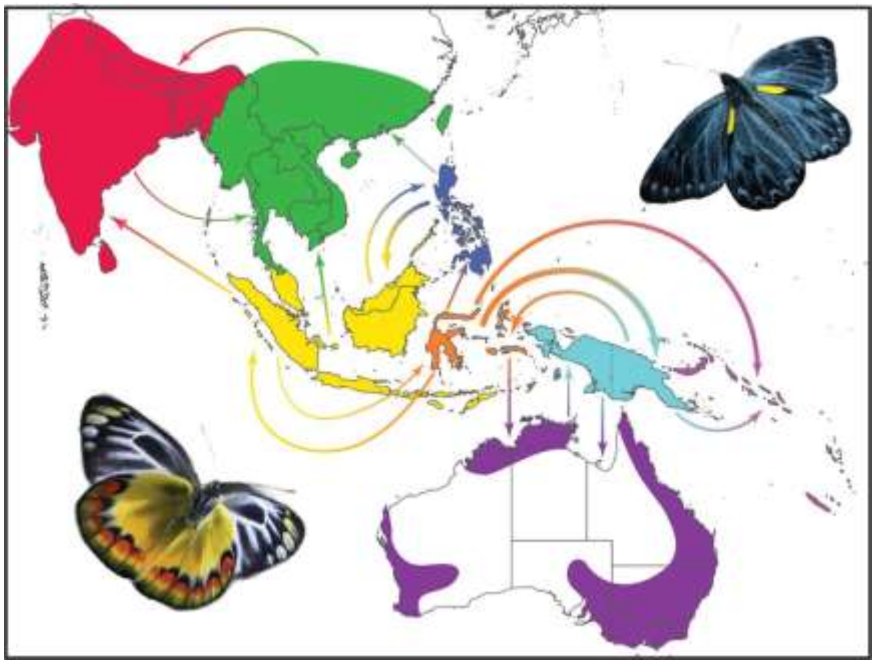
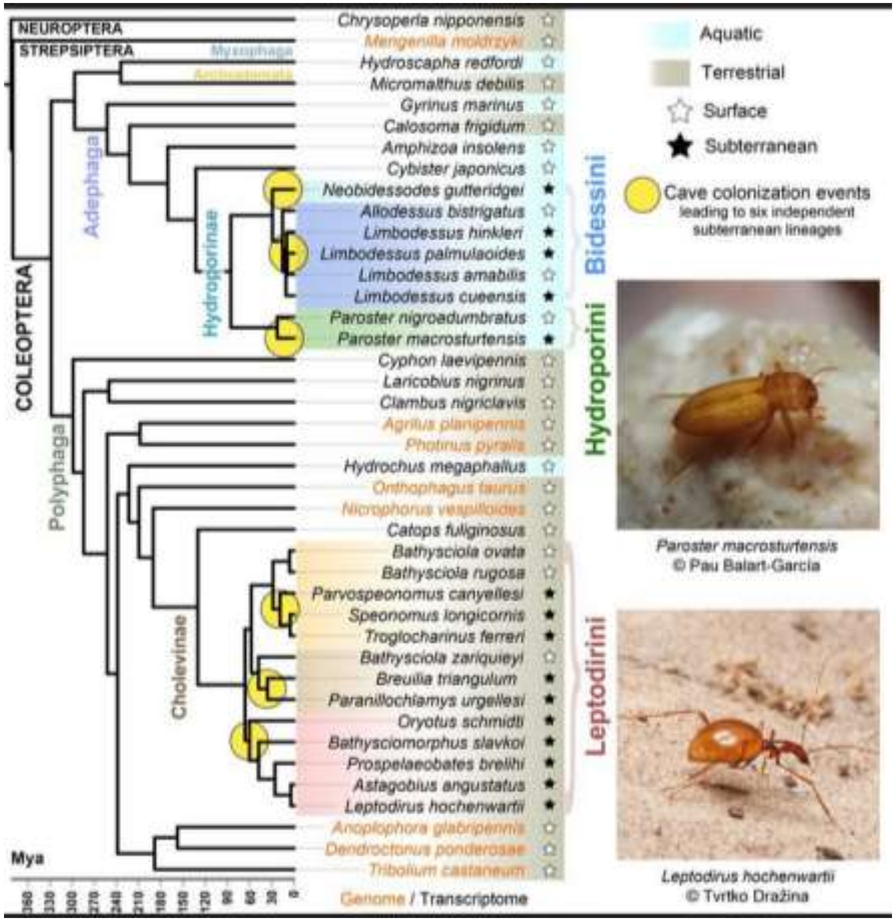
How to read and make phylogenetic trees, part 2
+ Use of molecular phylogetics in zoology





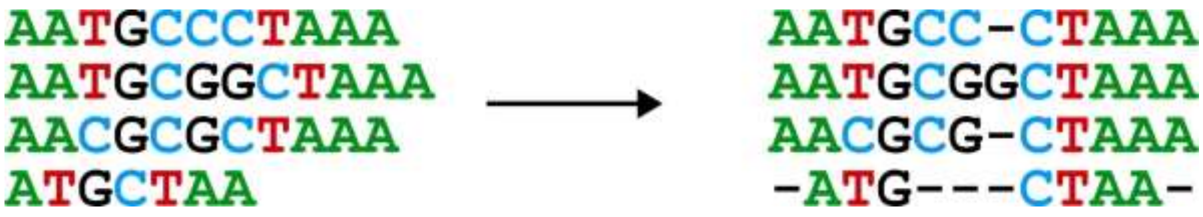
Zuzana Starostová, Department of Zoology, Faculty of Science, Charles University

How to make phylogenetic trees?

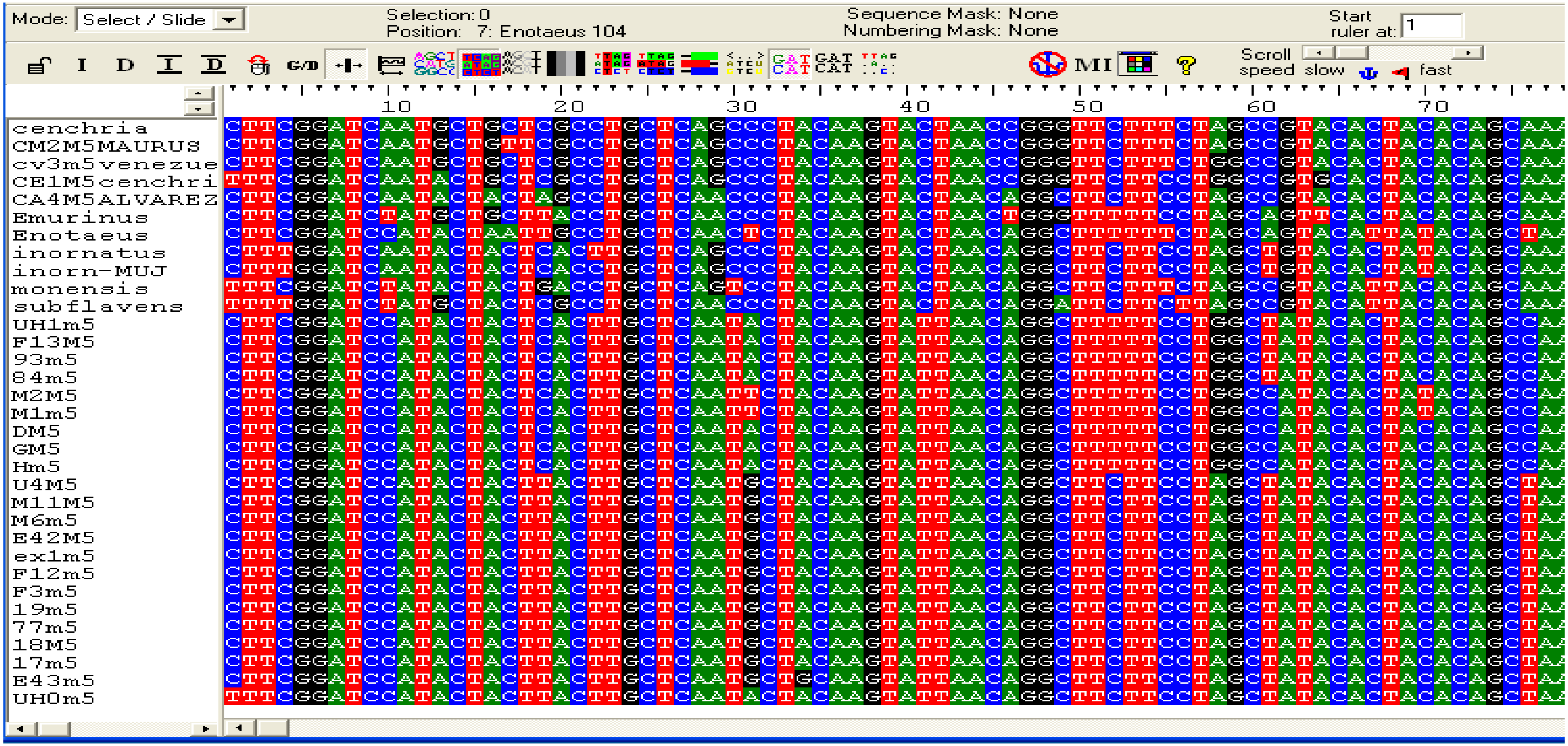
Workflow:

- ✓ obtain DNA sequence
- ✓ quality check
- ✓ sequence alignment
- calculating genetic distances
- phylogeny estimation – topology and branch length
- NJ, PM, ML, BA
- reliability test (bootstrap)
- tree visualization

Alignment



```
AATGCCCTAAA        AATGCC-CTAAA
AATGCGGCTAAA   →   AATGCGGCTAAA
AACGCGCTAAA        AACGCG-CTAAA
ATGCTAA            -ATG---CTAA-
```



Mode: Select / Slide

Selection: 0
Position: 7: Enotaeus 104

Sequence Mask: None
Numbering Mask: None

Start ruler at: 1

```
cenchria      CTTCGGATCAATGCTGCTCGCCTGCTCAGCCCTACAAGTACTAACCGGGTTCTTTCTAGCCGTACACTACACAGCAAA
CM2M5MAURUS   CTTCGGATCAATGCTGTTCGCCTGCTCAGCCCTACAAGTACTAACCGGGTTCTTTCTAGCCGTACACTACACAGCAAA
cv3m5venezue  CTTCGGATCAATGCTGCTCGCCTGCTCAGCCCTACAAGTACTAACCGGGTTCTTTCTGGCCGTACACTACACAGCAAA
CE1M5cenchri  TTTCGGATCAATACTGCTCGCCTGCTCAGCCCTACAAGTACTAACCGGGTTCTTCCTGGCCGTGCACTACACAGCAAA
CA4M5ALVAREZ  CTTCGGATCAATACTACTAGCCTGCTCAACCCTACAAGTACTAACAGGCTTCTTCCTAGCCGTACACTACACAGCAAA
Emurinus      CTTCGGATCTATGCTGCTTACCTGCTCAACCCTACAAGTACTAACTGGGTTTTTCCTAGCAGTTCACTACACAGCAAA
Enotaeus      CTTCGGATCCATACTAATTGCCTGCTCAACTCTACAAGTACTAACAGGCTTTTTTCTAGCAGTACATTATACAGCTAA
inornatus     CTTTGGATCAATACTACTCACTTGCTCAGCCCTACAAGTACTAACAGGCTTCTTCCTAGCTGTACACTATACAGCAAA
inorn-MUJ     CTTTGGATCAATACTACTCACCTGCTCAGCCCTACAAGTACTAACAGGCTTCTTCCTAGCTGTACACTATACAGCAAA
monensis      TTTCGGATCTATACTACTGACCTGCTCAGTCCTACAAGTACTAACAGGCTTCTTTCTAGCCGTACATTACACAGCAAA
subflavens    TTTTGGATCTATGCTACTCGCCTGCTCAACCCTACAAGTACTAACAGGATTCTTCTTAGCCGTACATTACACAGCAAA
UH1m5         CTTCGGATCCATACTACTCACTTGCTCAATACTACAAGTATTAACAGGCTTTTTCCTGGCTATACACTACACAGCCAA
F13M5         CTTCGGATCCATACTACTCACTTGCTCAATACTACAAGTATTAACAGGCTTTTTCCTGGCTATACACTACACAGCCAA
93m5          CTTCGGATCCATACTACTCACTTGCTCAATACTACAAGTATTAACAGGCTTTTTCCTGGCTATACACTACACAGCCAA
84m5          CTTCGGATCCATACTACTCACTTGCTCAATACTACAAGTATTAACAGGCTTTTTCCTGGCTATACACTACACAGCCAA
M2M5          CTTCGGATCCATACTACTCACTTGCTCAATTCTACAAGTATTAACAGGCTTTTTCCTGGCCATACACTATACAGCCAA
M1m5          CTTCGGATCCATACTACTCACTTGCTCAATTCTACAAGTATTAACAGGCTTTTTCCTGGCCATACACTATACAGCCAA
DM5           CTTCGGATCCATACTACTCACTTGCTCAATACTACAAGTATTAACAGGCTTTTTCCTGGCCATACACTACACAGCCAA
GM5           CTTCGGATCCATACTACTCACTTGCTCAATACTACAAGTATTAACAGGCTTTTTCCTGGCCATACACTACACAGCCAA
Hm5           CTTCGGATCCATACTACTCACTTGCTCAATACTACAAGTATTAACAGGCTTTTTCCTGGCCATACACTACACAGCCAA
U4M5          CTTCGGATCCATACTACTTACTTGCTCAATGCTACAAGTATTAACAGGCTTCTTCCTAGCTATACACTACACAGCTAA
M11M5         CTTCGGATCCATACTACTTACTTGCTCAATGCTACAAGTATTAACAGGCTTCTTCCTAGCTATACACTACACAGCTAA
M6m5          CTTCGGATCCATACTACTTACTTGCTCAATGCTACAAGTATTAACAGGCTTCTTCCTAGCTATACACTACACAGCTAA
E42M5         CTTCGGATCCATACTACTTACTTGCTCAATGCTACAAGTATTAACAGGCTTCTTCCTAGCTATACACTACACAGCTAA
ex1m5         CTTCGGATCCATACTACTTACTTGCTCAATGCTACAAGTATTAACAGGCTTCTTCCTAGCTATACACTACACAGCTAA
F12m5         CTTCGGATCCATACTACTTACTTGCTCAATGCTACAAGTATTAACAGGCTTCTTCCTAGCTATACACTACACAGCTAA
F3m5          CTTCGGATCCATACTACTTACTTGCTCAATGCTACAAGTATTAACAGGCTTCTTCCTAGCTATACACTACACAGCTAA
19m5          CTTCGGATCCATACTACTTACTTGCTCAATGCTACAAGTATTAACAGGCTTCTTCCTAGCTATACACTACACAGCTAA
77m5          CTTCGGATCCATACTACTTACTTGCTCAATGCTACAAGTATTAACAGGCTTCTTCCTAGCTATACACTACACAGCTAA
18M5          CTTCGGATCCATACTACTTACTTGCTCAATGCTACAAGTATTAACAGGCTTCTTCCTAGCTATACACTACACAGCTAA
17m5          CTTCGGATCCATACTACTTACTTGCTCAATGCTACAAGTATTAACAGGCTTCTTCCTAGCTATACACTACACAGCTAA
E43m5         CTTCGGATCCATACTACTTACTTGCTCAATGCTGCAAGTATTAACAGGCTTCTTCCTAGCTATACACTACACAGCTAA
UH0m5         TTTCGGATCCATACTACTTACTTGCTCAATACTACAAGTATTAACAGGCTTCTTCCTAGCTATACACTACACAGCTAA
```

Phylogeny estimation

two types of methods

character based (maximum parsimony, maximum likelihood, Bayesian analysis)

distance based (Neighbour-joining, UPGMA)

Two different approaches:
algorithm – number of specific steps resulting in one best tree
methods: UPGMA, Neighbour-joining

