	466
LSU	2,408	1,074	750
Combined	4,003	1,779	1,216

Medina et al. (2001) PNAS 98:9707

Kishino-Hasegawa test statistic

$$D = \sum D_i$$

D_i is the difference in the minimum number of substitutions at the i^{th} informative site

$$V(D) = \frac{n}{n-1} \sum_{i=1}^{n} \left(D_i - \frac{1}{n} \sum_{k=1}^{n} D_k \right)^2$$

$$t = \frac{D/n}{\sqrt{V(d)}\sqrt{n}}$$ paired t-test

Li and Graur, 2nd ed. eqn. 5.20 p. 211

The Universal Tree of Life (1997, ssRNA)

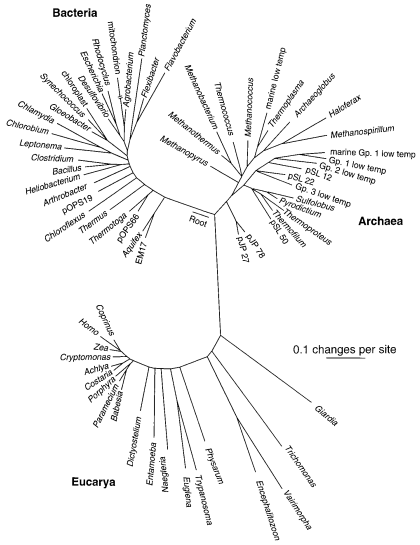
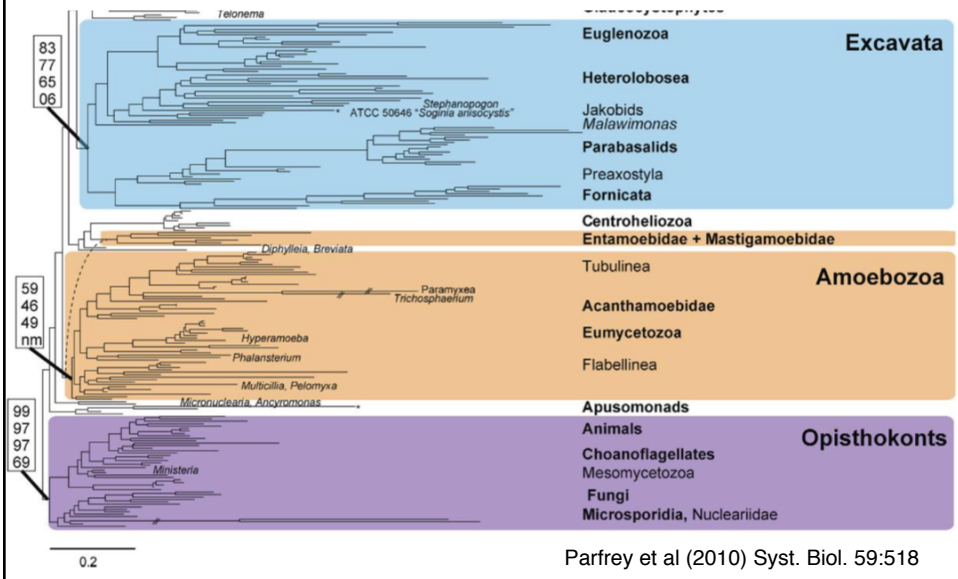


Fig. 1. Universal phylogenetic tree based on SSU rRNA sequences. Sixty-four rRNA sequences representative of all known phylogenetic domains were aligned, and a tree was produced using FASTDNAML (*43, 52*). That tree was modified, resulting in the composite one shown, by trimming lineages and adjusting branch points to incorporate results of other analyses. The scale bar corresponds to 0.1 changes per nucleotide.

Pace (1997) Science 276:734

The eukaryotic tree of life



Parfrey et al (2010) Syst. Biol. 59:518

The eukaryotic tree of life

FIGURE 1. Most likely eukaryotic tree of life reconstructed using all 451 taxa and all 16 genes (SSU-rDNA plus 15 protein genes). Major nodes in this topology are robust to analyses of subsets of taxa and genes, which include varying levels of missing data (Table 1). Clades in bold are monophyletic in analyses with 2 or more members except in all:15 in which taxa represented by a single gene were sometimes misplaced. Numbers in boxes represent support at key nodes in analyses with increasing amounts of missing data (10:16, 6:16, 4:16, and all:16 analyses; see Table 1 for more details). Given uncertainties around the root of the eukaryotic tree of life (see text), we have chosen to draw the tree rooted with the well-supported clade Opisthokonta. Dashed line indicates alternate branching pattern seen for Amoebozoa in other analyses. Long branches, indicated by //, have been reduced by half. The 6 lineages labeled by * represent taxa that are misplaced, probably due to LBA, listed from top to bottom with expected clade in parentheses. These are *Protoopalina japonica* (Stramenopiles), *Aggregata octopiana* (Apicomplexa), *Mikrocytos mackini* (Haplosporidia), *Centropyxis laevigata* (Tubulinea), *Marteilioides chungmuensis* (unplaced), and *Cochliopodium spiniferum* (Amoebozoa).

Parfrey et al (2010) Syst. Biol. 59:518

Bootstrapping and Tree reliability

- Trees describe events in the past. They cannot be confirmed for real data
  - simulations guarantee "correct" answer, but do they simulate biology?
- Tree space is enormous, and tree finding methods tend to find similar trees
  - are there "almost as good" trees that are very different topologically
- Do some methods prefer certain kinds of trees?
  - long branch attraction
- Trees with recent speciation are easier
- More data produces more robust trees