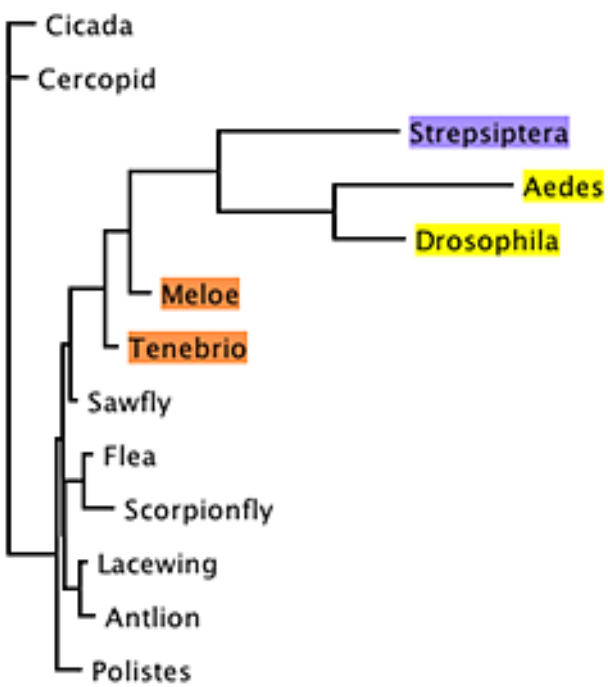
Tree and Branch Confidence Data Reliability

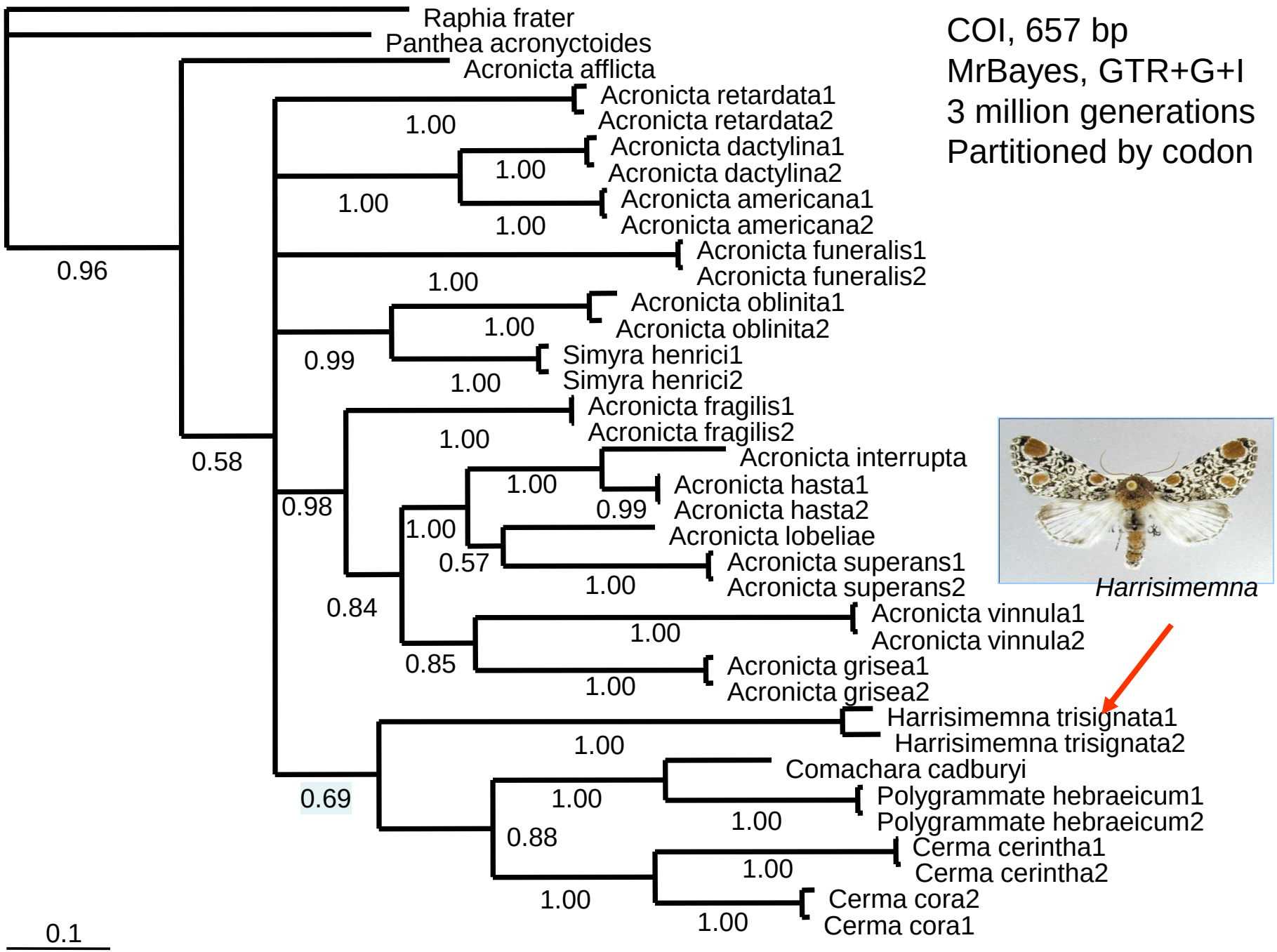


* How good is a given tree or hypothesis?
* How well does my tree reflect the evolutionary past?
* Could my tree (or parts thereof) have arisen by chance?
* How much support is there for the monophyly of a group?
* Is the placement of a given branch "good" (e.g., Strepsiptera with flies)
* What is the nature of the character support?

Taxonomic Distribution of Wood-tunneling (and Frass Ball-Rolling) Genera of Noctuoidea

My position:
- Family Noctuidae
  - Subfamily Acronictinae
    - *Comachara*
    - *Harrisimemna*
    - *Polygrammate*
    - *Cerma*

Prepupal Tunneling in Harris' Three-spot: *Harrisimemna trisignata*


COI, 657 bp
MrBayes, GTR+G+I
3 million generations
Partitioned by codon
Raphia frater
Panthea acronyctoides
Acronicta afflicta
Acronicta retardata1
Acronicta retardata2
Acronicta dactylina1
Acronicta dactylina2
Acronicta americana1
Acronicta americana2
Acronicta funeralis1
Acronicta funeralis2
Acronicta oblinita1
Acronicta oblinita2
Simyra henrici1
Simyra henrici2
Acronicta fragilis1
Acronicta fragilis2
Acronicta interrupta
Acronicta hasta1
Acronicta hasta2
Acronicta lobeliae
Acronicta superans1
Acronicta superans2
Acronicta vinnula1
Acronicta vinnula2
Acronicta grisea1
Acronicta grisea2
Harrisimemna trisignata1
Harrisimemna trisignata2
Comachara cadburyi
Polygrammate hebraeicum1
Polygrammate hebraeicum2
Cerma cerintha1
Cerma cerintha2
Cerma cora2
Cerma cora1
Harrisimemna
1.00
1.00
1.00
1.00
0.96
1.00
1.00
0.99
1.00
1.00
0.58
0.98
1.00
0.99
1.00
0.57
1.00
0.84
1.00
0.85
1.00
1.00
0.69
1.00
1.00
0.88
1.00
1.00
1.00
0.1

Tree and Branch Confidence

* Trees and/or clades are hypotheses
  - without measures of reliability trees are just summaries of data at hand

* trees or clades are hypotheses and must be viewed with caution, esp. if there is appreciable homoplasy
* with only four bases convergence is a certainty
* or too little data
* data correction is a best guess – we don’t really know what happened

Tree and Branch Confidence

* Every tree dependent on the
  - algorithm used
  - taxon sampling
  - outgroup selection
  - gene/data selection
  - homology/alignment decisions (and gap models)
  - data exclusion decisions
  - model choice (and fit with data)
  - analysis decisions (e.g., partitions)
  - etc.