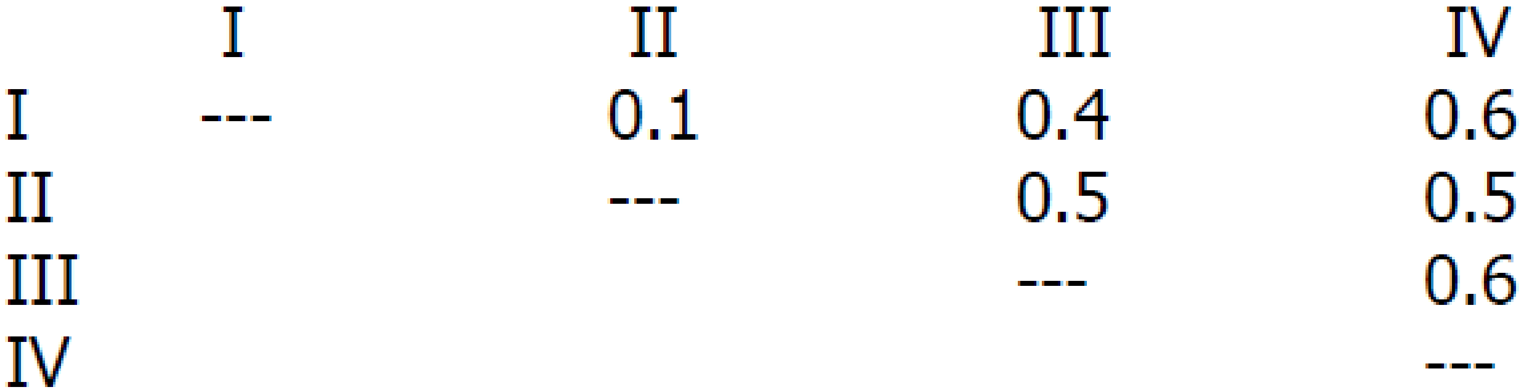
optimality criterion – consider and compare all theoretically possible trees based on selected criteria (number of evolutionary steps, likelihood value) and select the best one

distances

input is a matrix of distances between species

taxon

taxon



	I	II	III	IV
I	---	0.1	0.4	0.6
II		---	0.5	0.5
III			---	0.6
IV				---

proportional (p) distance

number of substitutions between sequences

p= total number of base differences/total no. of available sites

$$p = n_d/n$$

GATCATTAATGCGATAT
GACCGTTATTGCCATAT

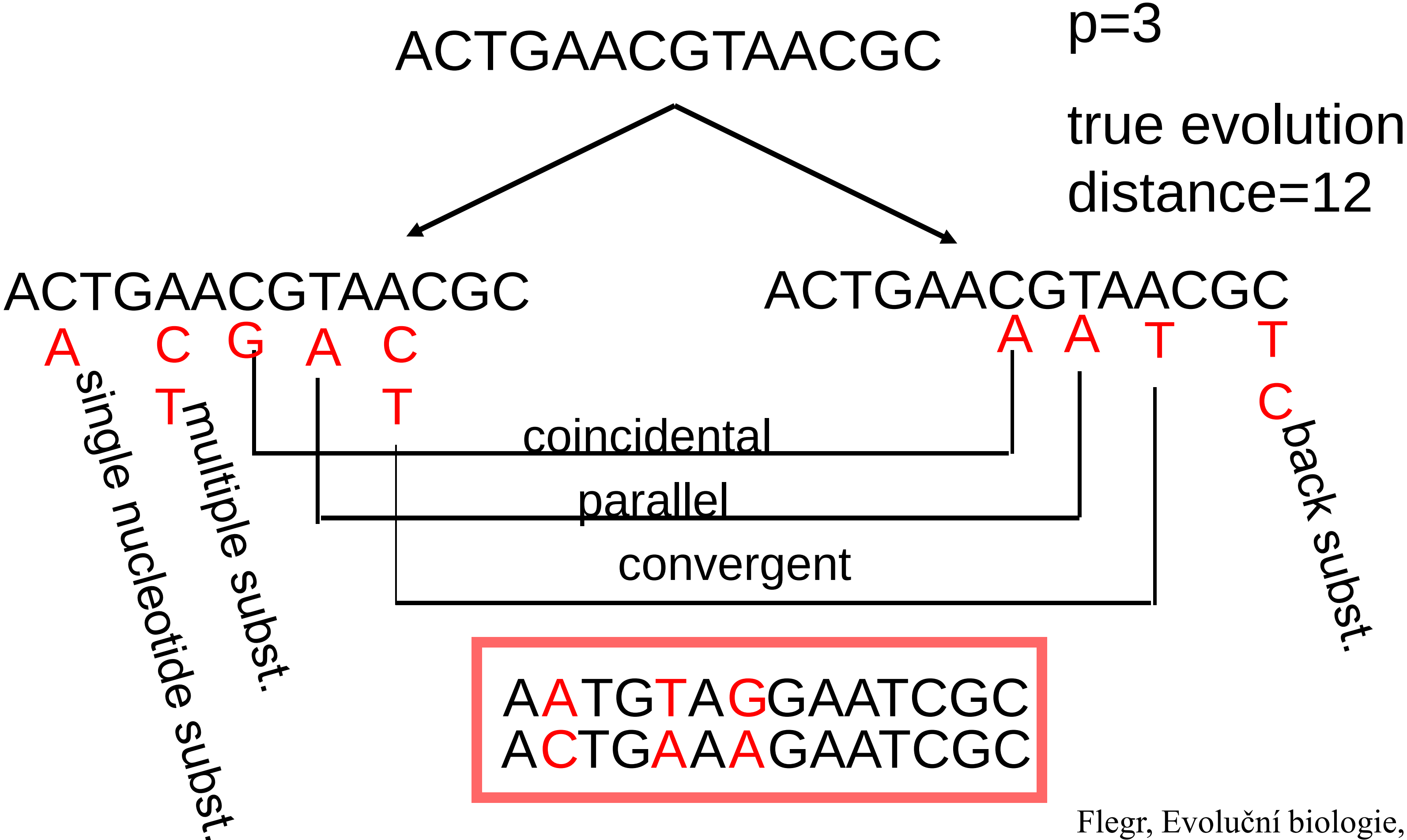
$$p = 4/17 = 0.23$$

real number of substitutions in the sequence over time is usually higher than observed p distance

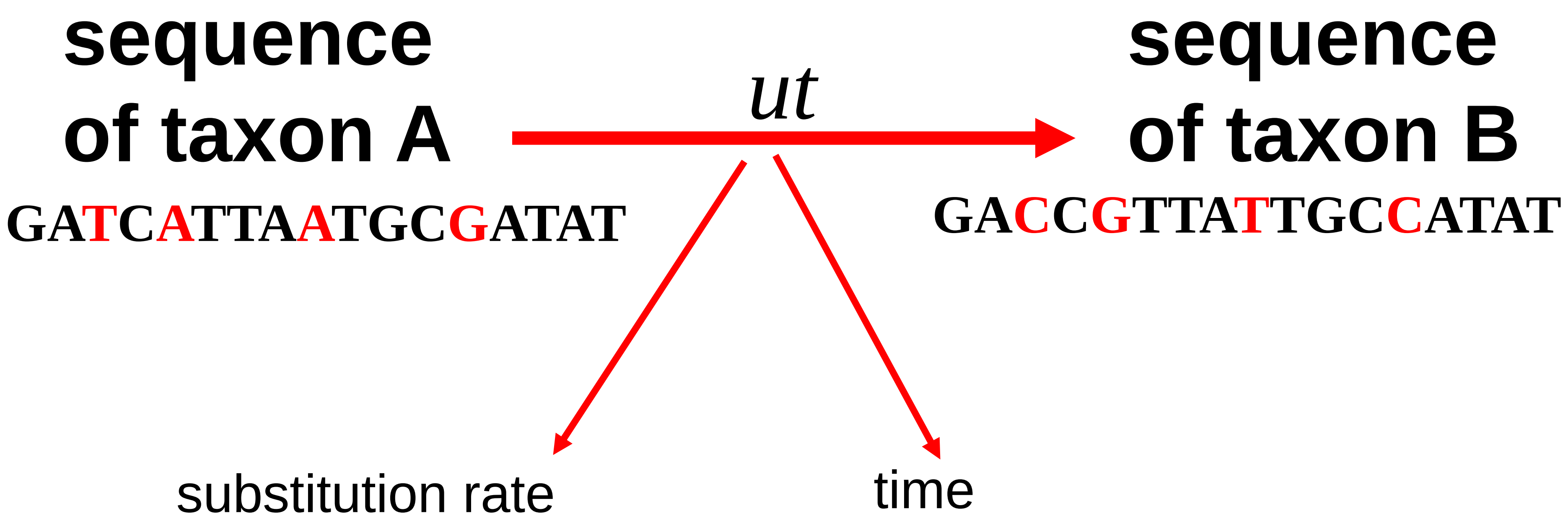
we can see just 3 differences (*p*), but in fact there was 12 substitutions

p=3

true evolutionary distance=12



Flegr, Evoluční biologie, Academia



sequence of taxon A

GATCATTAATGCGATAT

ut

sequence of taxon B

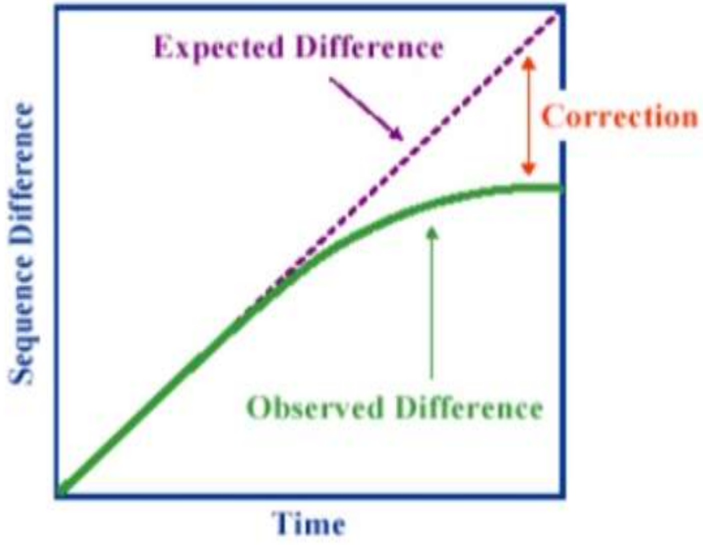
GACCGTTATTGCCATAT

substitution rate

time

→ in phylogenetic analyses we use “correction“ of observed distances to estimate number of hidden changes (multiple mutations etc.)

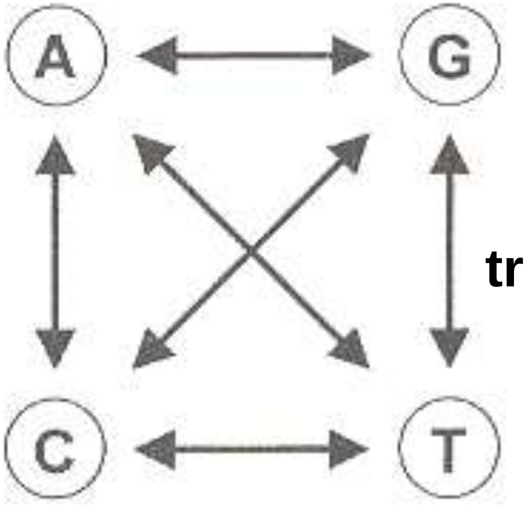
correction based on different substitution type (Ts, Tv), different substitution rate, frequencies of nucleotides



Examples:

Jukes-Cantor model (distance)

all substitution types and base frequencies are presumed equal



transition

transversion

transition

JC distance

$$d_{JC} = -\frac{3}{4}\ln(1-\frac{4}{3}p)$$

Kimura 2-parameter model (K2P):

transitions are more likely than transversions, equal base frequencies

K2P distance

$P = n_{TS} / n$

$Q = n_{TV} / n$

$$d_{K2P} = 0.5\ln\left(\frac{1}{1-2P-Q}\right) + 0.25\ln\left(\frac{1}{1-2Q}\right)$$

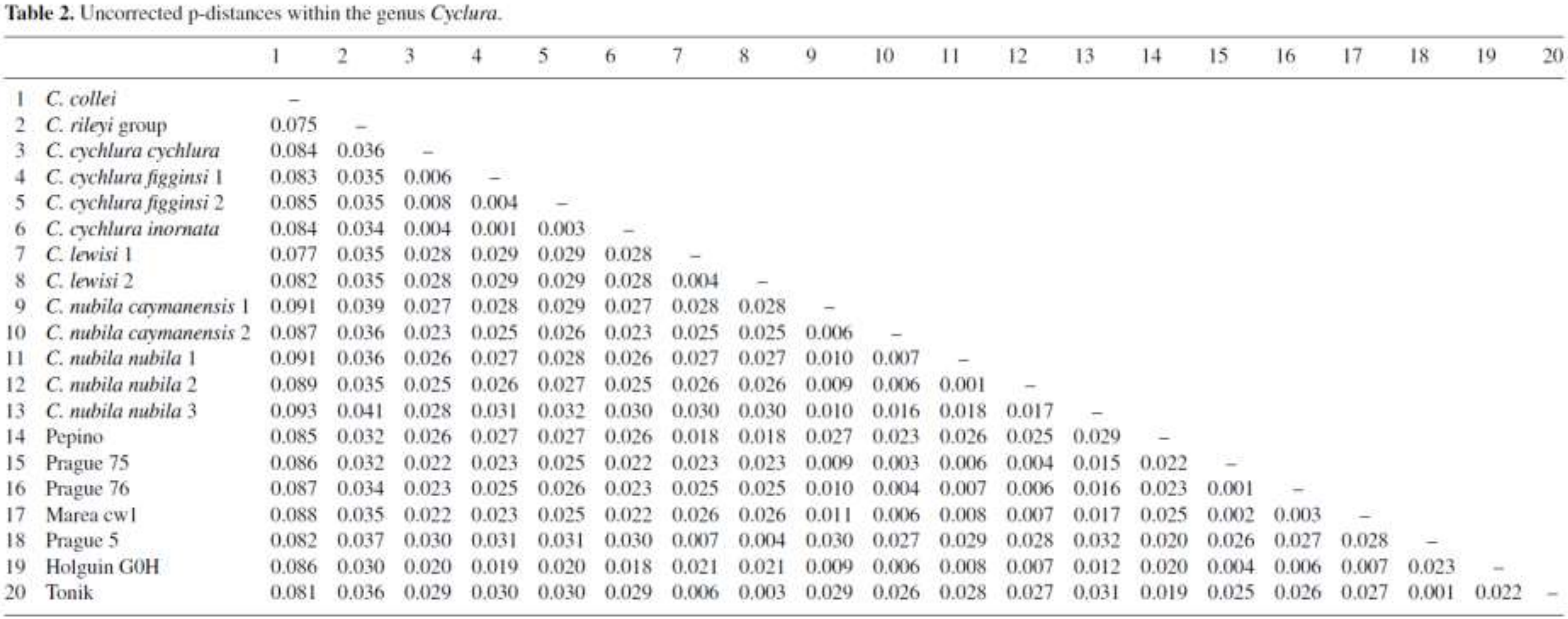


Table 2. Uncorrected p-distances within the genus *Cyclura*.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	*C. collei*	–																			
2	*C. rileyi* group	0.075	–																		
3	*C. cychlura cychlura*	0.084	0.036	–																	
4	*C. cychlura figginsi* 1	0.083	0.035	0.006	–																
5	*C. cychlura figginsi* 2	0.085	0.035	0.008	0.004	–															
6	*C. cychlura inornata*	0.084	0.034	0.004	0.001	0.003	–														
7	*C. lewisi* 1	0.077	0.035	0.028	0.029	0.029	0.028	–													
8	*C. lewisi* 2	0.082	0.035	0.028	0.029	0.029	0.028	0.004	–												
9	*C. nubila caymanensis* 1	0.091	0.039	0.027	0.028	0.029	0.027	0.028	0.028	–											
10	*C. nubila caymanensis* 2	0.087	0.036	0.023	0.025	0.026	0.023	0.025	0.025	0.006	–										
11	*C. nubila nubila* 1	0.091	0.036	0.026	0.027	0.028	0.026	0.027	0.027	0.010	0.007	–									
12	*C. nubila nubila* 2	0.089	0.035	0.025	0.026	0.027	0.025	0.026	0.026	0.009	0.006	0.001	–								
13	*C. nubila nubila* 3	0.093	0.041	0.028	0.031	0.032	0.030	0.030	0.030	0.010	0.016	0.018	0.017	–							
14	Pepino	0.085	0.032	0.026	0.027	0.027	0.026	0.018	0.018	0.027	0.023	0.026	0.025	0.029	–						
15	Prague 75	0.086	0.032	0.022	0.023	0.025	0.022	0.023	0.023	0.009	0.003	0.006	0.004	0.015	0.022	–					
16	Prague 76	0.087	0.034	0.023	0.025	0.026	0.023	0.025	0.025	0.010	0.004	0.007	0.006	0.016	0.023	0.001	–				
17	Marea cw1	0.088	0.035	0.022	0.023	0.025	0.022	0.026	0.026	0.011	0.006	0.008	0.007	0.017	0.025	0.002	0.003	–			
18	Prague 5	0.082	0.037	0.030	0.031	0.031	0.030	0.007	0.004	0.030	0.027	0.029	0.028	0.032	0.020	0.026	0.027	0.028	–		
19	Holguin G0H	0.086	0.030	0.020	0.019	0.020	0.018	0.021	0.021	0.009	0.006	0.008	0.007	0.012	0.020	0.004	0.006	0.007	0.023	–	
20	Tonik	0.081	0.036	0.029	0.030	0.030	0.029	0.006	0.003	0.029	0.026	0.028	0.027	0.031	0.019	0.025	0.026	0.027	0.001	0.022	–