  * Uncertainty necessitates measures of reliability

Tree and Branch Reliability

(1) Consistency and Retention Indices
(2) Decay Indices
(3) Nonparametric Bootstrapping
(4) Parametric Bootstrapping (see also comparing trees lecture)
(5) Jackknifing
(6) Bayesian Posterior Probabilities
(7) Data Exploration/Sensitivity Analyses

Consistency Index

* for character = minimum # changes/observ. # changes
  e.g., if two-state character evolves three times on a tree
  c.i. will be 1/3=0.333;

  e.g., if two-state character evolves fives times on a tree
  c.i. will be 1/5=0.2

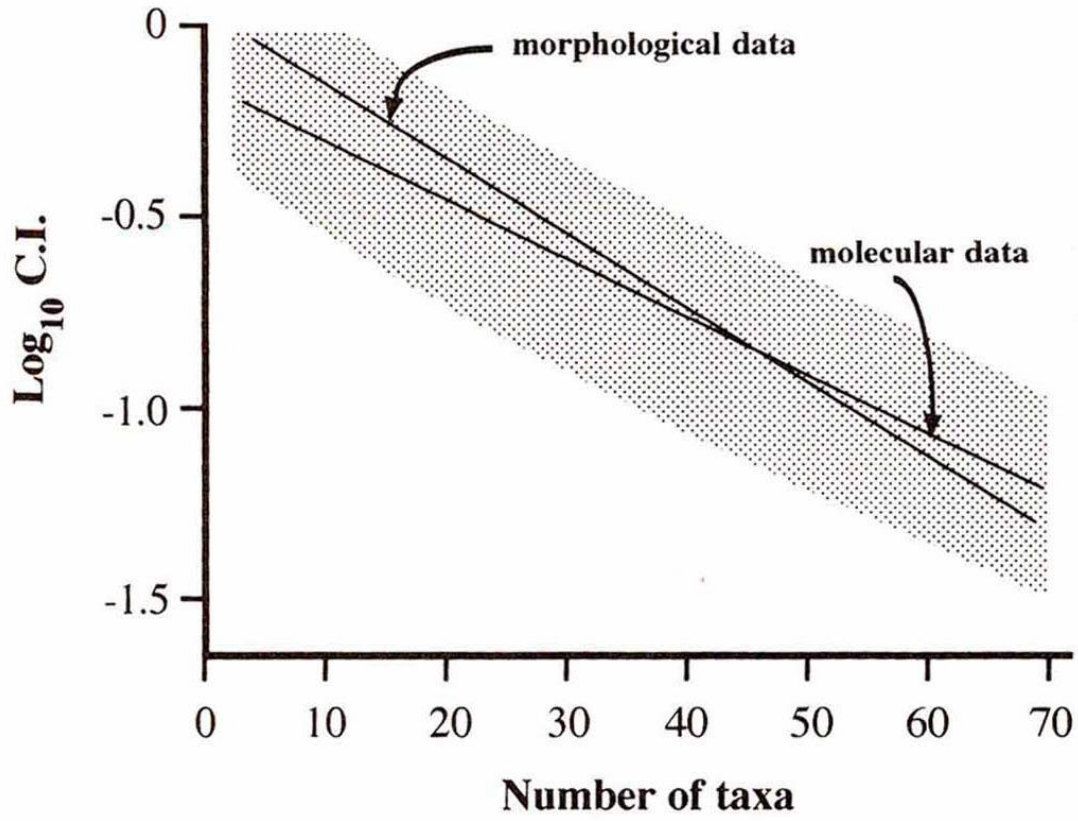
* morphologists (subconsciously) evaluating c.i. during their character selection (jettison characters with low cis)

* **c.i**. for cladogram tree = **ensemble c.i.**

$$\frac{\text{minimum \# steps for all characters}}{\text{\# observed steps for all characters}}$$

Problems with Consistency Indices

1) Upper bound is 1.0 (when there is no homoplasy)
   - but lower bound is undefined
2) c.i.s are correlated with numbers of characters as well as numbers of terminal taxa
   - greater numbers introduce chances for character conflict
   - thus c.i. falls off in larger data sets
3) can’t compare c.i. across data sets
4) redundancy in taxa or characters inflates c.i.
5) autapomorphic characters inflate the c.i.


morphological data
molecular data
0
-0.5
-1.0
-1.5
Log10 C.I.
0
10
20
30
40
50
60
70
Number of taxa

Retention Index

(Farris 1989)

* measures amount of synapomorphy

* r.i. = (g-s)/(g-m)

* where g = maximum homoplasy possible (on star tree)
  s = observed change (in a given tree)
  m = minimum amount of change possible
* measure homoplasy as a fraction of the *maximum possible* homoplasy

ensemble r.i. (of tree) calculated in an analogous fashion

$$R = \Sigma\ (g-s)/(g-m)$$

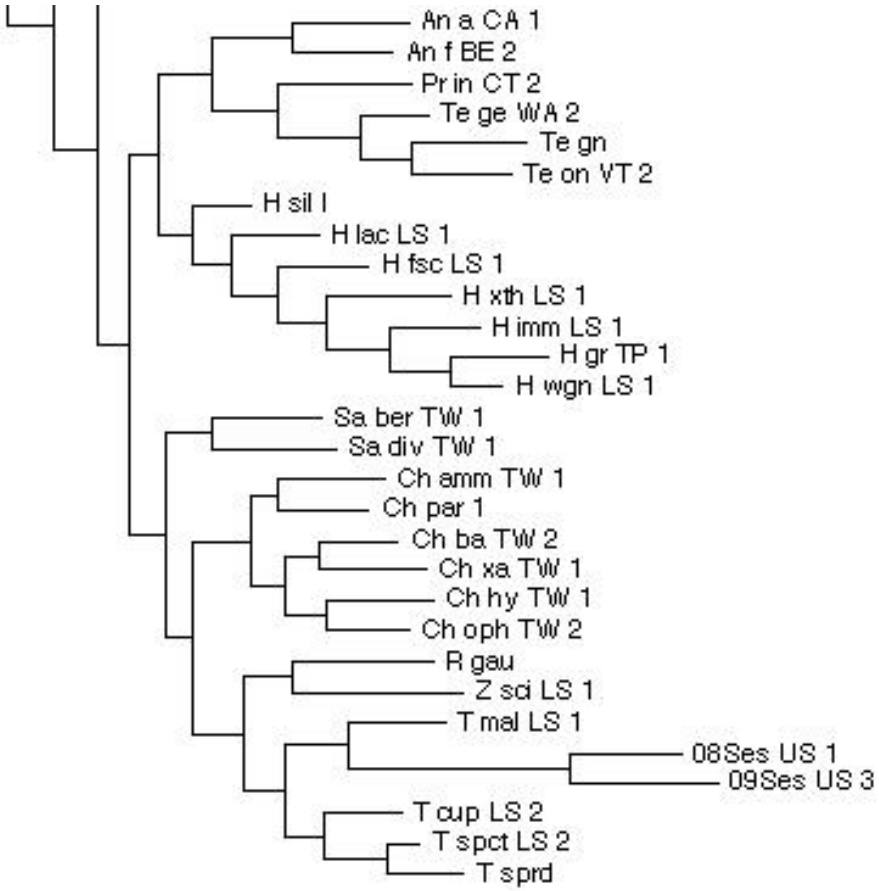
Retention Index

* retention index scales to zero
* Imagine a data set with 100 taxa and a binary character that is present in 10 of the 100 and it evolve 3 times: c.i. = 1/3 = .33; r.i.= 10-3/10-1 = 7/9 = 0.777

* But in smaller data set the r.i. would be smaller
  - character that changed three times among ten taxa would have a ci = 0.33
  - but r.i. would now fall to 3-3/3-1 = 0/2 = 0
* Excludes autapomorphies
  - because with autapomorphies: g, s, and m are all 0
* r.i. also tends to fall off monotonically with data set size

Rescaled consistency index = ci x ri

Tree #	1
Length	2428
CI	0.239
RI	0.401
RC	0.096
HI	0.761



10 changes

Bremer Support or Decay Index

(Bremer 1995)

Bremer support value = the # of extra steps needed to lose a branch in the consensus tree of all most parsimonious trees

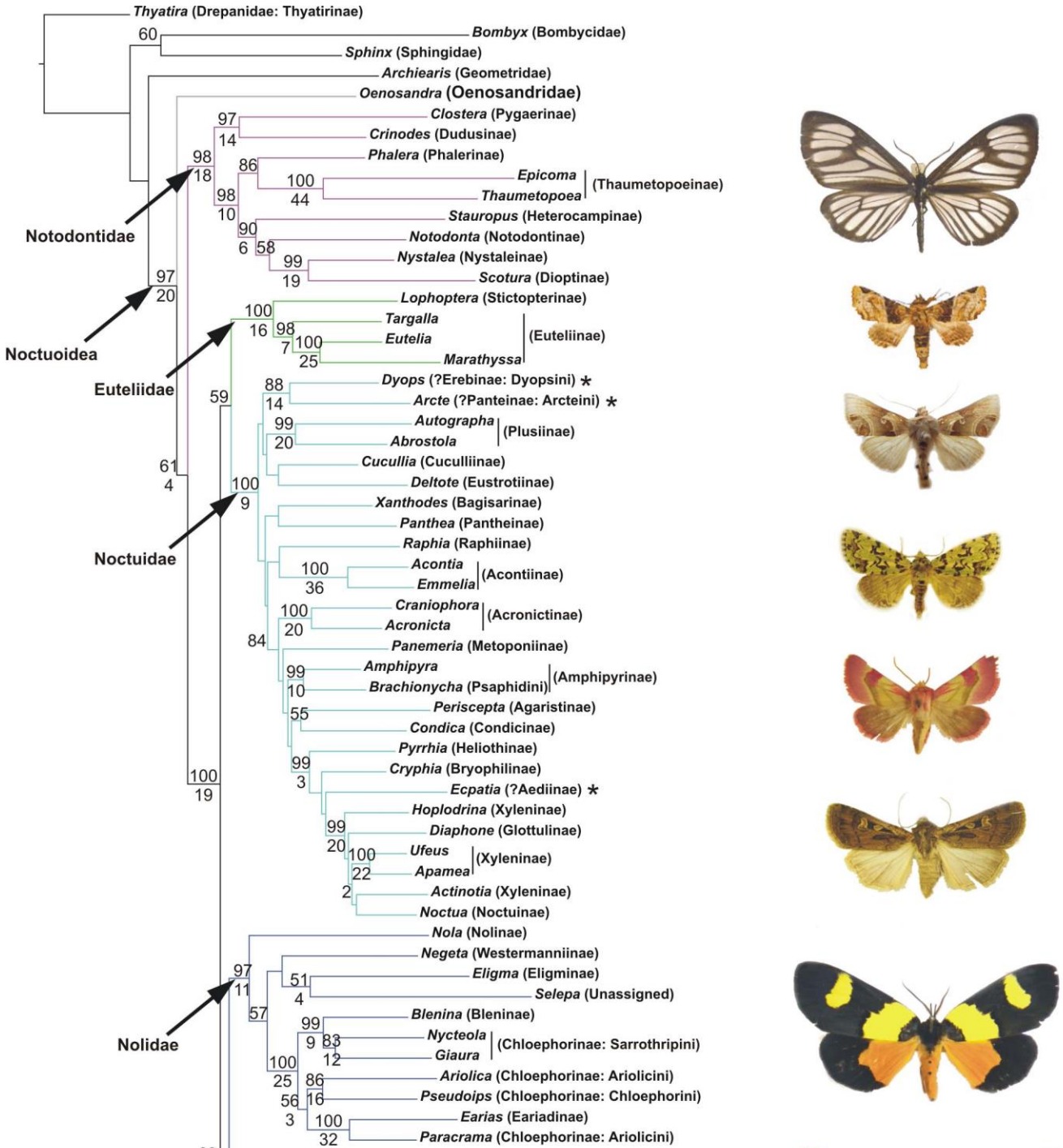
stated differently: difference in length between the most parsimonious tree(s) and the shortest trees in which the branch of interest is not resolved (Schuh and Brower 2009)

Watch for branch loss among strict consensus trees as steps are added to the tree lengths.

Bremer Support

* Goal: to eliminate support for a branch that could be due to homoplasy
  - because homoplastic characters will map on a tree in many ways, by considering trees that are one or two steps longer, groups held together by homoplasies will disappear early in a decay analysis
* Think of them as *successively relaxed parsimony*
* PAUP calculates by constraining tree and counting number of extra steps required for clade (autodecay/treerot option)

* Bremer support (decay indices) and bootstraps values often plotted together along a given branch



Zahiri, R., I. J. Kitching, J. D. Lafontaine, M. Mutanen, L. Kaila, J. D. Holloway, and N. Wahlberg. 2011. A new molecular phylogeny offers hope for a stable family-level classification of the Noctuoidea (Lepidoptera). *Zoologica Scripta* 40: 158-173.