Starostová et al. 2010, *Amphibia-Reptilia* 31:134-143

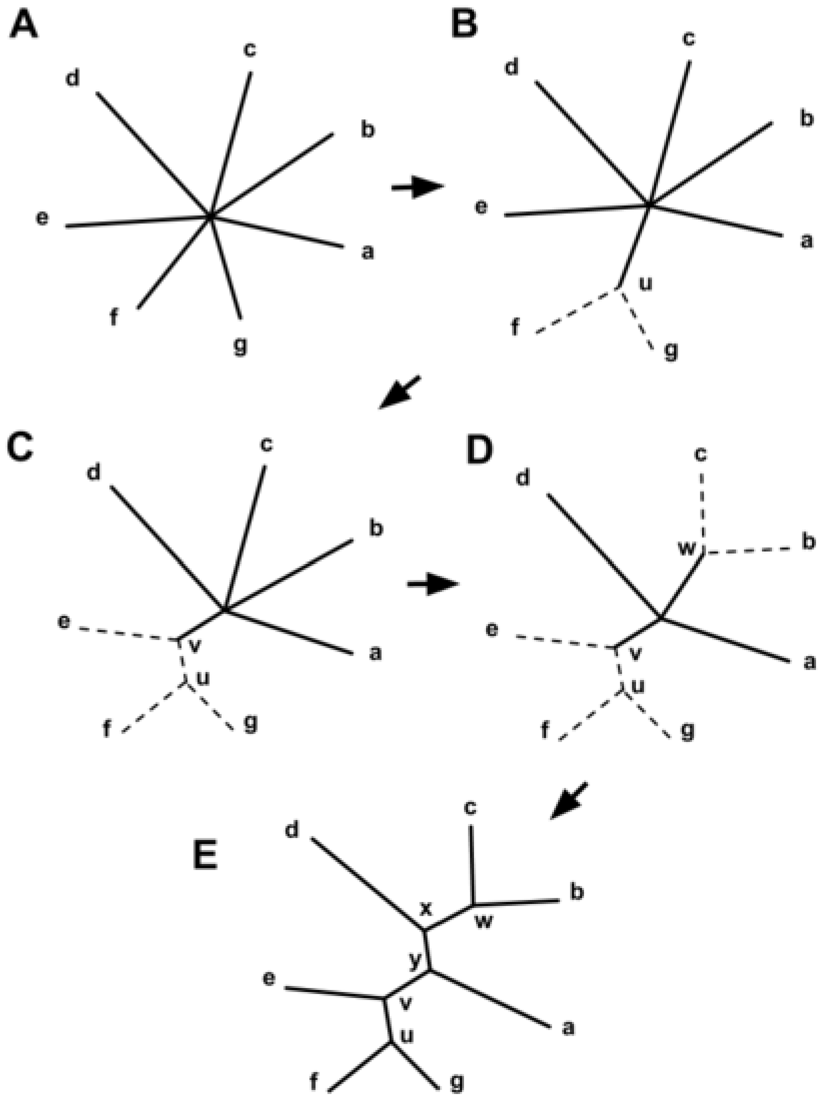
methods

Neighbour-joining (NJ) - the fully resolved tree is “decomposed” from a fully unresolved “star” tree by successively inserting

branches between a pair of closest neighbors and the

remaining terminals in the tree

result is one tree



- **other methods: UPGMA** (Unweighted Pair Group Method using Arithmetic means)**, Minimal evolution**

conclusion, pros and cons

distance methods rely on evolutionary models (distance corrections) to estimate the numbers of multiple/parallel… substitutions – the result is dependent on how well the accepted models match the actual evolutionary properties of the sequences

only one tree is derived

discards the primary character data

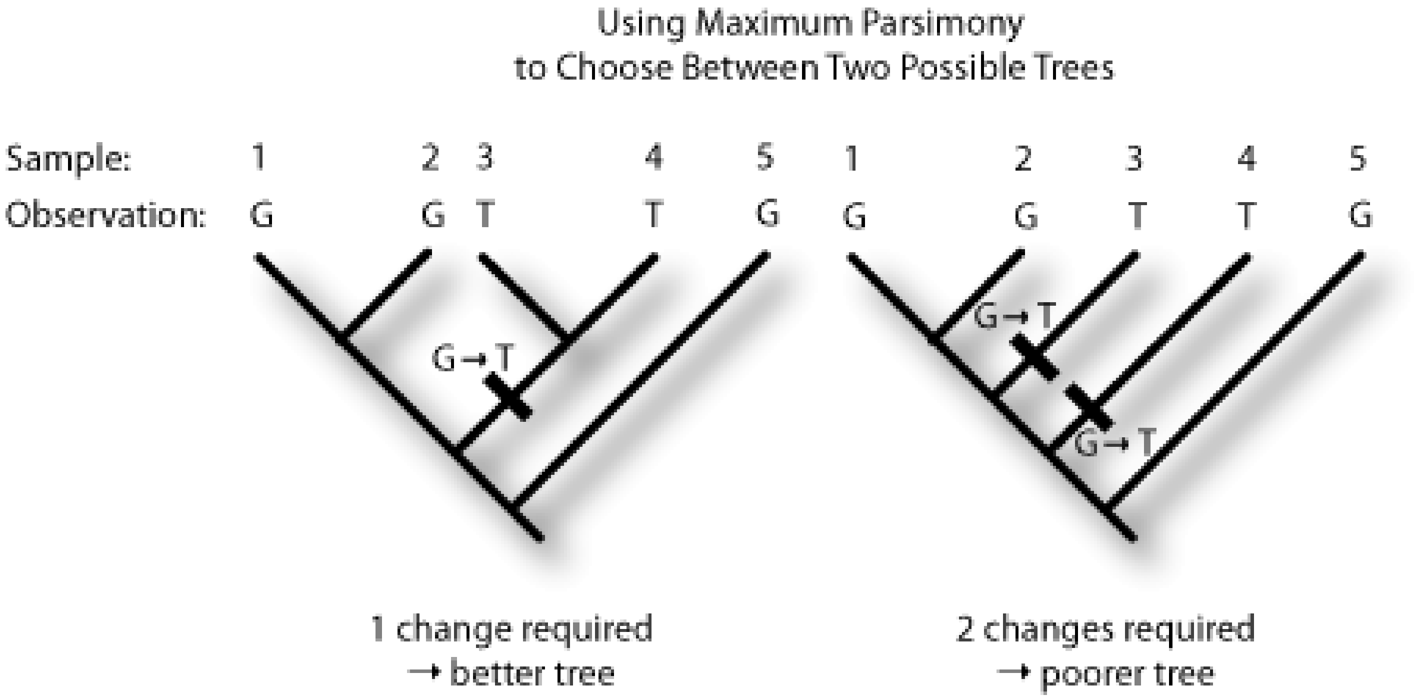
problem with interpretation of branch lengths

very fast, ideal for the first insight

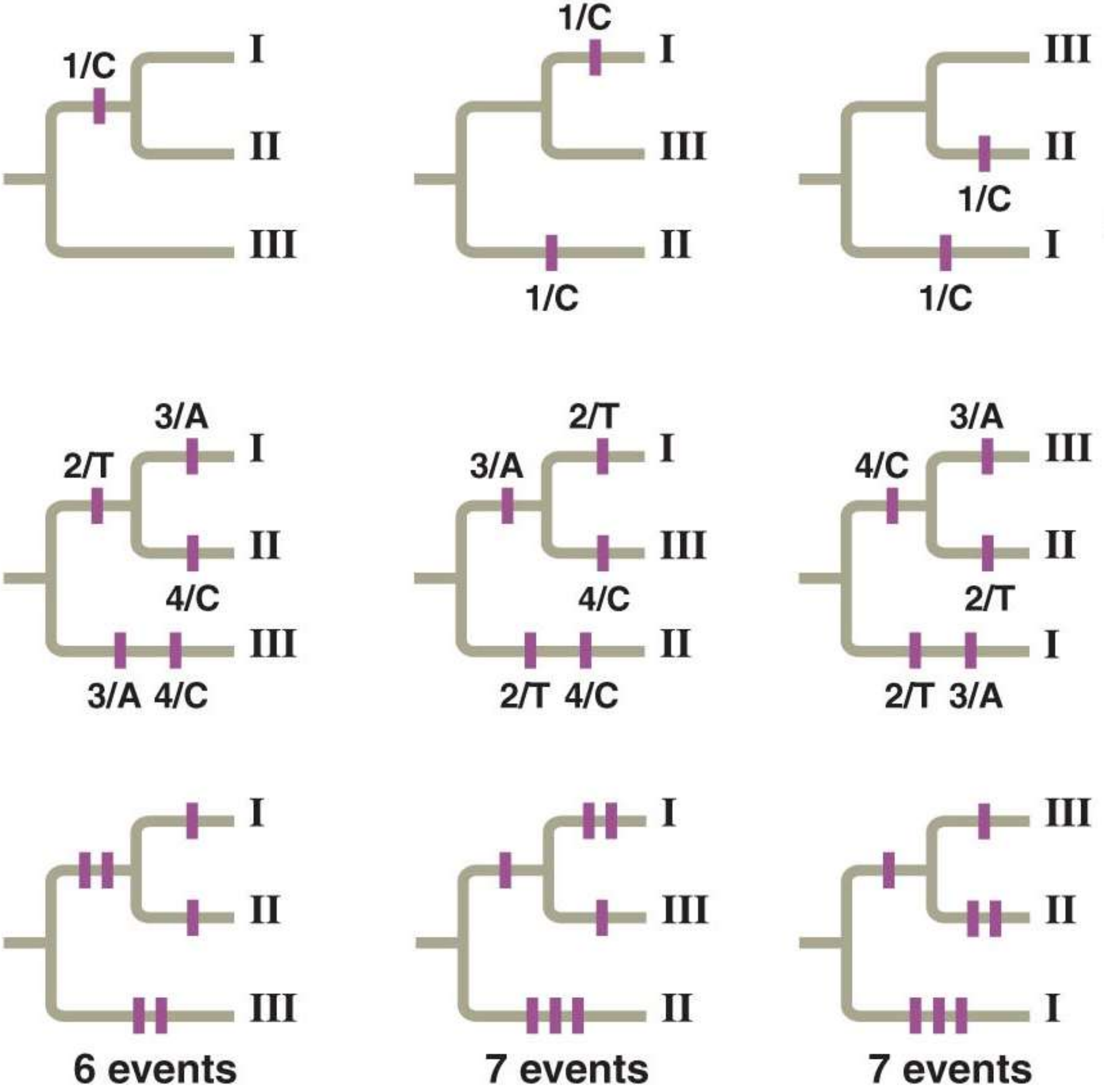
Maximum parsimony:

optimality criterion - parsimony score = minimum number of events (steps) required by a tree to explain the variation in the data

search for topologies that minimize the total tree length assuming a minimum number of base changes
“Occam’s Razor” – “keep it simple“



	Site 1	 2	 3	 4
Species I	C	T	A	T
Species II	C	T	T	C
Species III	A	G	A	C
Ancestral sequence	A	G	T	T



not all characters are good for parsimony:

the alignment is checked for **informative positions**

= a site must have at least two different character states (nucleotides for DNA), the same character states in at least two taxa

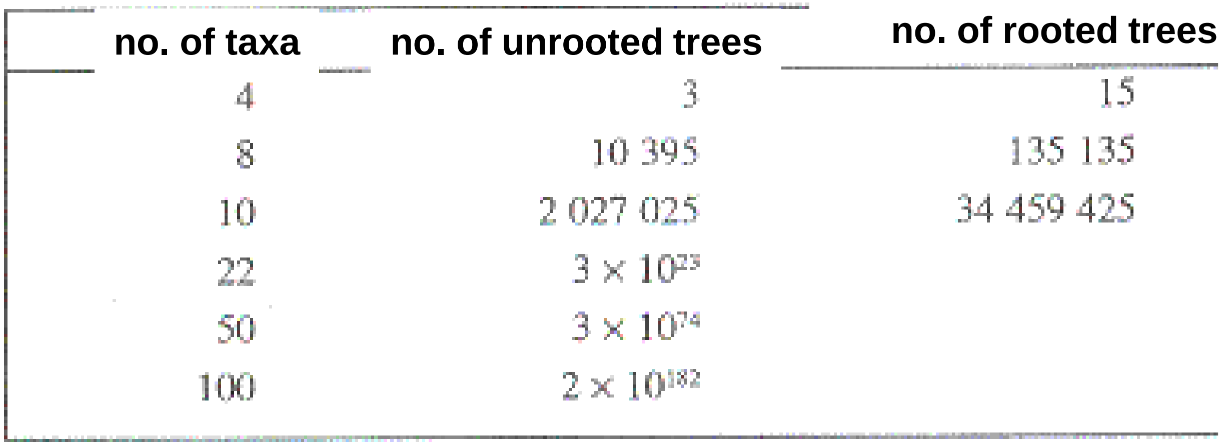
Maximum parsimony:

optimality criterion - parsimony score = minimum number of events (steps) required by a tree to explain the variation in the data

search for topologies that minimize the total tree length assuming a minimum number of base changes
“Occam’s Razor” – “keep it simple“

$$\frac{(2n - 3)!}{2^{n-2}(n-2)!}$$

We already know that there are a lot of possible trees- in most cases we can not compare all of them

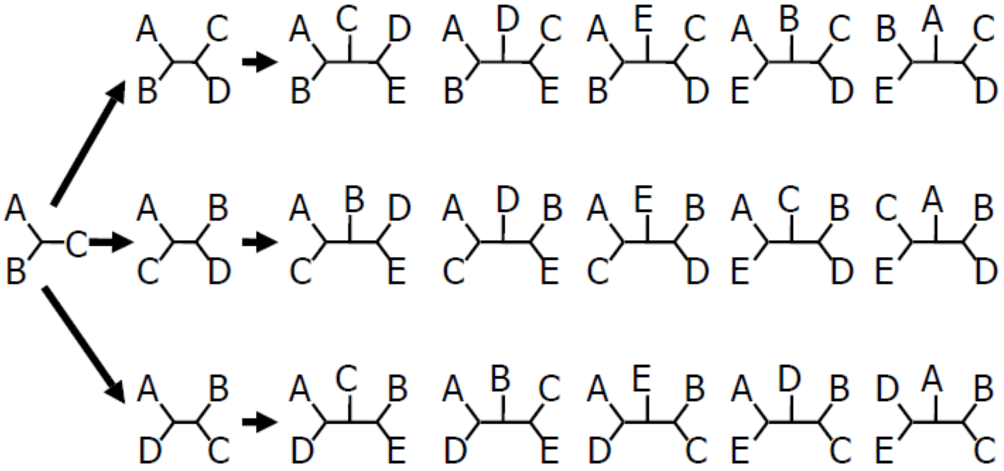


no. of taxa	no. of unrooted trees	no. of rooted trees
4	3	15
8	10 395	135 135
10	2 027 025	34 459 425
22	3×10^{23}	
50	3×10^{74}	
100	2×10^{182}	

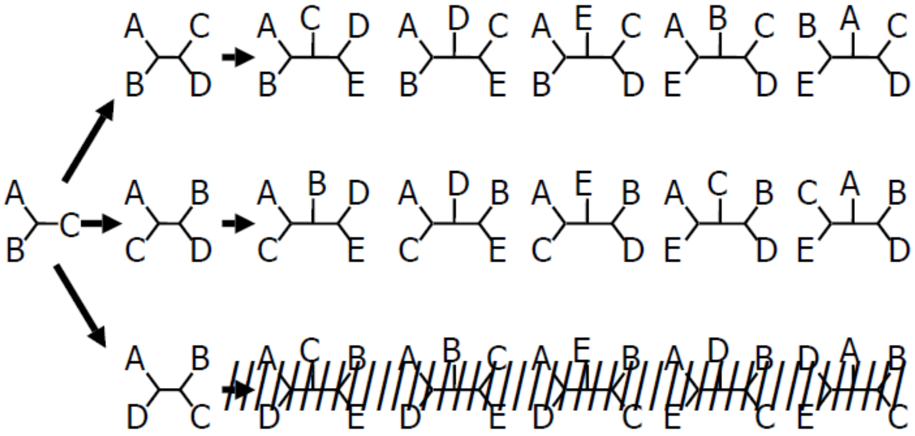
no. of trees exponentially increases

Tree searching

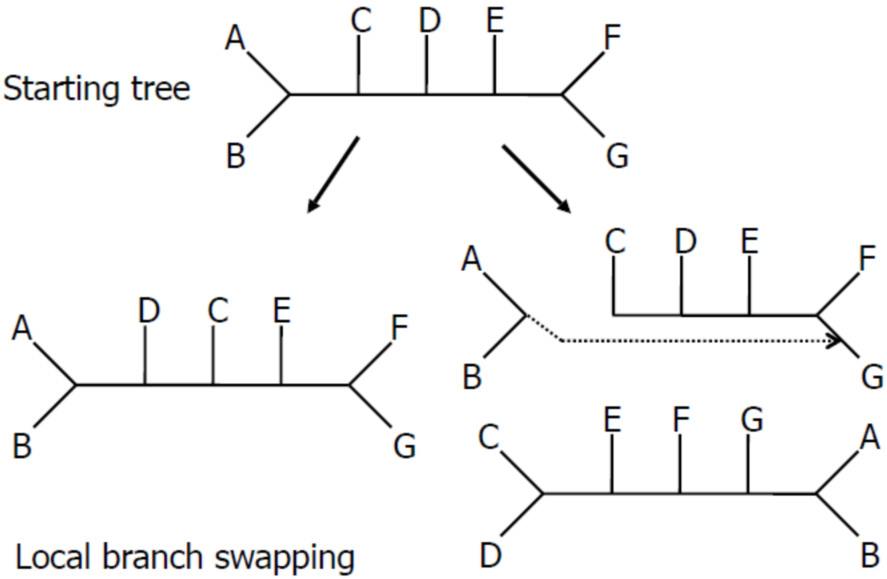
Exhaustive Searching



Branch and Bound Searching



Heuristic Searching



Local branch swapping
Swofford et al. (1996)

Global branch swapping

Maximum parsimony

in most cases we can not compare all trees

⇒ **e.g. heuristic search**

- create random tree
- calculate parsimony score
- rearranging of the tree,
- calculate parsimony score
- further the method works with the better (shorter) tree
- repeated rearranging and calculating scores
- at the end shortest tree