Bremer Support

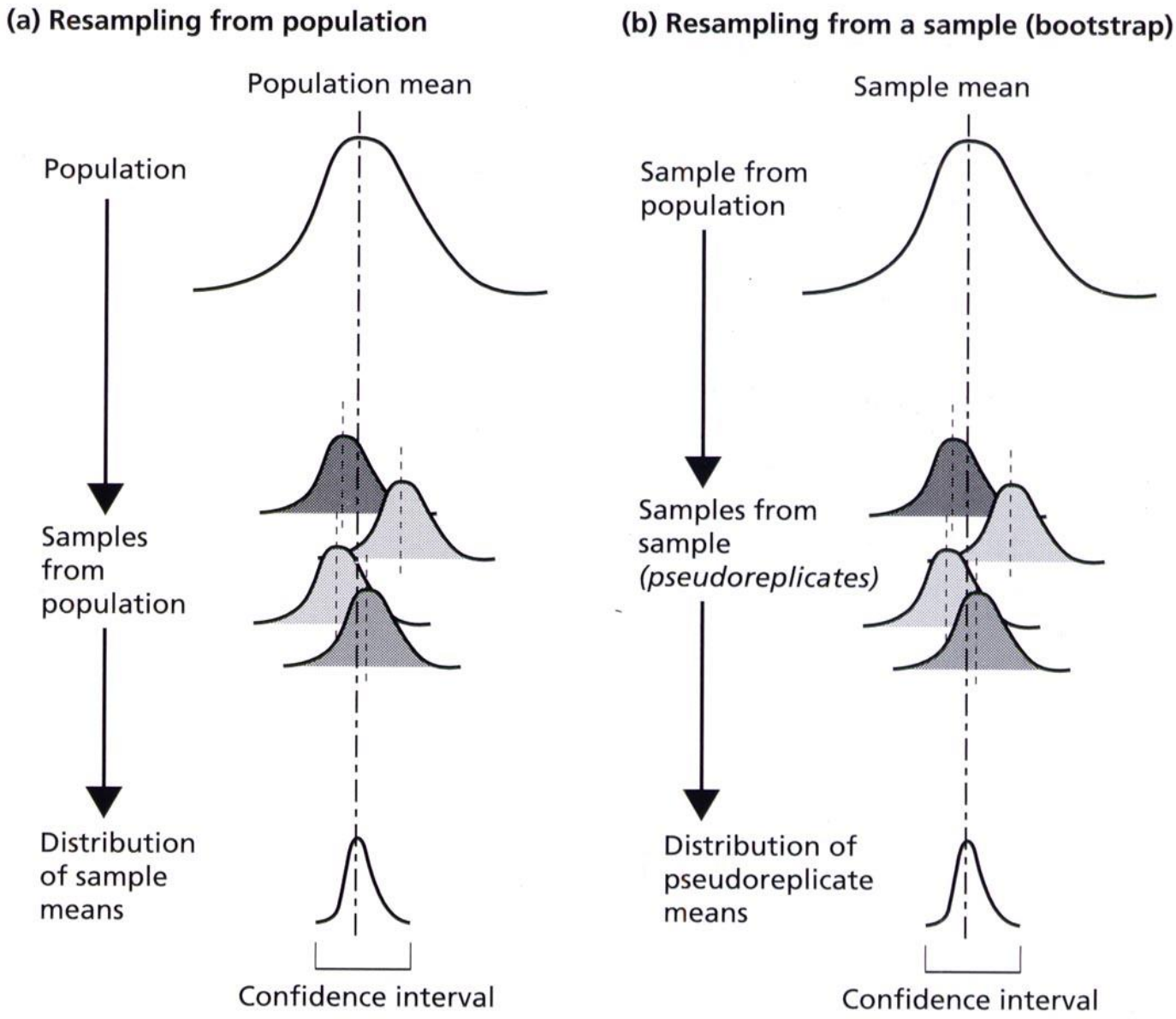
* Issues: meaning vague; not comparable
  - tend to be very large in molecular data sets
* commonly used in parsimony analyses, esp. for morphological data (because of fewer numbers of characters)

Non-parametric Bootstrap

(Felsenstein 1985)

* bootstrapping is a statistical technique used to estimate the variability of a statistic when the underlying distribution is unknown...it gets its name from having to pull one's self up by the bootstraps in a statistically difficult situation

* basically trying to estimate unknown mean and variance of a sample or population by random ***resampling*** of data
* you're forced to resample your sample of the distribution
* a matter of taking *pseudosamples* to estimate the true mean
* widely used "reliability" measure, esp. for molecular data


(a) Resampling from population
Population mean
Population
Samples from population
Distribution of sample means
Confidence interval
(b) Resampling from a sample (bootstrap)
Sample mean
Sample from population
Samples from sample (pseudoreplicates)
Distribution of pseudoreplicate means
Confidence interval

Non-parametric Bootstrap Method

(Felsenstein 1985)

(1) characters are randomly sampled from data set with replacement until data set of equal size is obtained

* make pseudoreplicate by randomly sampling columns from your data matrix to make new sample of equal size
* some characters sampled more than once, others not at all

(2) generate tree from each pseudoreplicate

(3) repeat process, i.e., resample data and generate tree from each pseudoreplicate100 to 1000 times

(4) generate majority rule consensus tree

(5) record fraction of times each branch was recovered:

e.g., if 760 trees out of a 1000 bootstraps had a given clade bootstrap value for clade = 76

Bootstrap values are not confidence intervals

* they are a measure of internal branch support “or branch reliability”

* they measure not whether a branch is real but the probability of getting this same branch if more data were collected

* *systematic* error in data (or analysis) could result in high bootstrap value as more data is collected but the node (grouping) could be wrong

* another way to think about bootstrap branch estimates:

* 1-P = probability of getting that much evidence if the group, in fact, did not exist

* Thus, if a branch comes up supported 93% of the time: 1-93%=7%
  - 7% of the time you can expect to see this branch (this well) supported when in fact the group does *not* exist)

Non-parametric Bootstrap

* in a (statistical) bootstrap every character is supposed to be independent and identically distributed, they are not in phylogenetics, but we still use them (meaningfully)

* if multiple comparisons are being made (i.e., branch lengths being evaluated) one should employ Bonferroni correction (see Felsenstein 2004)
  - e.g., if you care about the monophyly and branch support for twenty nodes on your tree, statistically you should expect 1 of the 20 to be well supported by chance alone

* low bootstrap values (ca. under 70%) tend to be overestimates of signal; and high bootstrap values (ca. over 70%) tend to be low estimates of phylogenetic signal (Hillis and Bull 1993; Zharkikh and Li 1995, Li and Zharkikh 1995)

Parametric Bootstrap

(Efron 1985; Huelsenbeck et al. 1996)