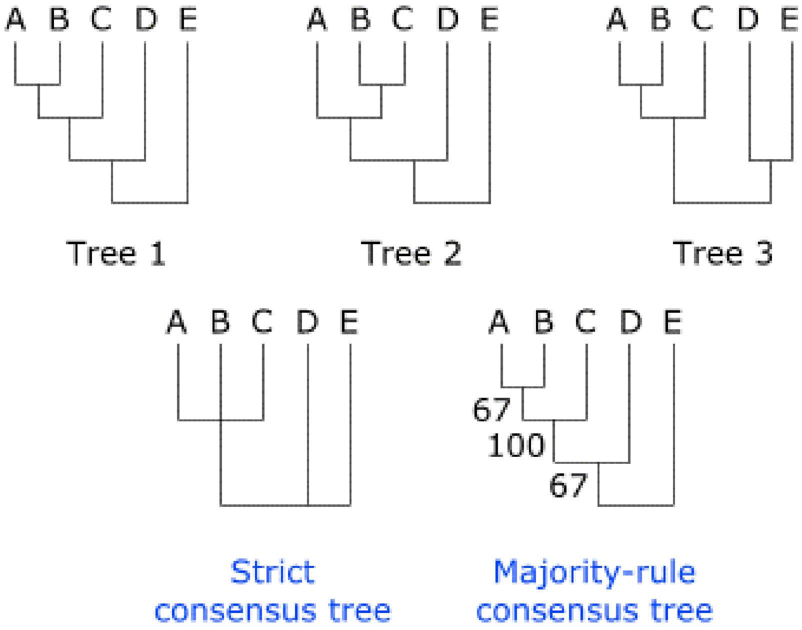
Sometimes (quite often) we find more equal trees →

Consensus tree:

when multiple phylogenies are supported - a consensus tree shows only those relationships common to all trees (based on our settings)

- **strict consensus** (only relationships common to all trees)

- **majority-rule consensus** (relationships common to more than 50 % of trees are shown)



Parsimony: pros and cons

⊕

works directly with characters

straightforward, well understood principle

relatively fast

does not need a model of evolution (but not really model free – change is rare)

⊖

performs weakly on distantly related data

long branch attraction

can produce many trees with the same parsimony score

long branch attraction (LBA)

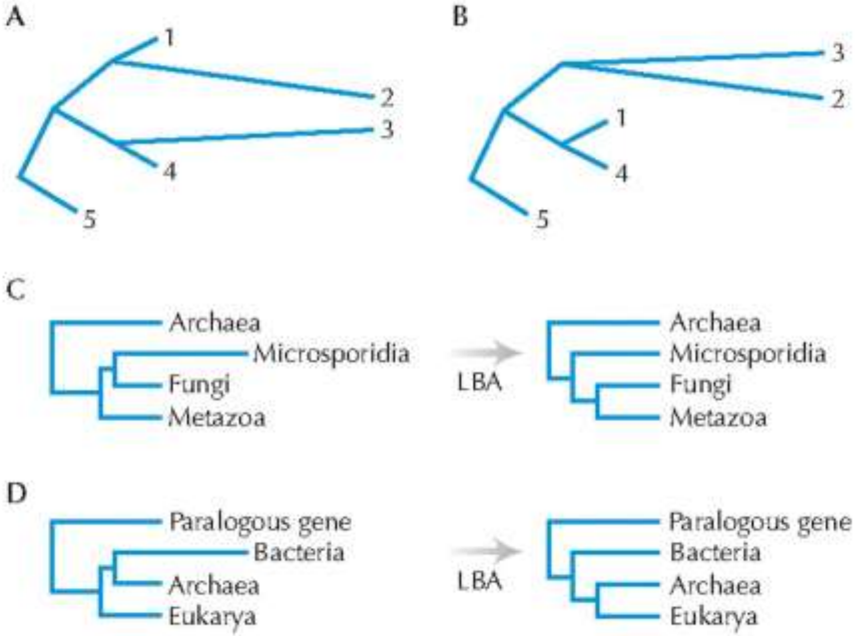


FIGURE 5.22. Long-branch attraction is a methodological artifact that can cause phylogenetic trees to inaccurately portray evolutionary history. The phenomenon causes errors in phylogenetic reconstruction when two (or more) of the entities being studied lie on the end of long branches in their "real" tree but are not sister taxa. (*A*) In this hypothetical "real" tree of five species, species 2 and 3 (which are not sister taxa, as indicated) have undergone higher rates of evolution than the other three, and thus sit at the end of longer branches. Many phylogenetic reconstruction methods used to infer the evolution of species will cause the long branches to appear to be closely related and thus produce an incorrect tree (as shown in *B*). (*C*) In studies of the evolution of microsporidia (a relative of fungi, *left tree*), long-branch attraction (LBA) is believed to have erroneously identified them as deeply branching eukaryotes (*right tree*). (The evolution of microsporidia is discussed in more detail on p. 198.) (*D*) In trees of anciently duplicated genes, long-branch attraction might have pulled bacteria down to the paralogs used to root the tree, because the paralogs are at the end of a long branch (*right tree*). This would occur if bacteria evolved at a higher rate than archaea and eukaryotes (as suggested in the *left tree*).

5.22C,D, redrawn from Forterre P. et al., *Bioessays* **21:** 871–879, © 1999 Wiley-Liss, Inc.

Evolution © 2007 Cold Spring Harbor Laboratory Press

Maximum likelihood - ML

- **method compares possible phylogenetic trees on the basis of their ability to predict the observed data. The tree that has the highest probability of producing the observed sequences is preferred.**

- maximum likelihood reconstructs ancestors at all nodes of each considered tree, but it also assigns branch lengths based on the probabilities of mutations. For each possible tree topology, the assumed substitution rates are varied to find the parameters that give the highest likelihood of producing the observed sequences.

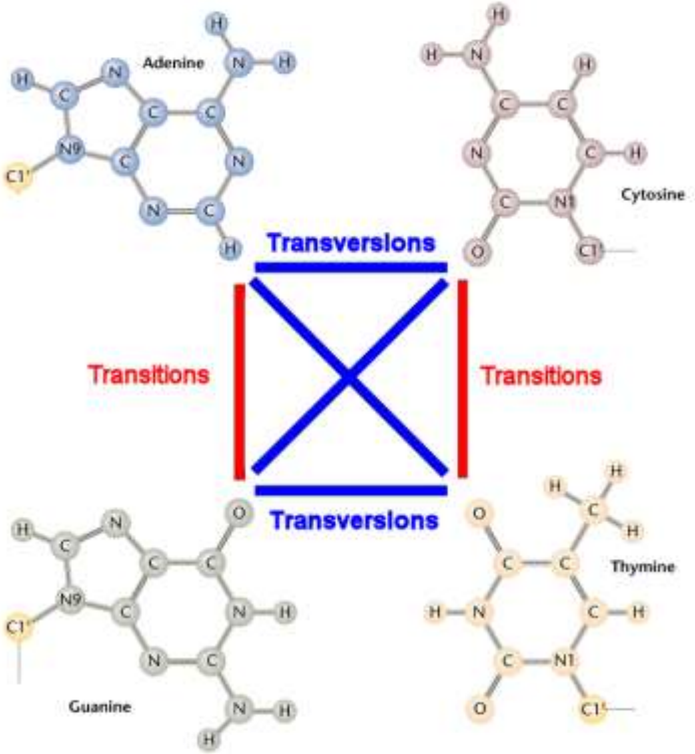
Likelihood describes how well the model predicts the data
- it prefers higher likelihood above the lower one
ML results in only 1 tree with branch lengths

Maximum likelihood - ML

- **ML uses model of sequence evolution (substitution model)**
- several programs (Modeltest, jModeltest, MrAIC…)
  programs examine the goodness of fit of the model to the data

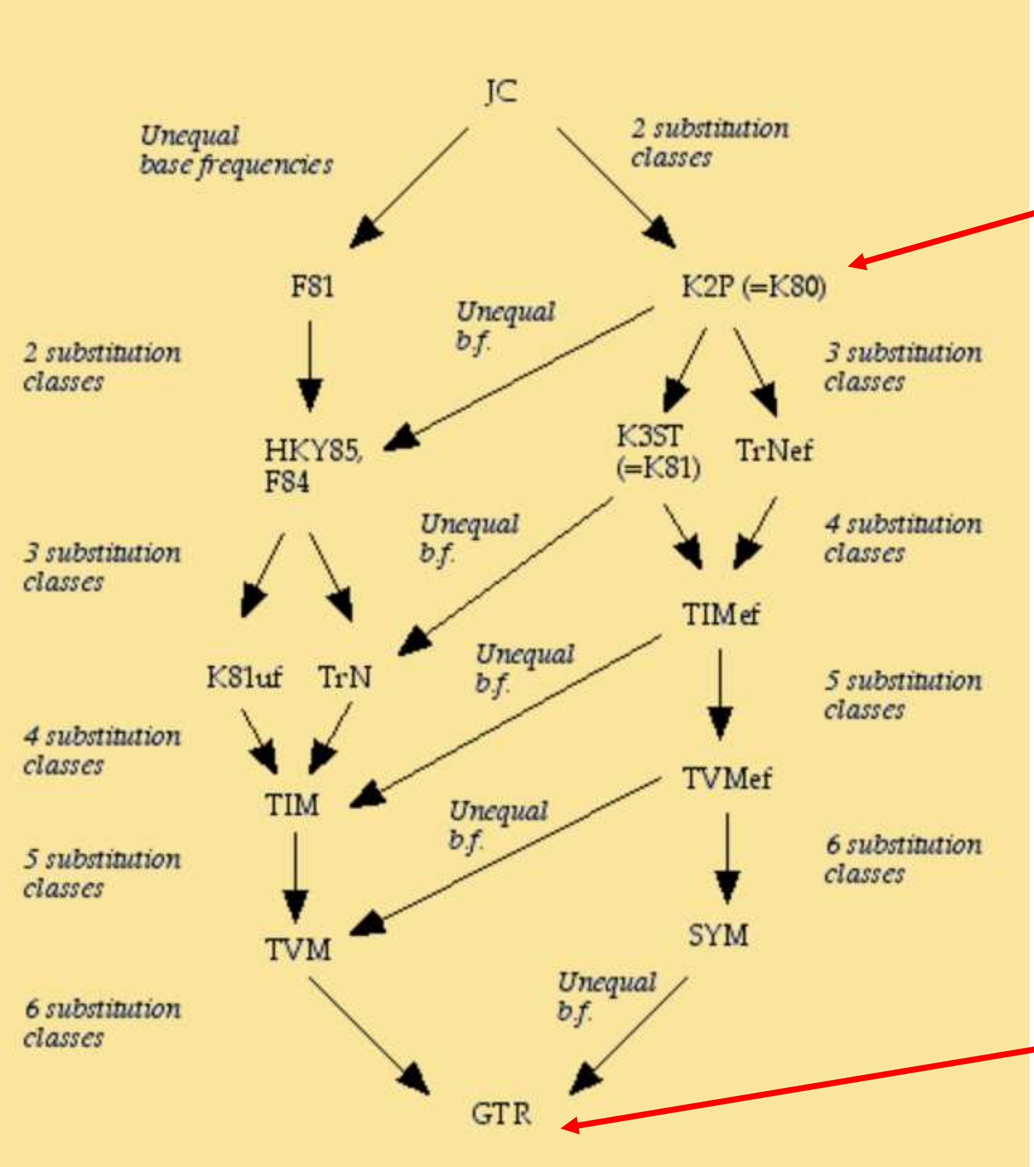
- models differ in:
- base frequencies
- probability of nucleotides changes (transition x transversion)
- heterogeneousness in different parts of sequence or in different position

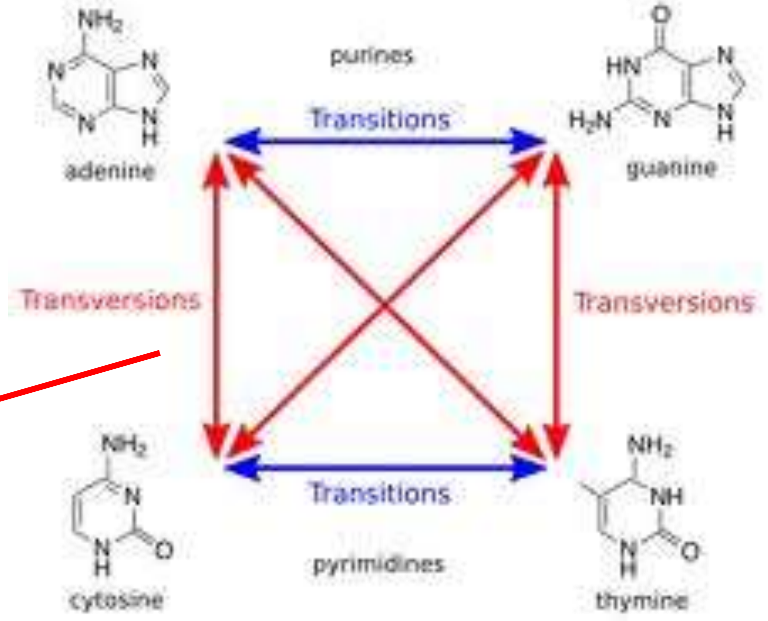
Model examples:
Jukes-Cantor (JC),
Kimura 2-parametres model (K2P),
General time-reversible model (GTR)



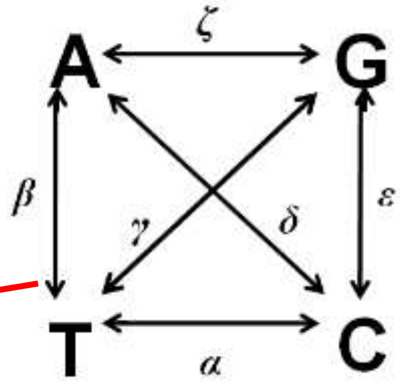
Models of sequence evolution

- models are nested, one is a special case of the other





Kimura 2-parametres model



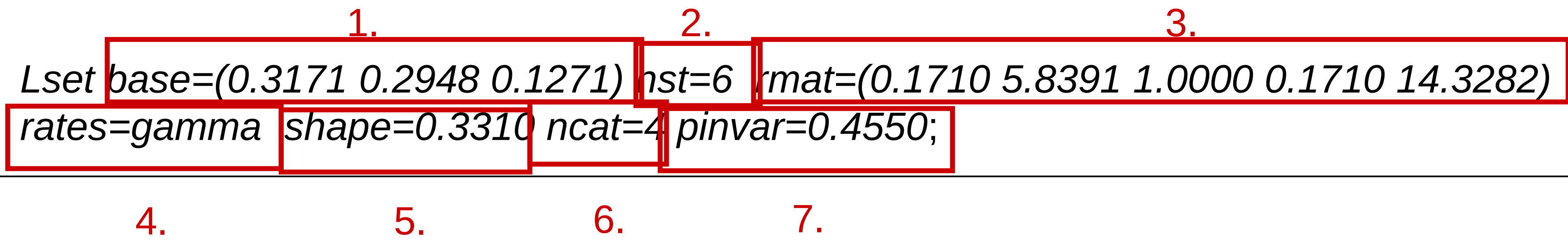
General time-reversible model

Best model selection

program jModeltest (Modeltest)

Example of model:

Lset base=(0.3171 0.2948 0.1271) nst=6 rmat=(0.1710 5.8391 1.0000 0.1710 14.3282)
rates=gamma shape=0.3310 ncat=4 pinvar=0.4550;



1. 2. 3.

Lset base=(0.3171 0.2948 0.1271) nst=6 rmat=(0.1710 5.8391 1.0000 0.1710 14.3282)
rates=gamma shape=0.3310 ncat=4 pinvar=0.4550;

4. 5. 6. 7.