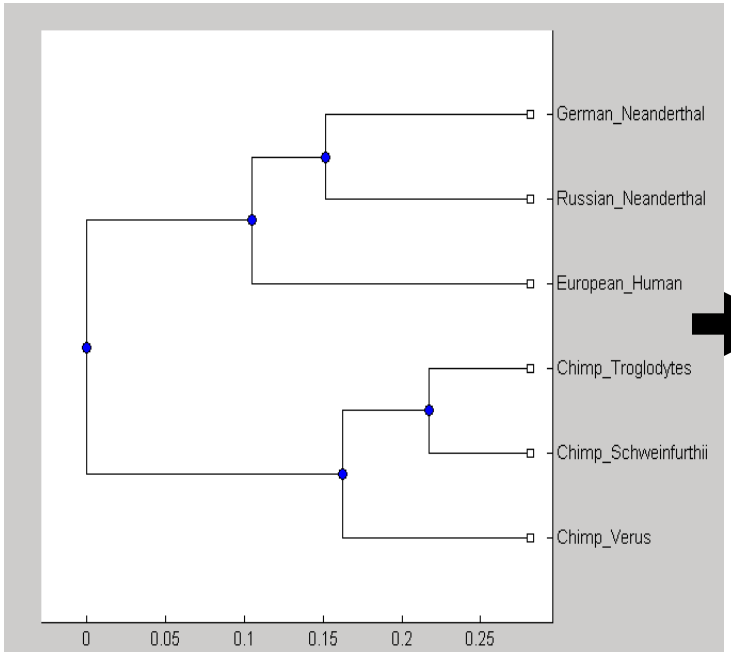
* a hybrid between simulation and bootstrapping
* with parametric (an non-parametric) bootstrapping the goal is to mimic the variability one would get if you were taking independent phylogenetic estimates of the true tree (mean)
* in parametric bootstrapping one simulates data…to generate estimates in the vicinity of our best tree…using same (statistical) model of evolution to generate simulated data matrices → trees

Parametric Bootstrap Method

(Efron 1985; Huelsenbeck et al. 1996)

1) build best tree
2) generate *simulated data* sets using estimated branch lengths and other parameters (e.g., alpha and substitution model) (from tree/data)
3) build new tree for simulated data set
4) replicate 100 to 1000 times
5) generate majority rule consensus or tally fraction of the trees that come out with each topology

Parametric Bootstrap



Use tree branches and model to simulate new data matrices

Simulation 1

Germ Neand	CCTGGCATAA ATCGCATACG
Rus Neand	CCTGGCATAA ATCGCATACG
Europ. Hum	CCTGGCATTA ATCGCATTCG
Chimp trog	ACTGGCTTTA ATCGCATTCG
Chimp Schw	ACTGGCTTTA ATCGCATTCG
Chimp venus	ACTGGCATTA ATCGCATTCG

→ T_1

Simulation 2

Germ Neand	ACAGGCATAA ATCGCATACG
Rus Neand	ACAGGCATAA ATCGCATACG
Europ. Hum	ACAGGCATTA ATCGCATTCG
Chimp trog	ACTGGCTTTA ATCGCATTCG
Chimp Schw	ACTGGCTTTA ATCGCATTCG
Chimp venus	ACTGGCATTA ATCGCATTCG

→ T_2

Simulation 3

Germ Neand	ACAGGCATAA ATCGCATACG
Rus Neand	ACAGGCATAA ATCGCATACG
Europ. Hum	ACAGGCATTA ATCGCATTCG
Chimp trog	ACTGGCTTTA ATCGCATTCG
Chimp Schw	ACTGGCTTTA ATCGCATTCG
Chimp venus	ACTGGCATTA ATCGCATTCG

→ T_3

⇓

To n replicates

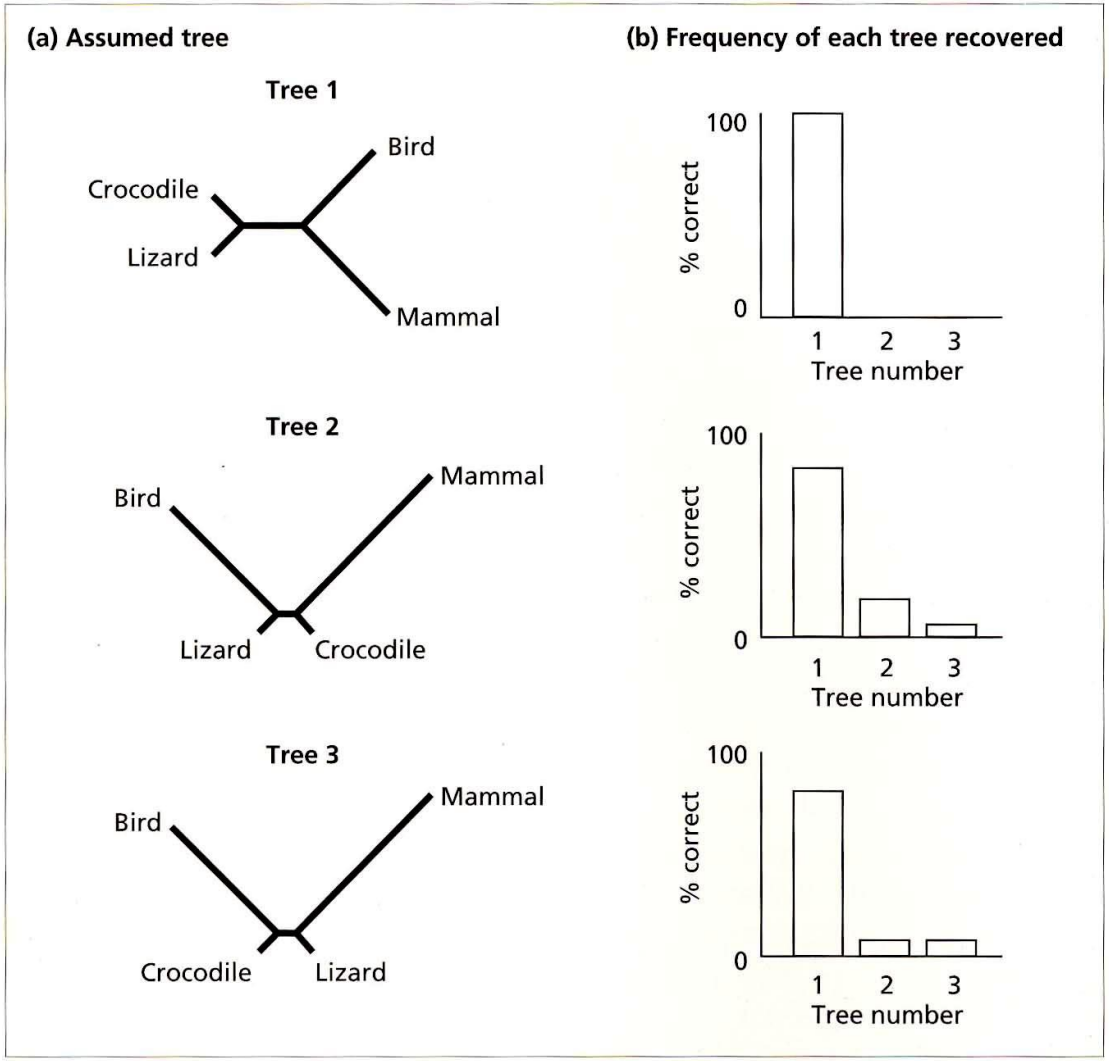


Fig. 6.38 Parametric bootstrapping. Three alternative trees for 18S rRNA sequences from a bird, mammal, crocodile and lizard are shown on the left. For each tree 1000 artificial data sets of the same length as the original 18S rRNA data were generated using parameters derived from that tree. On the right is shown the proportion of times each tree was the most parsimonious tree for the data sets derived from each tree. Note that no matter which tree was used to generate the data, tree 1 is most often recovered as the most parsimonious tree. After Huelsenbeck *et al.* (1996).

Jackknifing

* a resampling procedure *without* replacement
* trees built from smaller data sets
* compares trees built from random subsets of the data
  - can delete characters or
  - other delete taxa

Jackknifing

Method:

1) delete portion of characters (or taxa) and generate tree
   - half jackknife deletes half the characters
2) replace characters (or taxa) then repeat step 1, n times
3) construct majority rule consensus and plot number of times a clade is supported on each node
   - again, clades that appear less than 70% of the time should be viewed with a bit of caution

Bayesian Posterior Probabilities

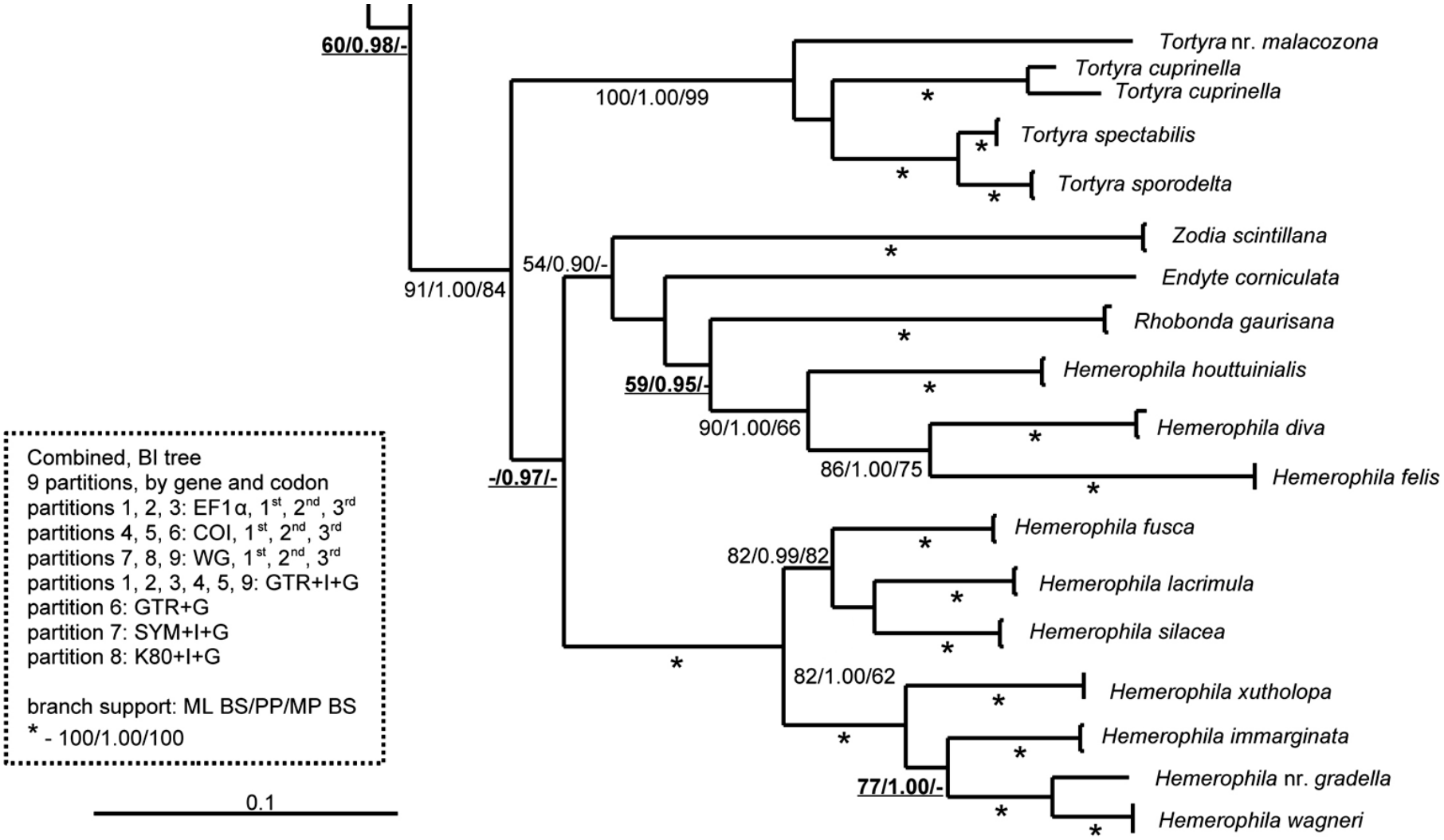
* sample tree space, changing one parameter at a time, build a tree, then change another parameter, build a tree, and so forth (tweak and build)

* algorithm encouraged to find most likely tree given the data (and a model of evolution)

* Bayesian approach yields a set of trees that is most likely to be explained by the sequences, or formally, "*the probability of the hypothesis being correct given the data*" (P[H|D])

Bayesian Posterior Probabilities

* save a tree each time one of the parameters in the model is changed, i.e., at every interval determined by “samplefreq” command
* common to generate 5-10 million trees, and save/sample one tree every 1000 generations
* makes a tree file from sampled trees
* builds a majority rule consensus tree
* number tells us what proportion of the trees had a given clade
* unlike bootstraps Bayes posterior probabilities will be an estimate of the true probabilities of that clade


60/0.98/-
100/1.00/99
Tortyra nr. malacozona
Tortyra cuprinella
Tortyra cuprinella
Tortyra spectabilis
Tortyra sporodelta
91/1.00/84
54/0.90/-
Zodia scintillana
Endyte corniculata
Rhobonda gaurisana
59/0.95/-
Hemerophila houttuinialis
90/1.00/66
Hemerophila diva
86/1.00/75
Hemerophila felis
-/0.97/-
Hemerophila fusca
82/0.99/82
Hemerophila lacrimula
Hemerophila silacea
82/1.00/62
Hemerophila xutholopa
Hemerophila immarginata
77/1.00/-
Hemerophila nr. gradella
Hemerophila wagneri
Combined, BI tree
9 partitions, by gene and codon
partitions 1, 2, 3: EF1α, 1st, 2nd, 3rd
partitions 4, 5, 6: COI, 1st, 2nd, 3rd
partitions 7, 8, 9: WG, 1st, 2nd, 3rd
partitions 1, 2, 3, 4, 5, 9: GTR+I+G
partition 6: GTR+G
partition 7: SYM+I+G
partition 8: K80+I+G
branch support: ML BS/PP/MP BS
* - 100/1.00/100
0.1