1. - relative base composition (4th is 1-(fr1st+fr2nd+fr3rd))

2. - No. of substitution types (1 = same probability for all bases,
6 = every substitution has different probability)

3. - substitution rate matrix – rate of changes of each type of bases in alignment

4. - probability of changes distribution in individual positions
(equal = equal for all position, gamma = with different
gama distribution, invgamma)

5. - shape of gamma distribution

6. - gamma distribution category

7. - ratio (proportion) of invariable sites

Maximum likelihood

⊕

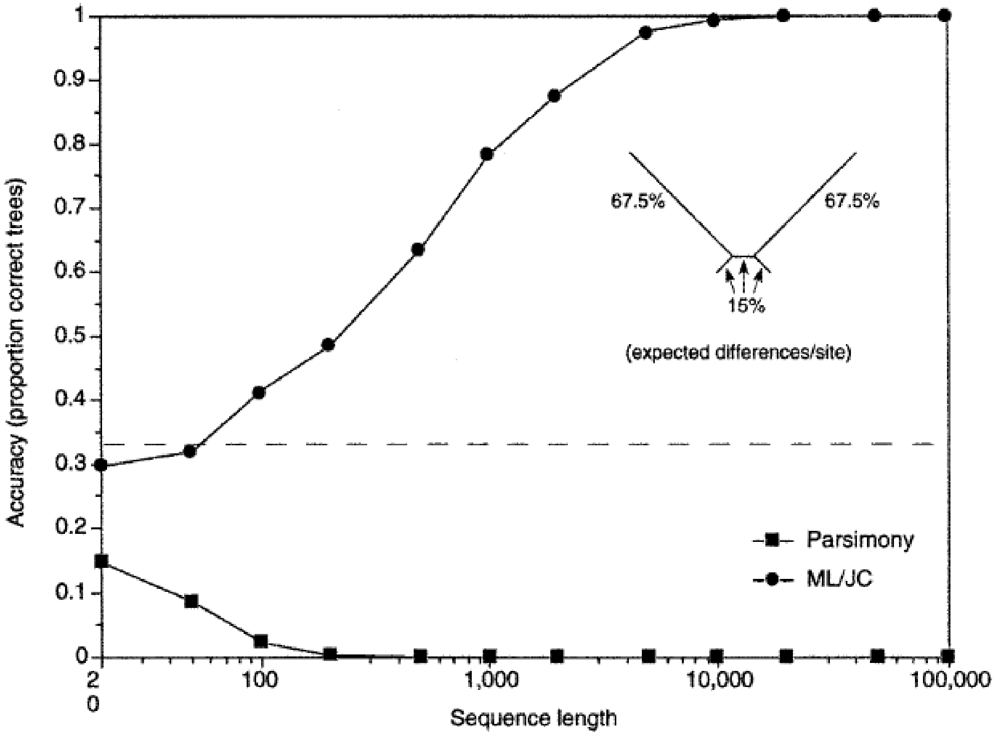
pros

- a lot of possible models of sequence evolution, robust to deviations from the model

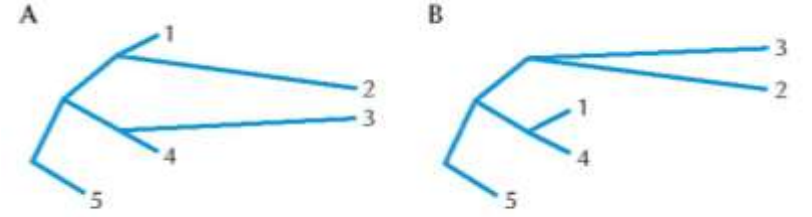
⊖

cons

- computationally demanding, slow (nowadays not so big problem)



ML method can decrease effect of LBA



Swofford et al,. *Systematic Biology*, 2001

reliability tests

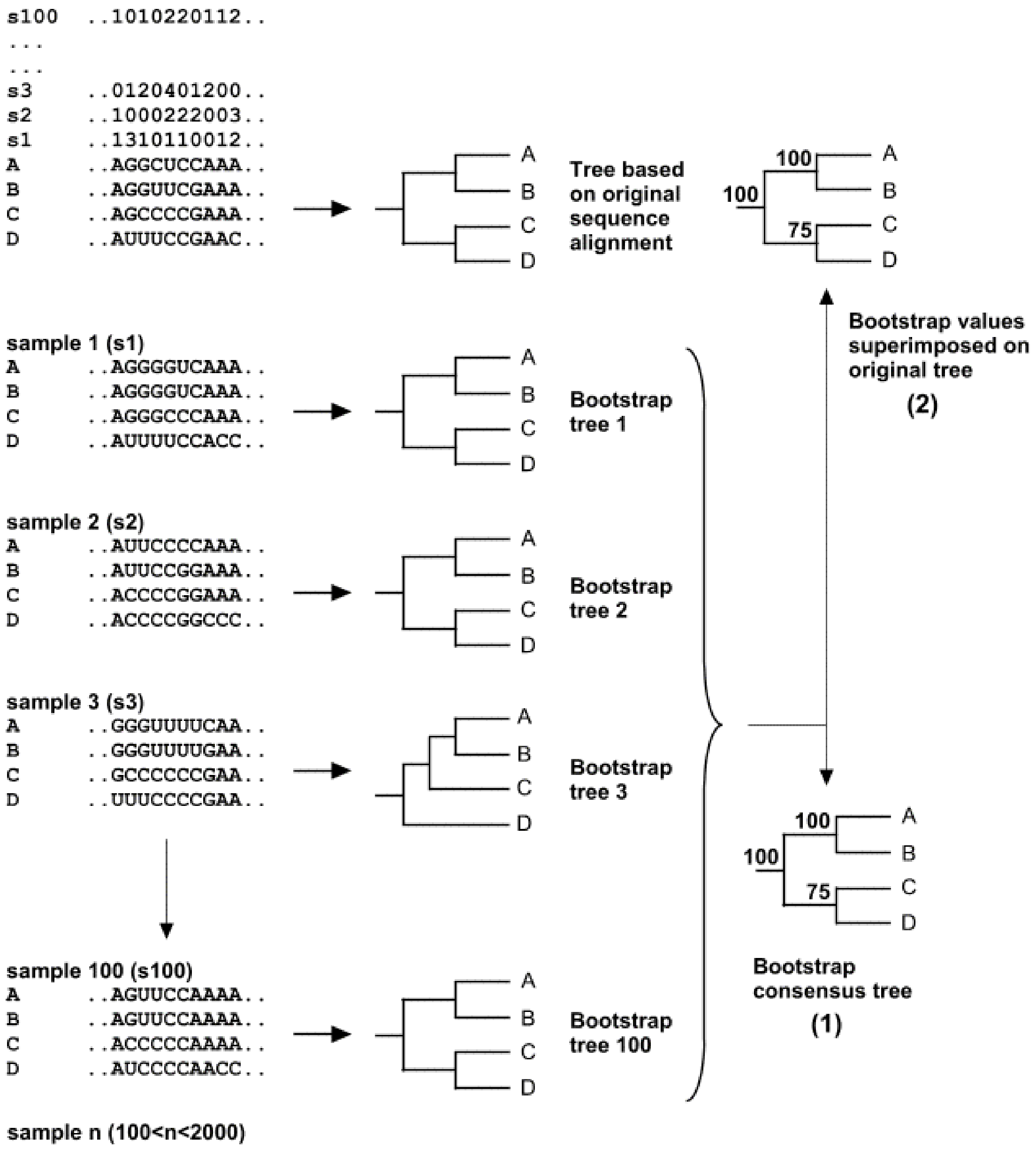
-nonparametric resampling methods - bootstrapping, jackknifing

→ new data sets are created from the original data set by sampling columns of characters at random

- each site can be sampled again with the same probability as any of the other sites

Box 3

Bootstrap Analysis (Felsenstein, 1985)



```
s100   ..1010220112..
...
...
s3     ..0120401200..
s2     ..1000222003..
s1     ..1310110012..
A      ..AGGCUCCAAA..
B      ..AGGUUCGAAA..
C      ..AGCCCCGAAA..
D      ..AUUUCCGAAC..
```

sample 1 (s1)

```
A      ..AGGGGUCAAA..
B      ..AGGGGUCAAA..
C      ..AGGGCCCAAA..
D      ..AUUUUCCACC..
```

sample 2 (s2)

```
A      ..AUUCCCCAAA..
B      ..AUUCCGGAAA..
C      ..ACCCCGGAAA..
D      ..ACCCCGGCCC..
```

sample 3 (s3)

```
A      ..GGGUUUUCAA..
B      ..GGGUUUUGAA..
C      ..GCCCCCCGAA..
D      ..UUUCCCCGAA..
```

sample 100 (s100)

```
A      ..AGUUCCAAAA..
B      ..AGUUCCAAAA..
C      ..ACCCCCAAAA..
D      ..AUCCCCAACC..
```

sample n (100<n<2000)

Tree based on original sequence alignment

Bootstrap tree 1

Bootstrap tree 2

Bootstrap tree 3

Bootstrap tree 100

Bootstrap values superimposed on original tree

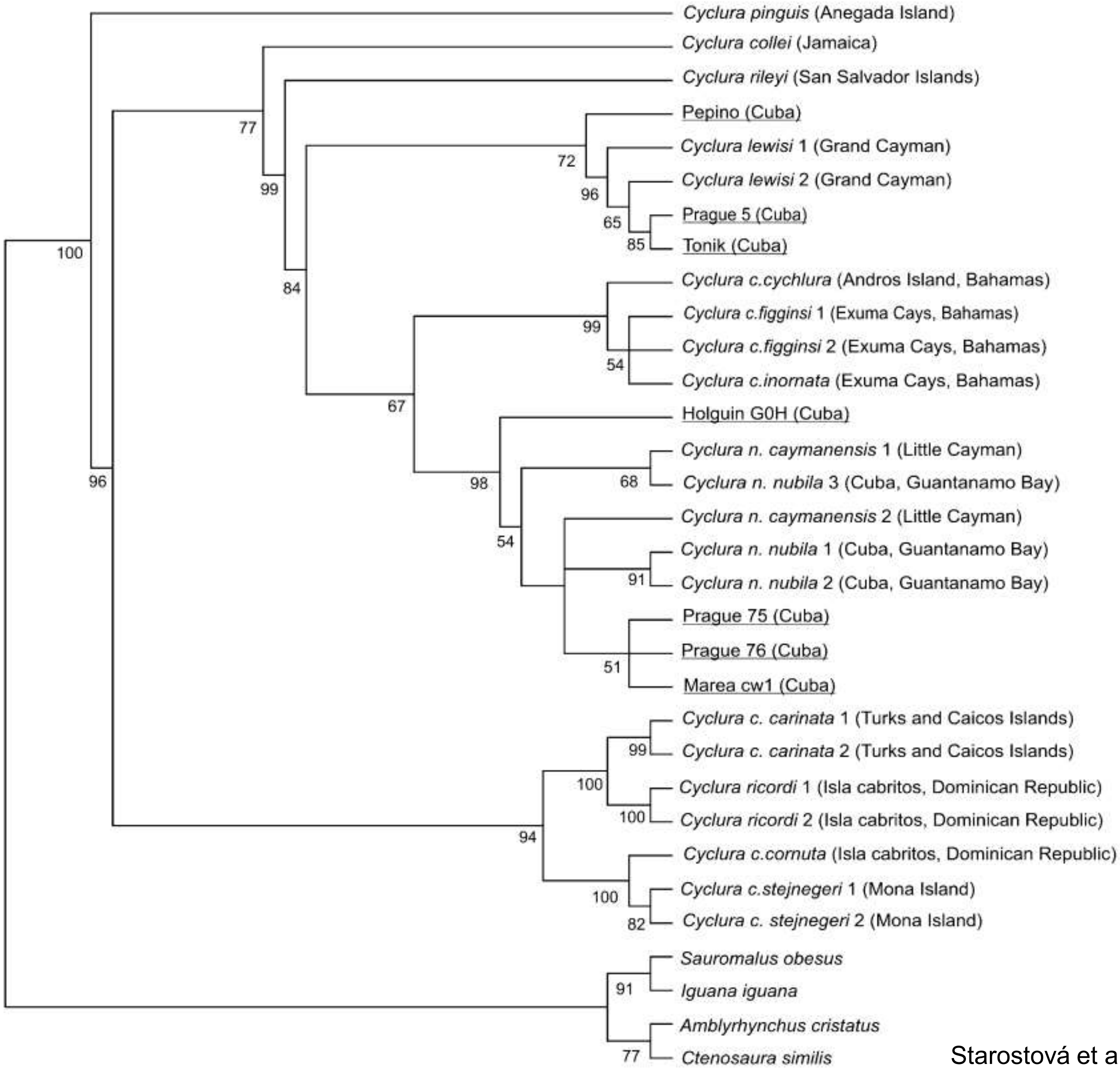
(2)

Bootstrap consensus tree

(1)

Bootstrap values:

< 50% - no - just by chance ; > 75% ok; 95-100% great


Cyclura pinguis (Anegada Island)
Cyclura collei (Jamaica)
Cyclura rileyi (San Salvador Islands)
Pepino (Cuba)
77
Cyclura lewisi 1 (Grand Cayman)
72
Cyclura lewisi 2 (Grand Cayman)
99
96
Prague 5 (Cuba)
65
85
Tonik (Cuba)
100
Cyclura c.cychlura (Andros Island, Bahamas)
84
Cyclura c.figginsi 1 (Exuma Cays, Bahamas)
99
Cyclura c.figginsi 2 (Exuma Cays, Bahamas)
54
Cyclura c.inornata (Exuma Cays, Bahamas)
67
Holguin G0H (Cuba)
Cyclura n. caymanensis 1 (Little Cayman)
96
98
68
Cyclura n. nubila 3 (Cuba, Guantanamo Bay)
Cyclura n. caymanensis 2 (Little Cayman)
54
Cyclura n. nubila 1 (Cuba, Guantanamo Bay)
91
Cyclura n. nubila 2 (Cuba, Guantanamo Bay)
Prague 75 (Cuba)
Prague 76 (Cuba)
51
Marea cw1 (Cuba)
Cyclura c. carinata 1 (Turks and Caicos Islands)
99
Cyclura c. carinata 2 (Turks and Caicos Islands)
100
Cyclura ricordi 1 (Isla cabritos, Dominican Republic)
100
Cyclura ricordi 2 (Isla cabritos, Dominican Republic)
94
Cyclura c.cornuta (Isla cabritos, Dominican Republic)
Cyclura c.stejnegeri 1 (Mona Island)
100
82
Cyclura c. stejnegeri 2 (Mona Island)
Sauromalus obesus
91
Iguana iguana
Amblyrhynchus cristatus
77
Ctenosaura similis

Starostová et al. 2010, *Amphibia-Reptilia* 31:134-143

Bayesian inference/analysis

Bayesian inference of phylogeny uses a likelihood function to create a quantity called the **posterior probability** of trees using a model of evolution (substitution model), based on some prior probabilities (priors), producing the most likely phylogenetic tree for the given data

Based on theorem of Thomas Bayes (18. century) – Bayesian theorem

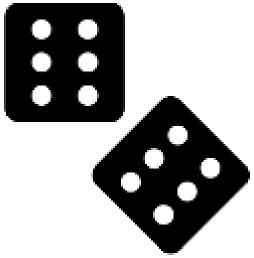
- describes the probability of an event, based on prior knowledge of conditions that might be related to the event

$$\underset{\text{Posterior probability}}{\mathrm{Pr}\ (H \mid D)} = \frac{\overset{\text{Likelihood}}{\mathrm{Pr}\ (D \mid H)} \times \overset{\text{Prior}}{\mathrm{Pr}\ (H)}}{\underset{\text{Probability of data}}{\mathrm{Pr}\ (D)}}$$

Pr (D) is not possible to calculate as this is the Σ_H Pr (D | H) x Pr (H). Too many different hypothesis.

The hypothesis H is a combination of τ, ν, Π, R, α.

We will illustrate Bayesian inference using a simple example involving dice. Consider a box with 100 dice, 90 of which are fair and 10 of which are biased. The probability of observing some number of pips after rolling a fair or biased die is given in the following table:



Observation	Fair	Biased
⚀	$\frac{1}{6}$	$\frac{1}{21}$
⚁	$\frac{1}{6}$	$\frac{2}{21}$
⚂	$\frac{1}{6}$	$\frac{3}{21}$
⚃	$\frac{1}{6}$	$\frac{4}{21}$
⚄	$\frac{1}{6}$	$\frac{5}{21}$
⚅	$\frac{1}{6}$	$\frac{6}{21}$

The probability of a high roll is larger for the biased dice than for the fair dice. Suppose that you draw a die at random from the box and roll it twice, observing a four on the first roll and a six on the second roll. What is the probability that the die is biased?

A Bayesian analysis combines ones prior beliefs about the probability of a hypothesis with the likelihood. The likelihood is the vehicle that carries the information about the hypothesis contained in the observations. In this case, the likelihood is simply the probability of observing a four and a six given that the die is biased or fair. Assuming independence of the tosses, the probability of observing a four and a six is

$$\Pr[\text{⚃}, \text{⚅} | \text{ Fair}] = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

for a fair die and

$$\Pr[\text{⚃}, \text{⚅} | \text{ Biased}] = \frac{4}{21} \times \frac{6}{21} = \frac{24}{441}$$

for a biased die. The probability of observing the data is 1.96 times greater under the hypothesis that the die is biased. In other words, the ratio of the likelihoods under the two hypotheses suggests that the die is biased.

Bayesian inferences are based upon the posterior probability of a hypothesis. The posterior probability that the die is biased can be obtained using Bayes' *(1)* formula:

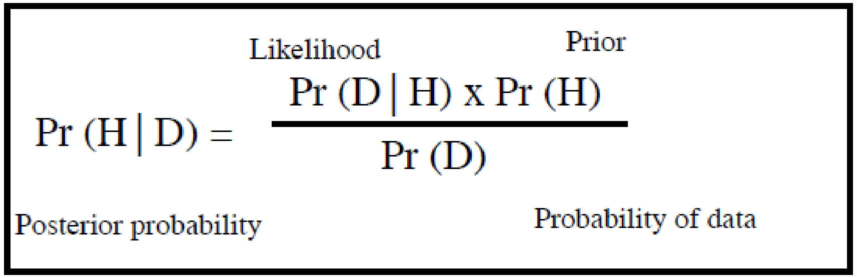
$$\Pr[\text{Biased} \mid \text{⚃}, \text{⚅}] = \frac{\Pr[\text{⚃}, \text{⚅} | \text{ Biased}] \times \Pr[\text{Biased}]}{\Pr[\text{⚃}, \text{⚅} | \text{ Biased}] \times \Pr[\text{Biased}] + \Pr[\text{⚃}, \text{⚅} | \text{ Fair}] \times \Pr[\text{Fair}]}$$

where Pr[Biased] and Pr[Fair] are the prior probabilities that the die is biased or fair, respectively. As we set up the problem, a reasonable prior probability that the die is biased would be the proportion of the dice in the box that were biased. The posterior probability is then

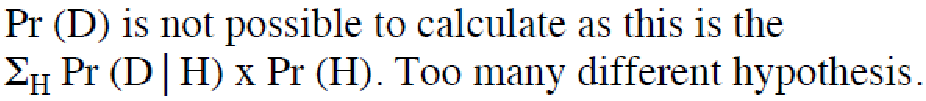
$$\Pr[\text{Biased} \mid \text{⚃}, \text{⚅}] = \frac{\frac{24}{441} \times \frac{1}{10}}{\frac{24}{441} \times \frac{1}{10} + \frac{1}{36} \times \frac{9}{10}} = 0.179$$

This means that our opinion that the die is biased changed from 0.1 to 0.179 after observing the four and six.

Bayesian inference of phylogeny uses a likelihood function to create a quantity called the **posterior probability** of trees using a model of evolution (substitution model), based on some prior probabilities (priors), producing the most likely phylogenetic tree for the given data



$$\underset{\text{Posterior probability}}{\mathrm{Pr}\,(H\,|\,D)} = \frac{\overset{\text{Likelihood}}{\mathrm{Pr}\,(D\,|\,H)} \text{ x } \overset{\text{Prior}}{\mathrm{Pr}\,(H)}}{\underset{\text{Probability of data}}{\mathrm{Pr}\,(D)}}$$



Pr (D) is not possible to calculate as this is the Σ_H Pr (D | H) x Pr (H). Too many different hypothesis.

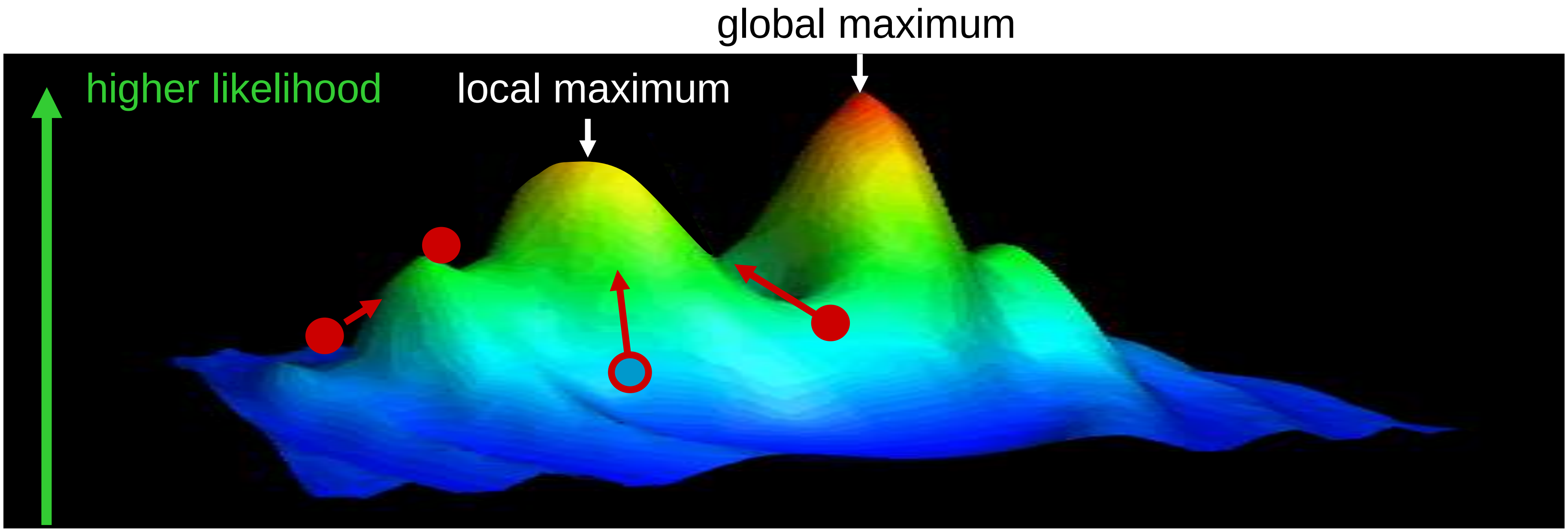
- the hypothesis H is a combination of topology of branches, branch length and parameter of the substitution model

- we may approximate the posterior distribution for H using Marcov Chain Monte Carlo (MCMC) methods

Bayesian analysis

Bayesian analysis step-by-step:

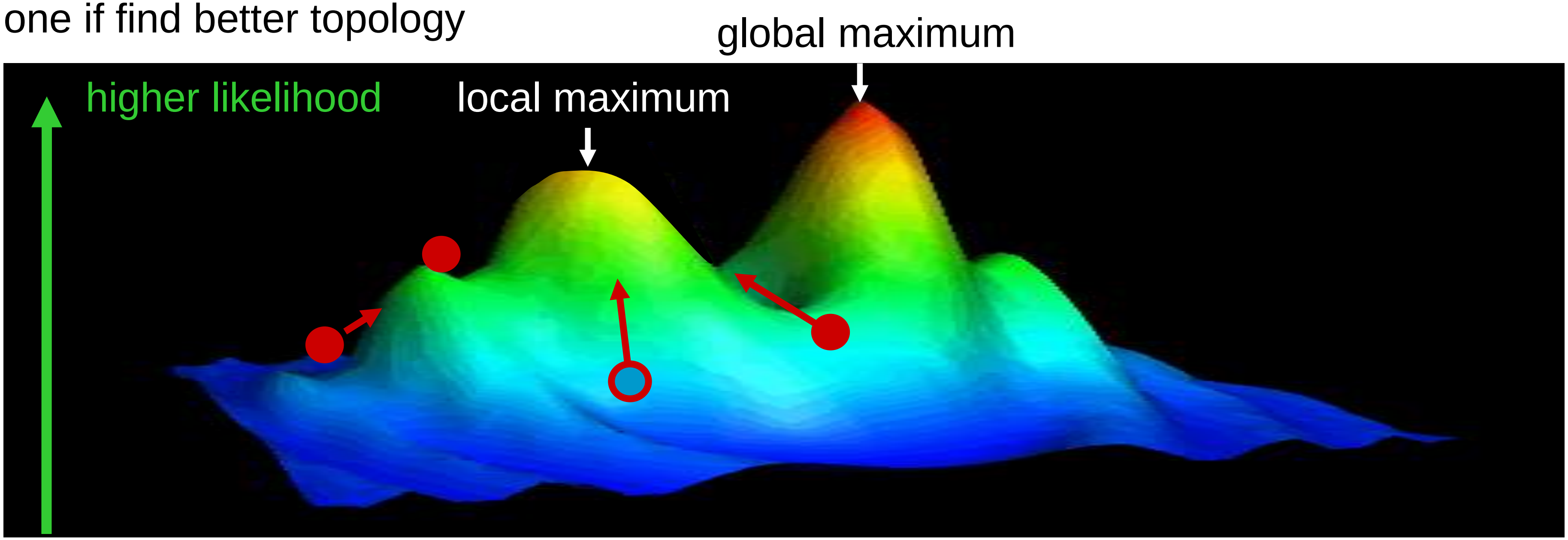
- 4 chains
- 3D space (area) with all possible trees
- find (built) first tree, compute likelihood (L)
- second tree, compute L
- if L is better, jump to the second tree, if not, stay with the first one

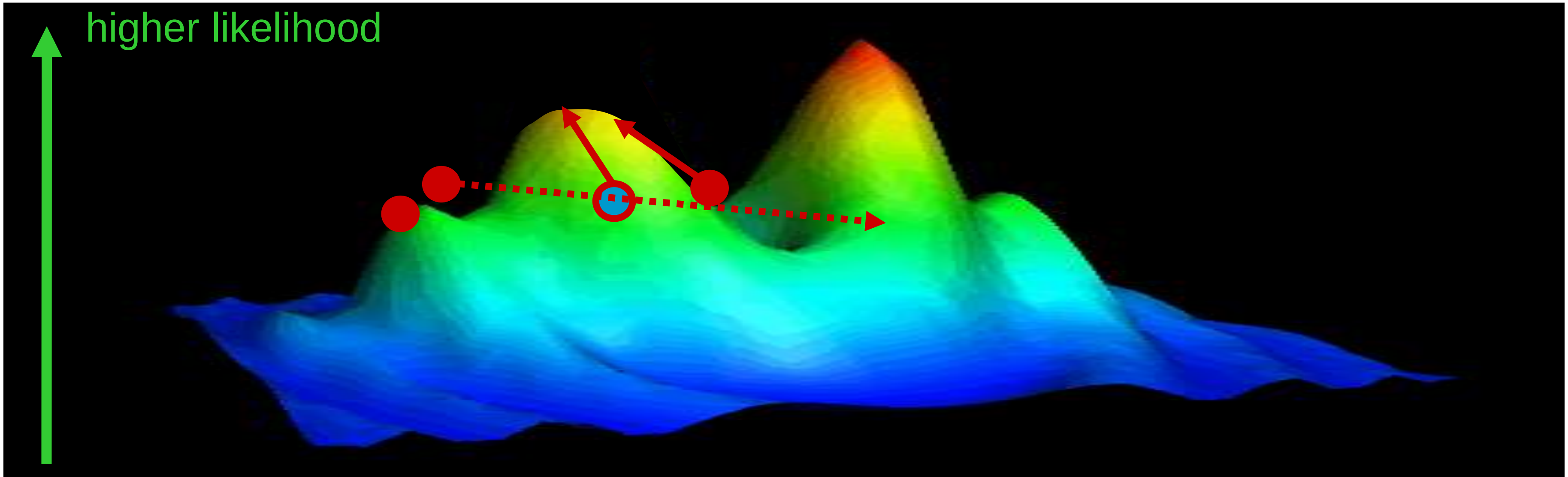


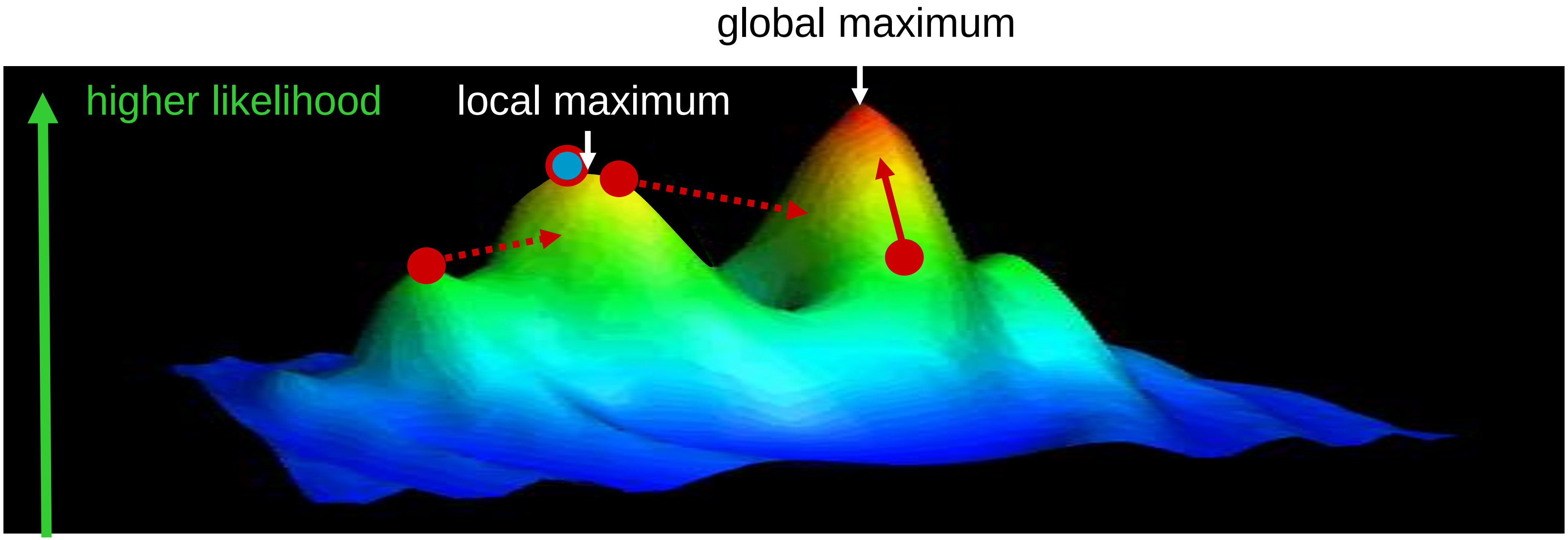
Two types of chains:

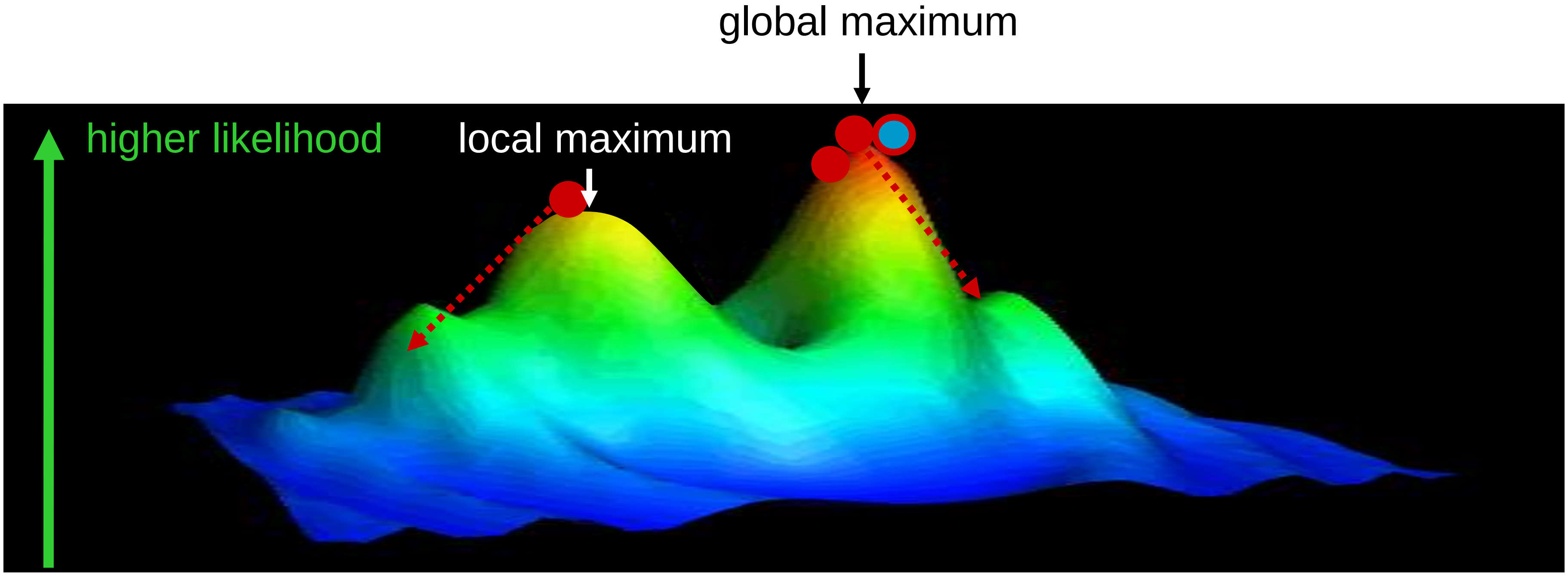
Cold – conservative one, can jump only upwards, if finds better L value

Warm – three chains – can jump also downwards + jump accidentaly + call cold one if find better topology



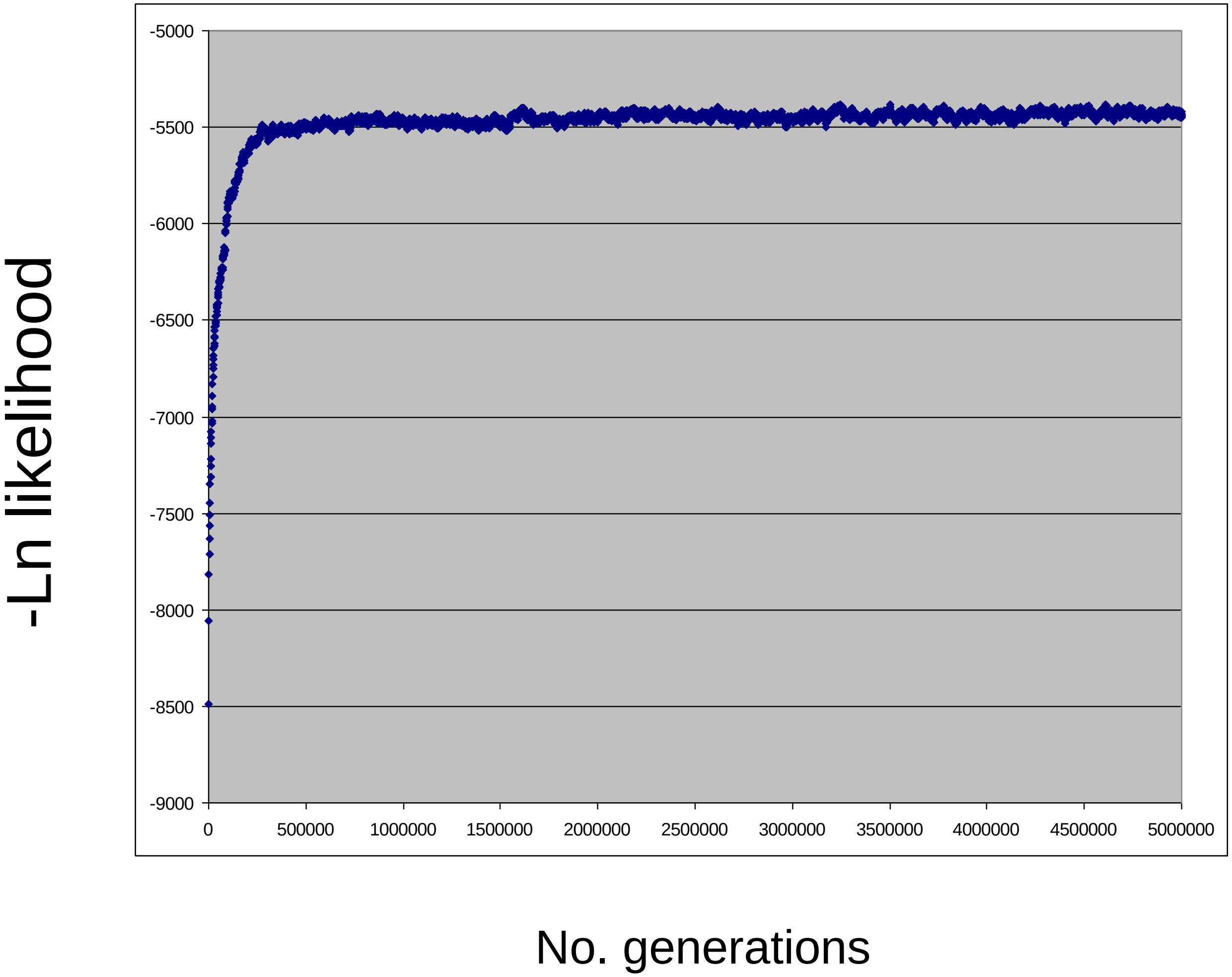



global maximum
higher likelihood
local maximum



- If there are enough generations (i.e. search steps) cold chain finds the highes global L