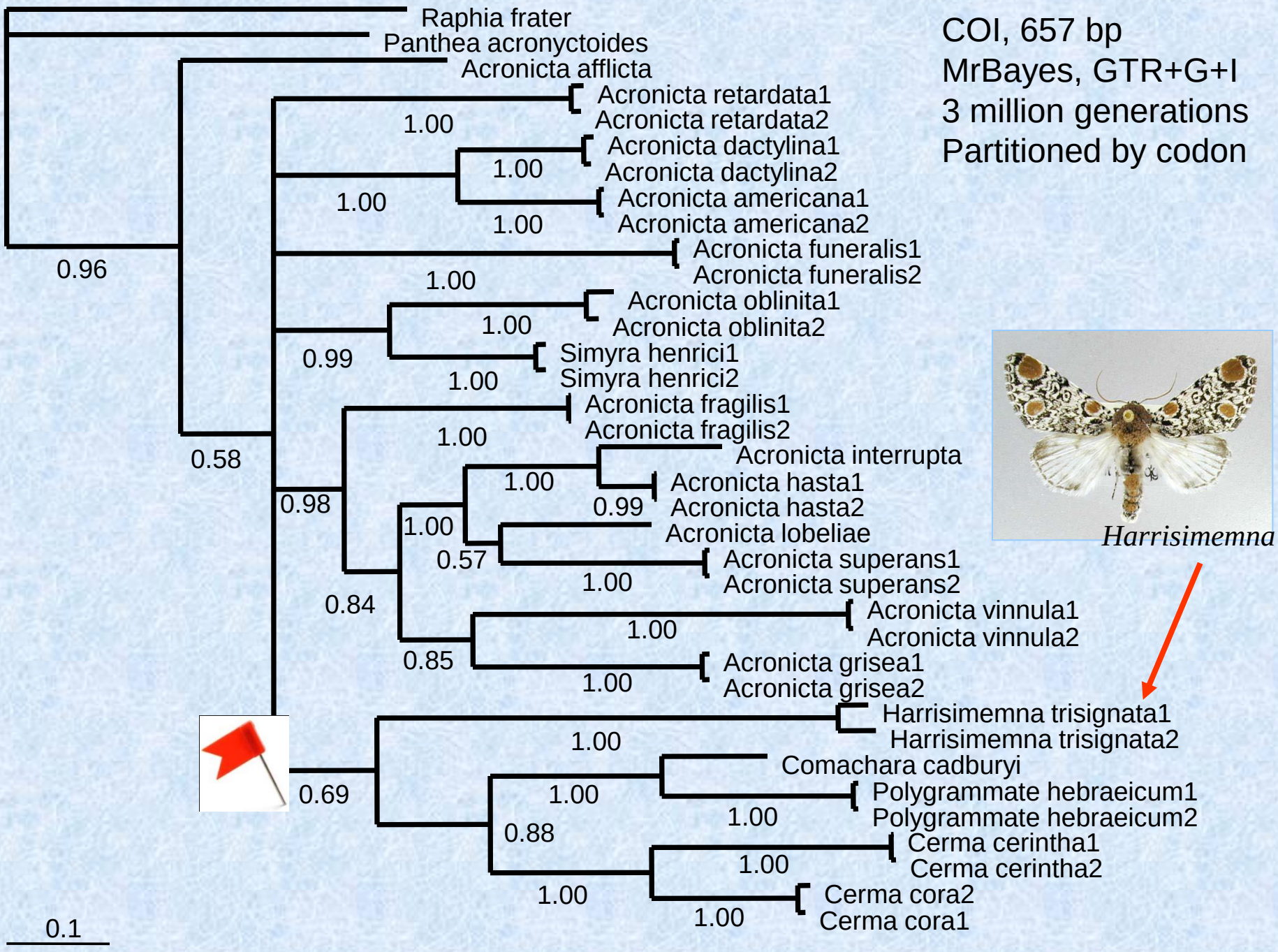
End 21 October

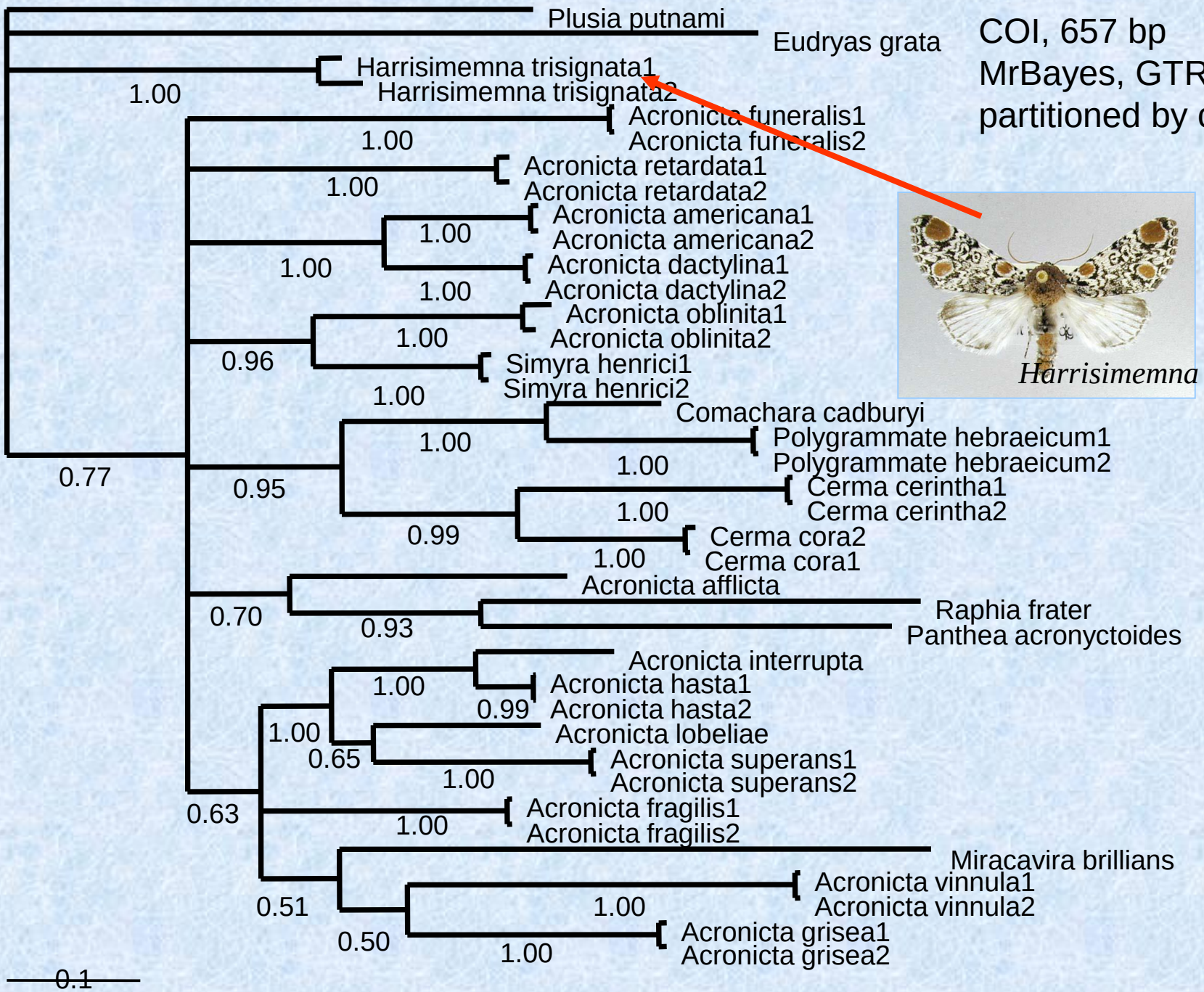
Tree and Branch Reliability

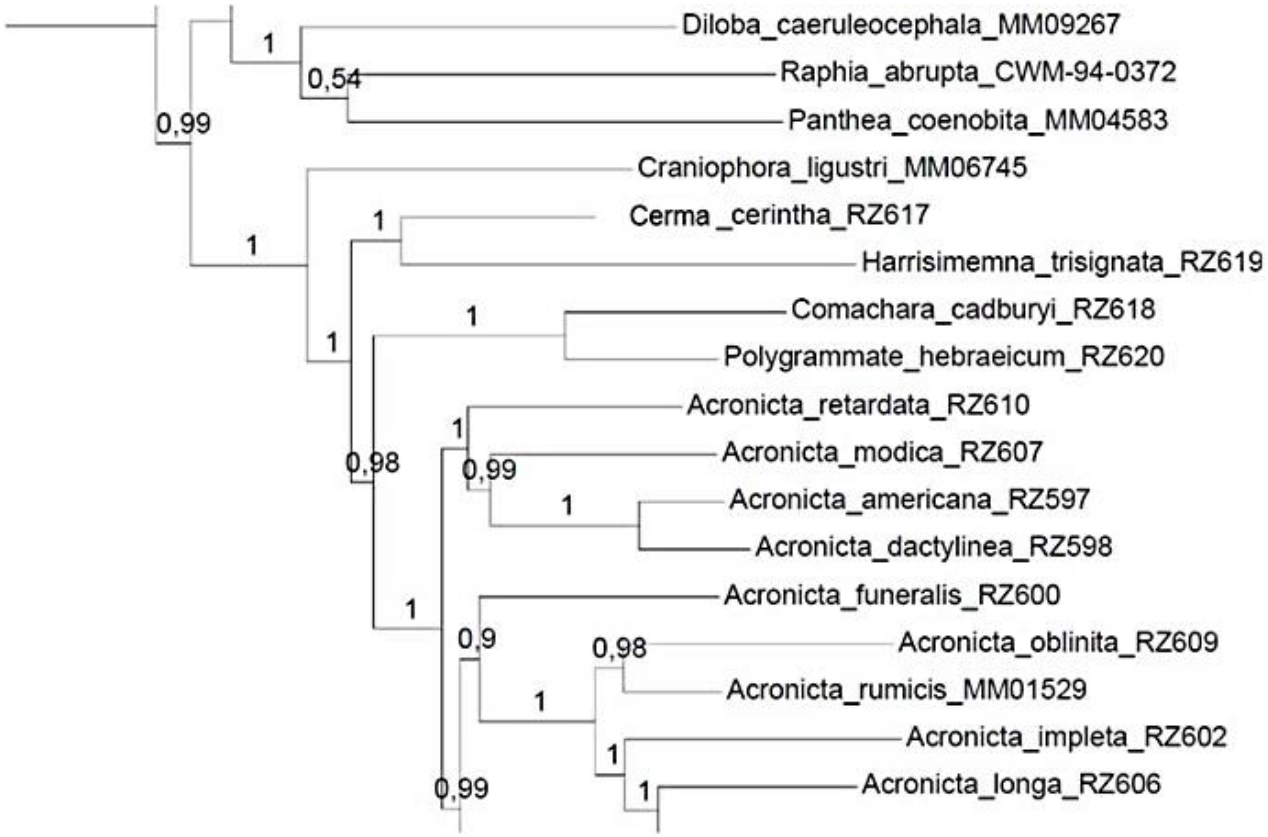
(1) Consistency and Retention Indices
(2) Decay Indices
(3) Nonparametric Bootstrapping
(4) Parametric Bootstrapping (see also comparing trees lecture)
(5) Jackknifing
(6) Bayesian Posterior Probabilities
(7) Data Exploration/Sensitivity Analyses

Data Exploration:

* engage in some “sensitivity” analyses
* compare results from different algorithms
* try different outgroups
* delete ambiguously aligned characters
* try deleting “rogue” taxon (which can change groupings and especially branch support)


COI, 657 bp
MrBayes, GTR+G+I
3 million generations
Partitioned by codon
Raphia frater
Panthea acronyctoides
Acronicta afflicta
Acronicta retardata1
Acronicta retardata2
Acronicta dactylina1
Acronicta dactylina2
Acronicta americana1
Acronicta americana2
Acronicta funeralis1
Acronicta funeralis2
Acronicta oblinita1
Acronicta oblinita2
Simyra henrici1
Simyra henrici2
Acronicta fragilis1
Acronicta fragilis2
Acronicta interrupta
Acronicta hasta1
Acronicta hasta2
Acronicta lobeliae
Acronicta superans1
Acronicta superans2
Acronicta vinnula1
Acronicta vinnula2
Acronicta grisea1
Acronicta grisea2
Harrisimemna trisignata1
Harrisimemna trisignata2
Comachara cadburyi