MrBayes run



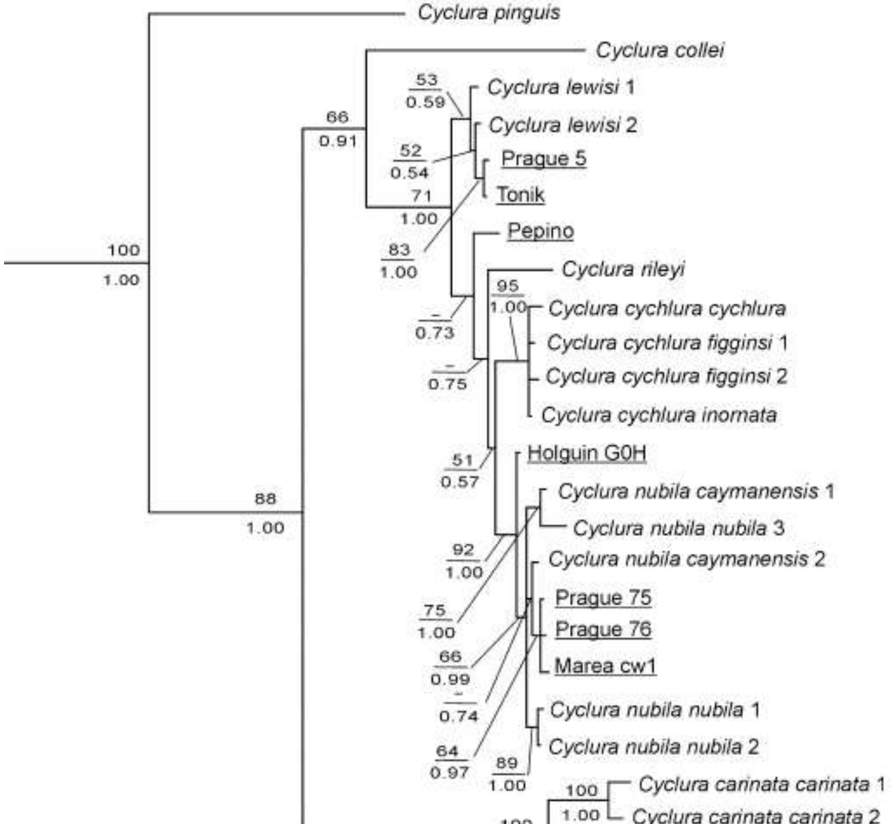
- Output of **MrBayes** is file with all trees found by cold chain during the procedure Usually every 100th tree from milions generation is saved
- Usually we have two runs

Trees at the beginning of run are not OK – we have to cut them (burnin)

Posterior probabilty

BPP (PP) is parameter of Bayesian analysis – instead of bootstraps

- BPP: represent the probability that the corresponding clade is true conditional on the model, the priors, and the data
- **below 0.95 – 0.9 topology is considered unreliable**



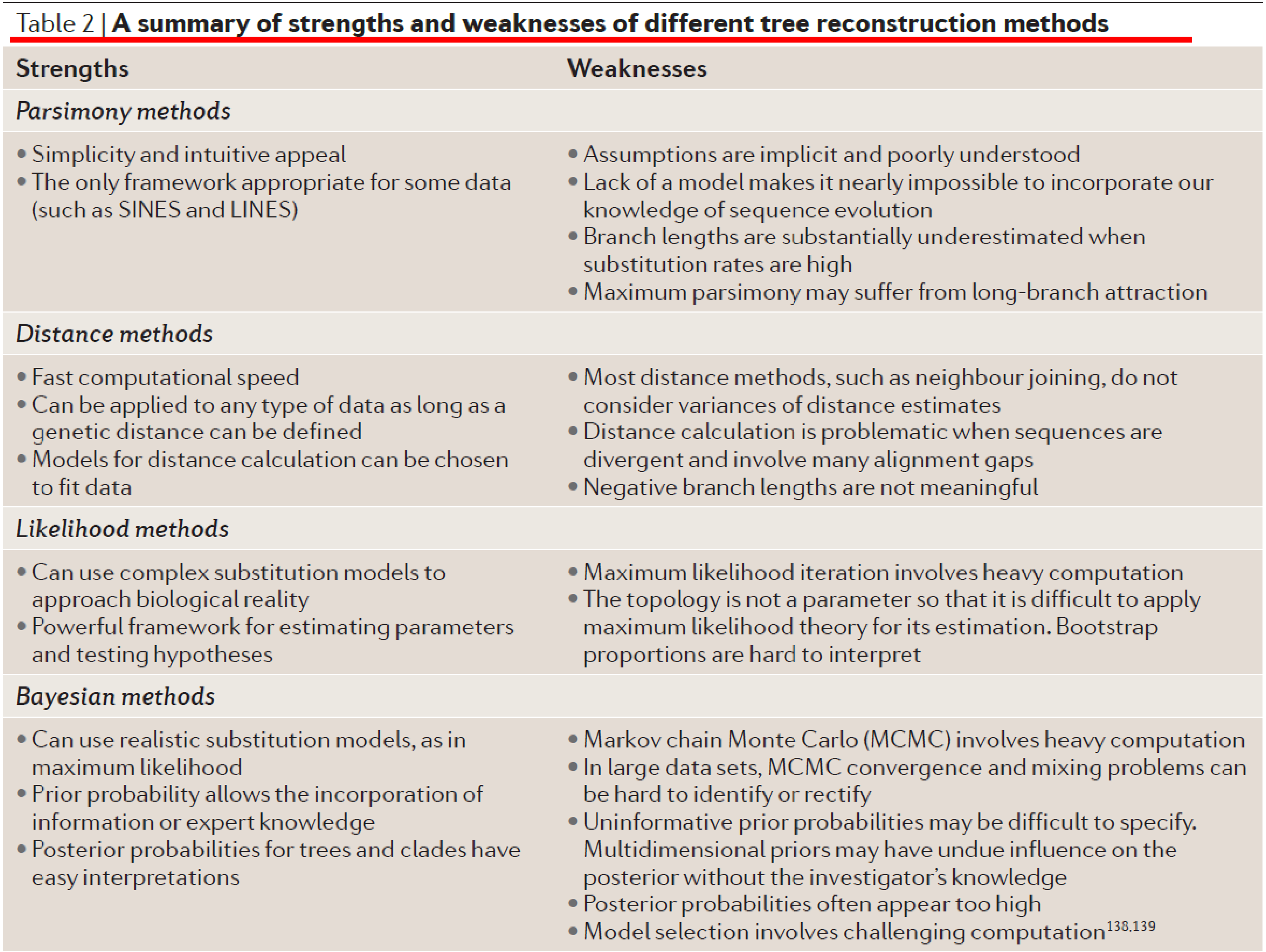
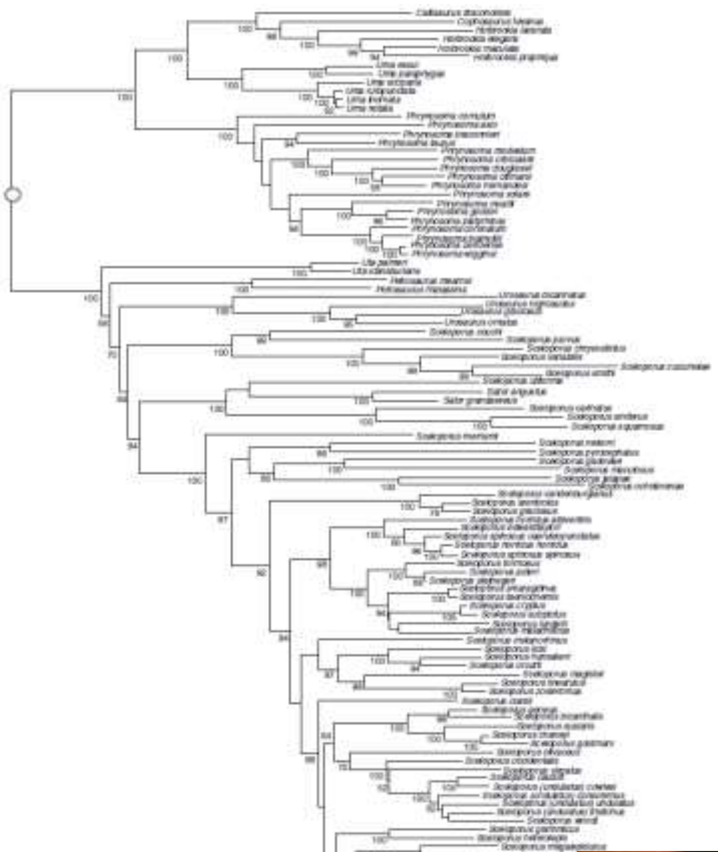


Table 2 | **A summary of strengths and weaknesses of different tree reconstruction methods**

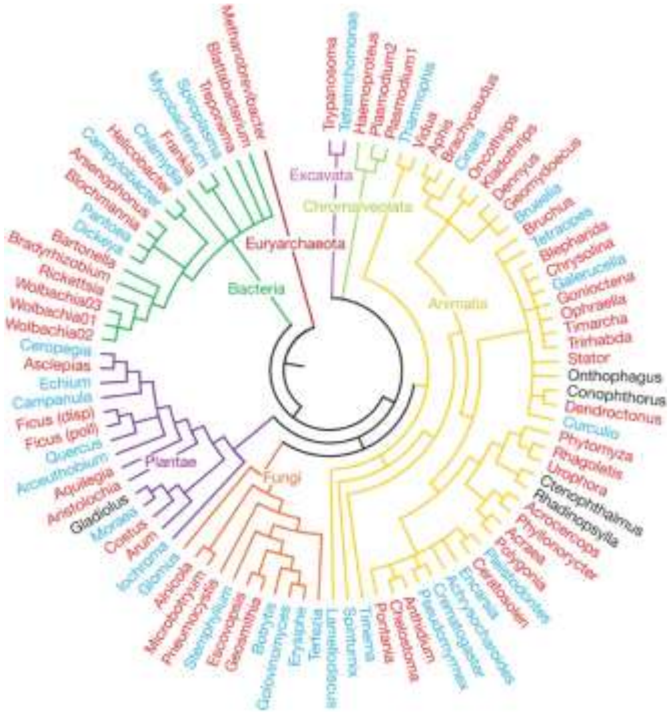
Strengths	Weaknesses
Parsimony methods	
• Simplicity and intuitive appeal • The only framework appropriate for some data (such as SINES and LINES)	• Assumptions are implicit and poorly understood • Lack of a model makes it nearly impossible to incorporate our knowledge of sequence evolution • Branch lengths are substantially underestimated when substitution rates are high • Maximum parsimony may suffer from long-branch attraction
Distance methods	
• Fast computational speed • Can be applied to any type of data as long as a genetic distance can be defined • Models for distance calculation can be chosen to fit data	• Most distance methods, such as neighbour joining, do not consider variances of distance estimates • Distance calculation is problematic when sequences are divergent and involve many alignment gaps • Negative branch lengths are not meaningful
Likelihood methods	
• Can use complex substitution models to approach biological reality • Powerful framework for estimating parameters and testing hypotheses	• Maximum likelihood iteration involves heavy computation • The topology is not a parameter so that it is difficult to apply maximum likelihood theory for its estimation. Bootstrap proportions are hard to interpret
Bayesian methods	
• Can use realistic substitution models, as in maximum likelihood • Prior probability allows the incorporation of information or expert knowledge • Posterior probabilities for trees and clades have easy interpretations	• Markov chain Monte Carlo (MCMC) involves heavy computation • In large data sets, MCMC convergence and mixing problems can be hard to identify or rectify • Uninformative prior probabilities may be difficult to specify. Multidimensional priors may have undue influence on the posterior without the investigator's knowledge • Posterior probabilities often appear too high • Model selection involves challenging computation[138,139]

Ziheng Yang & Bruce Rannala, 2012, Nature Reviews Genetics

Tree visualization:



nuclear and mitochndrial DNA maximum likelihood

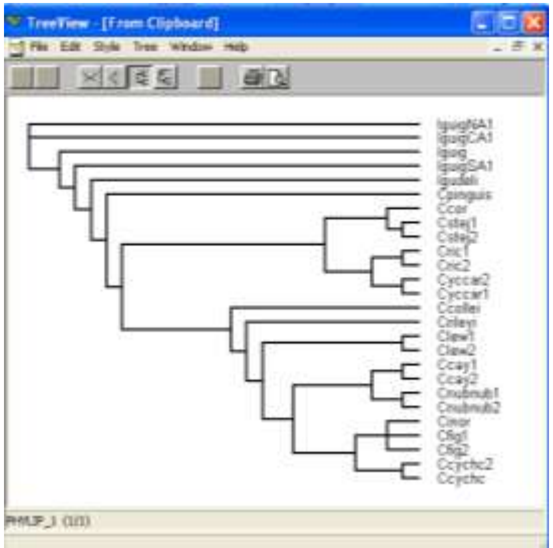


Tree visualization:

- Newick format

(A,(B,(C,(D,E))))

(IguigNA1_:0.00221,(Iguig:0.01733,(((Cpinguis:0.05228,((((Cstej1:0.00012, Cstej2:0.00098):0.00354,Ccor:0.00543):0.03863,((Cric1:0.00184, Cric2:0.00184):0.01853,(Cyccar2:0.00636,Cyccar1:0.00498):0.01702):0.03298):0.01722, (Ccollei:0.04462,(Crileyi:0.01300,((Clew1:0.00221,Clew2:0.00221):0.00885, (((Ccay1:0.00442,Ccay2:0.00111):0.00111,(Cnubnub1:0.00111,Cnubnub2:0.00000):0.00332):0.00885,((Cinor:0.00000,Cfig1:0.00111,Cfig2:0.00332):0.00221,(Ccychc2:0.00000, Ccychc:0.00000):0.00221):0.01051):0.00442):0.01023):0.02222):0.01447):0.02213):0.06215,Igudeli:0.05379):0.02871,IguigSA1_:0.02231):0.01069):0.02471,IguigCA1_:0.00553);



Different programs for tree visualization: TreeView, FigTree, Dendroscope

Take -Home Message!

- there are more methods how to calculate tree
- a phylogenetic tree is a hypothesis
- we have to test the reliability
- obtaining a good alignment is one of the most crucial steps towards a good phylogenetic tree

Software: MP: PAUP*, TNT, Phylip, MEGA, …
ML: PAUP*, PHYML, GARLI, RAxML, Phylip, MEGA,…
BA: MrBayes
NJ: PAUP*, Phylip, MEGA, …