Polygrammate hebraeicum1
Polygrammate hebraeicum2
Cerma cerintha1
Cerma cerintha2
Cerma cora2
Cerma cora1
Harrisimemna
1.00
1.00
1.00
1.00
0.96
1.00
1.00
0.99
1.00
1.00
0.58
0.98
1.00
0.99
1.00
0.57
1.00
0.84
1.00
0.85
1.00
1.00
0.69
1.00
1.00
0.88
1.00
1.00
1.00
0.1


Plusia putnami
Eudryas grata
Harrisimemna trisignata1
Harrisimemna trisignata2
1.00
Acronicta funeralis1
Acronicta funeralis2
1.00
Acronicta retardata1
Acronicta retardata2
1.00
Acronicta americana1
Acronicta americana2
1.00
Acronicta dactylina1
Acronicta dactylina2
1.00
Acronicta oblinita1
Acronicta oblinita2
1.00
0.96
Simyra henrici1
Simyra henrici2
1.00
Comachara cadburyi
Polygrammate hebraeicum1
Polygrammate hebraeicum2
1.00
0.77
0.95
Cerma cerintha1
Cerma cerintha2
1.00
0.99
Cerma cora2
Cerma cora1
1.00
Acronicta afflicta
0.70
Raphia frater
0.93
Panthea acronyctoides
Acronicta interrupta
1.00
Acronicta hasta1
0.99
Acronicta hasta2
1.00
Acronicta lobeliae
0.65
Acronicta superans1
Acronicta superans2
1.00
0.63
Acronicta fragilis1
Acronicta fragilis2
1.00
Miracavira brillians
Acronicta vinnula1
Acronicta vinnula2
0.51
1.00
Acronicta grisea1
Acronicta grisea2
0.50
1.00
0.1
COI, 657 bp
MrBayes, GTR+G+I,
partitioned by codon
Harrisimemna



Harrisimemna

COI + 7 nuclear genes: CAD, wingless, EFI (two regions), GAPDH, IDH, MDH, and Ids5.
Bayes tree: GTR+G+I