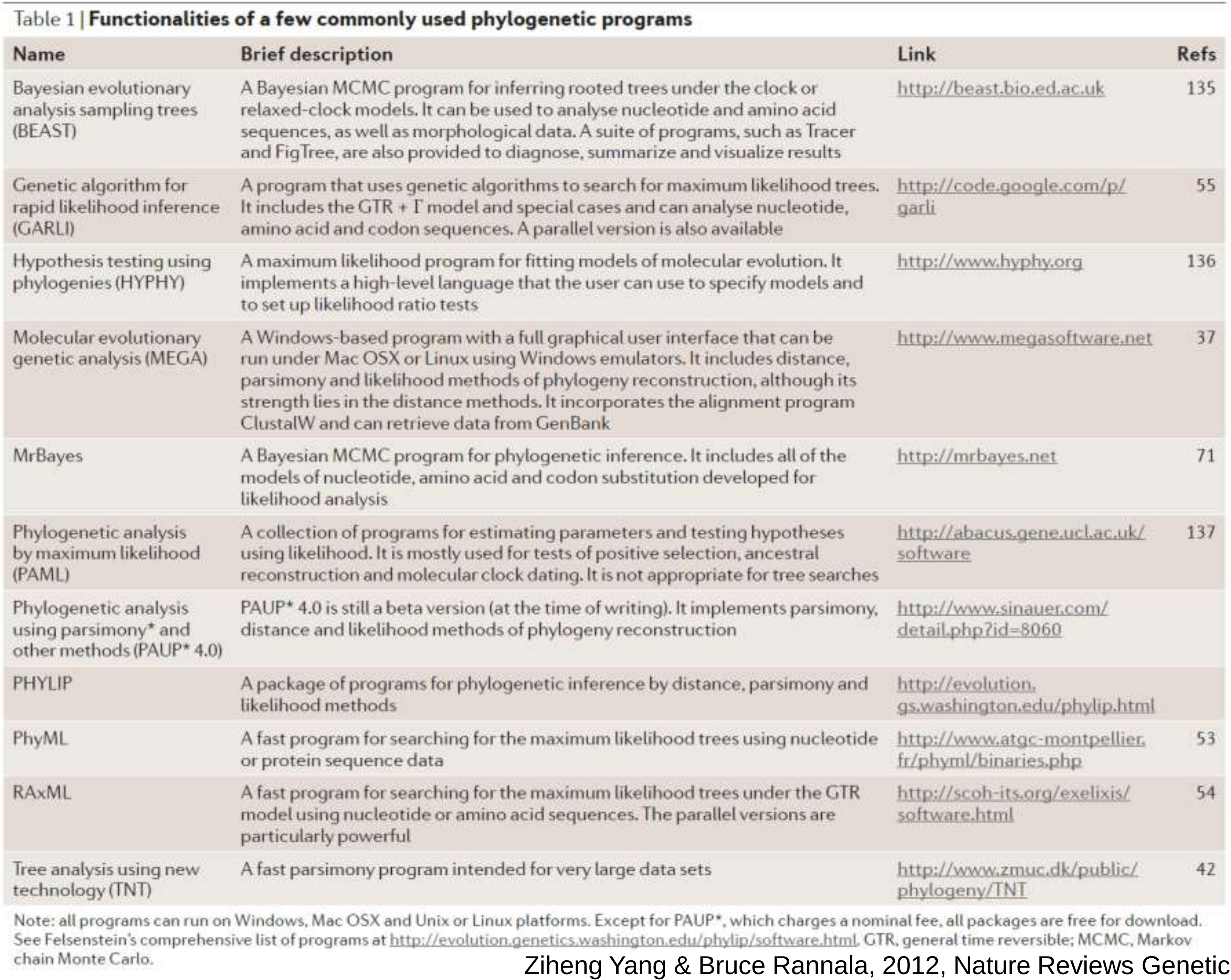


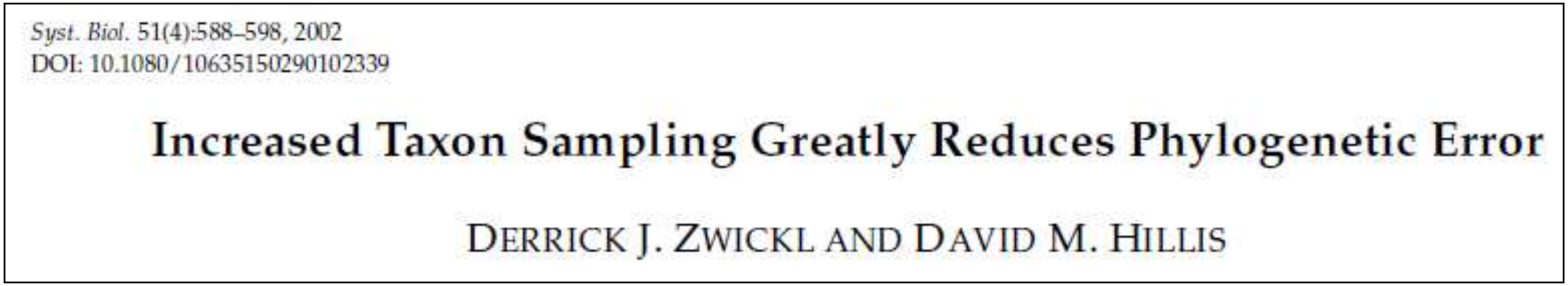
Table 1 | **Functionalities of a few commonly used phylogenetic programs**

Name	Brief description	Link	Refs
Bayesian evolutionary analysis sampling trees (BEAST)	A Bayesian MCMC program for inferring rooted trees under the clock or relaxed-clock models. It can be used to analyse nucleotide and amino acid sequences, as well as morphological data. A suite of programs, such as Tracer and FigTree, are also provided to diagnose, summarize and visualize results	http://beast.bio.ed.ac.uk	135
Genetic algorithm for rapid likelihood inference (GARLI)	A program that uses genetic algorithms to search for maximum likelihood trees. It includes the GTR + Γ model and special cases and can analyse nucleotide, amino acid and codon sequences. A parallel version is also available	http://code.google.com/p/garli	55
Hypothesis testing using phylogenies (HYPHY)	A maximum likelihood program for fitting models of molecular evolution. It implements a high-level language that the user can use to specify models and to set up likelihood ratio tests	http://www.hyphy.org	136
Molecular evolutionary genetic analysis (MEGA)	A Windows-based program with a full graphical user interface that can be run under Mac OSX or Linux using Windows emulators. It includes distance, parsimony and likelihood methods of phylogeny reconstruction, although its strength lies in the distance methods. It incorporates the alignment program ClustalW and can retrieve data from GenBank	http://www.megasoftware.net	37
MrBayes	A Bayesian MCMC program for phylogenetic inference. It includes all of the models of nucleotide, amino acid and codon substitution developed for likelihood analysis	http://mrbayes.net	71
Phylogenetic analysis by maximum likelihood (PAML)	A collection of programs for estimating parameters and testing hypotheses using likelihood. It is mostly used for tests of positive selection, ancestral reconstruction and molecular clock dating. It is not appropriate for tree searches	http://abacus.gene.ucl.ac.uk/software	137
Phylogenetic analysis using parsimony* and other methods (PAUP* 4.0)	PAUP* 4.0 is still a beta version (at the time of writing). It implements parsimony, distance and likelihood methods of phylogeny reconstruction	http://www.sinauer.com/detail.php?id=8060	
PHYLIP	A package of programs for phylogenetic inference by distance, parsimony and likelihood methods	http://evolution.gs.washington.edu/phylip.html	
PhyML	A fast program for searching for the maximum likelihood trees using nucleotide or protein sequence data	http://www.atgc-montpellier.fr/phyml/binaries.php	53
RAxML	A fast program for searching for the maximum likelihood trees under the GTR model using nucleotide or amino acid sequences. The parallel versions are particularly powerful	http://scoh-its.org/exelixis/software.html	54
Tree analysis using new technology (TNT)	A fast parsimony program intended for very large data sets	http://www.zmuc.dk/public/phylogeny/TNT	42

Note: all programs can run on Windows, Mac OSX and Unix or Linux platforms. Except for PAUP*, which charges a nominal fee, all packages are free for download. See Felsenstein's comprehensive list of programs at http://evolution.genetics.washington.edu/phylip/software.html. GTR, general time reversible; MCMC, Markov chain Monte Carlo.

Ziheng Yang & Bruce Rannala, 2012, Nature Reviews Genetics

But what to do if tree does not look „good“?

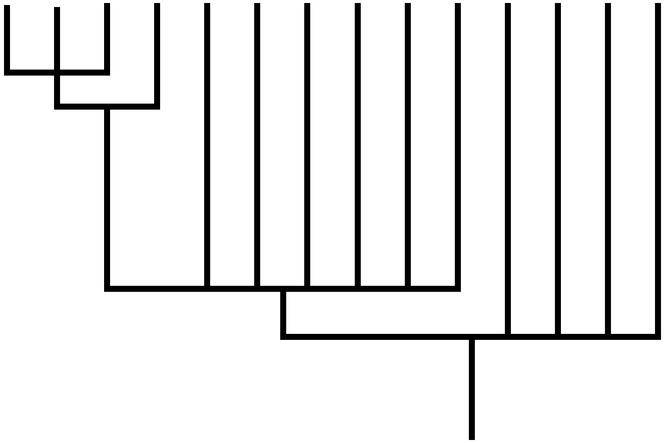


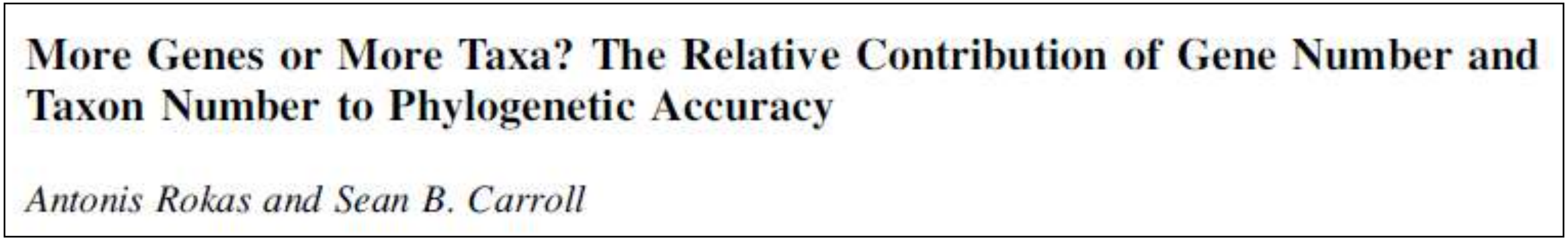
Syst. Biol. 51(4):588–598, 2002
DOI: 10.1080/10635150290102339

Increased Taxon Sampling Greatly Reduces Phylogenetic Error

DERRICK J. ZWICKL AND DAVID M. HILLIS

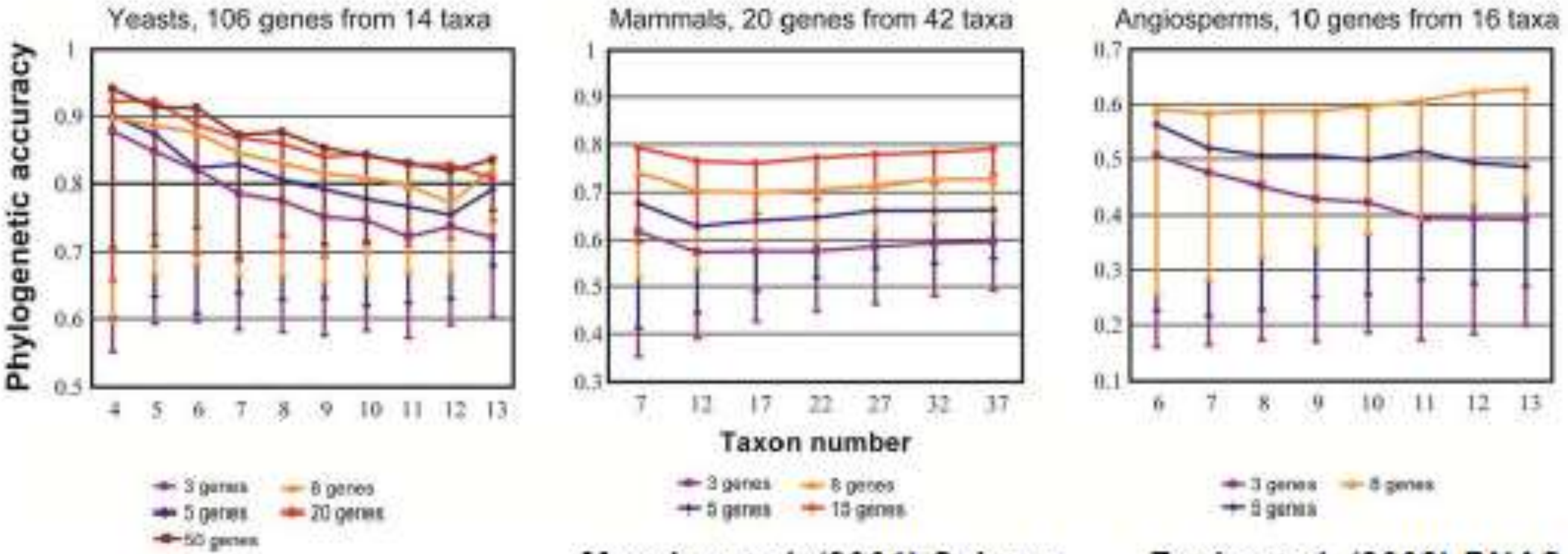
- add some more genes/sequences
- add some taxa





More Genes or More Taxa? The Relative Contribution of Gene Number and Taxon Number to Phylogenetic Accuracy

Antonis Rokas and Sean B. Carroll



Murphy et al. (2001) Science

Zanis et al. (2002) PNAS

Rokas & Carroll (2005) Mol. Biol. Evol.

- majority of taxa should have the most complete dataset!

But what to do if the tree does not look „good“?

- add some taxa
- add some more genes/sequences
- change alignment parameters – the most important
- different model of sequence evolution for each gene
- model of sequence evolution can be different for all position in coding genes

→ partitioning analysis

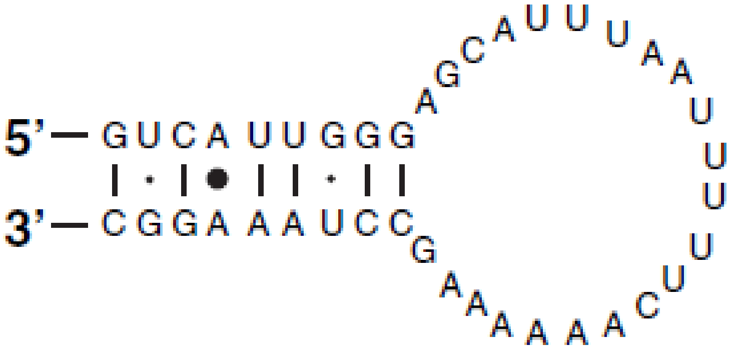
- **do not overpartition (Partition Finder v1.1.0 – Lanfear et al. 2012)**

But what to do if the tree does not look „good“?

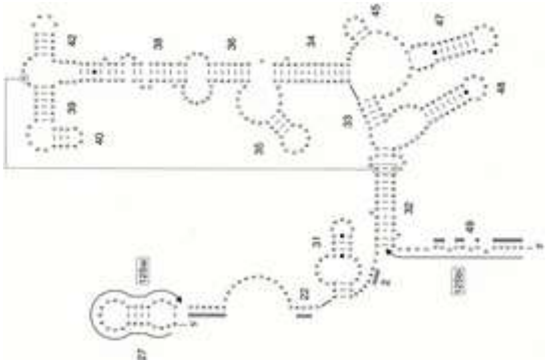
- add some taxa
- add some more genes/sequences
- change alignment parameters – the most important (contamination, „strange“ taxa)
- different model of sequence evolution for each gene
- model of sequence evolution can be different for all position in coding genes

- knowledge of secondary structure

- try more than one phylogenetic method (usually BA and ML (MP))



secondary structure of the part of DNA (RNA) of 28S rDNA

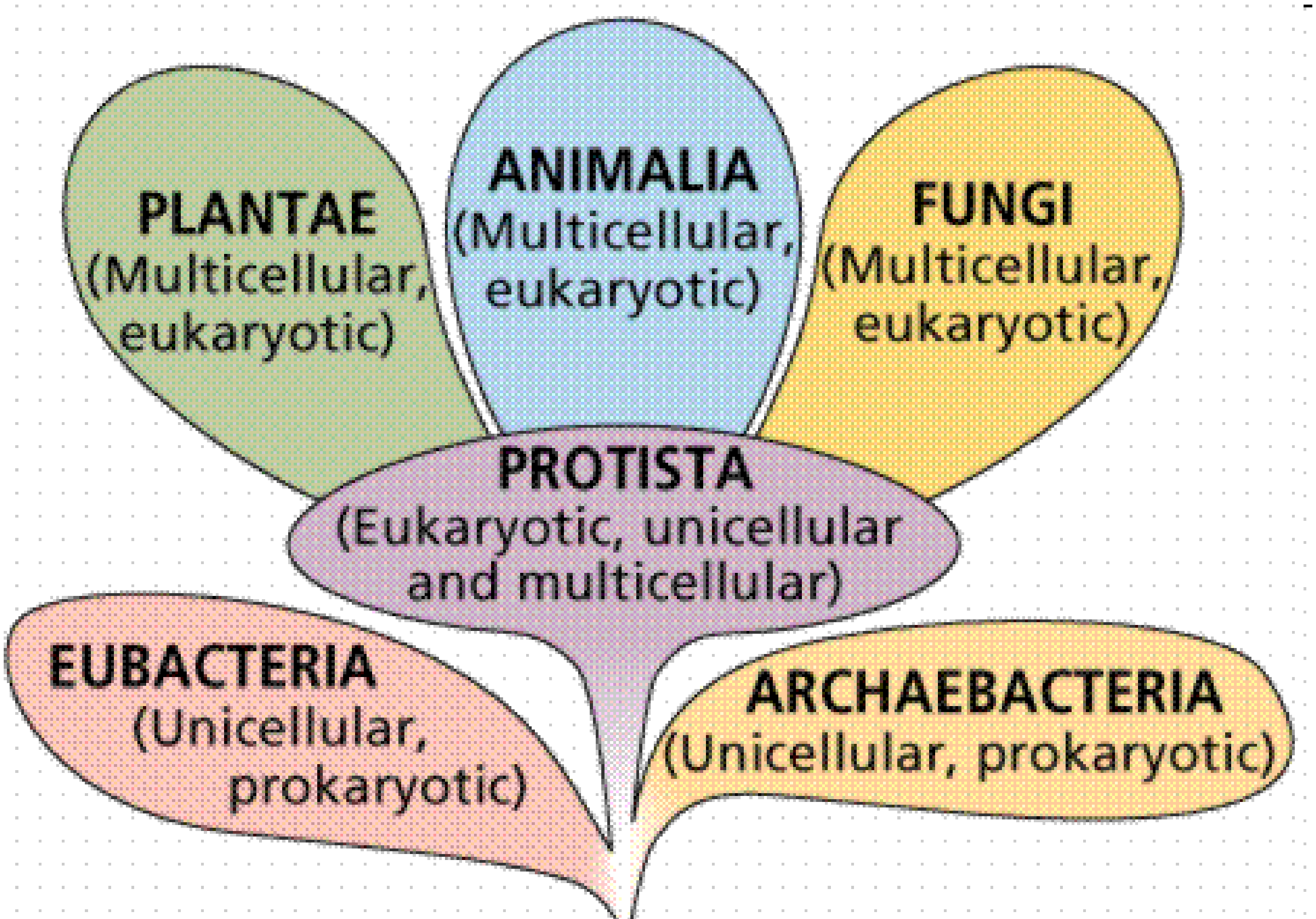


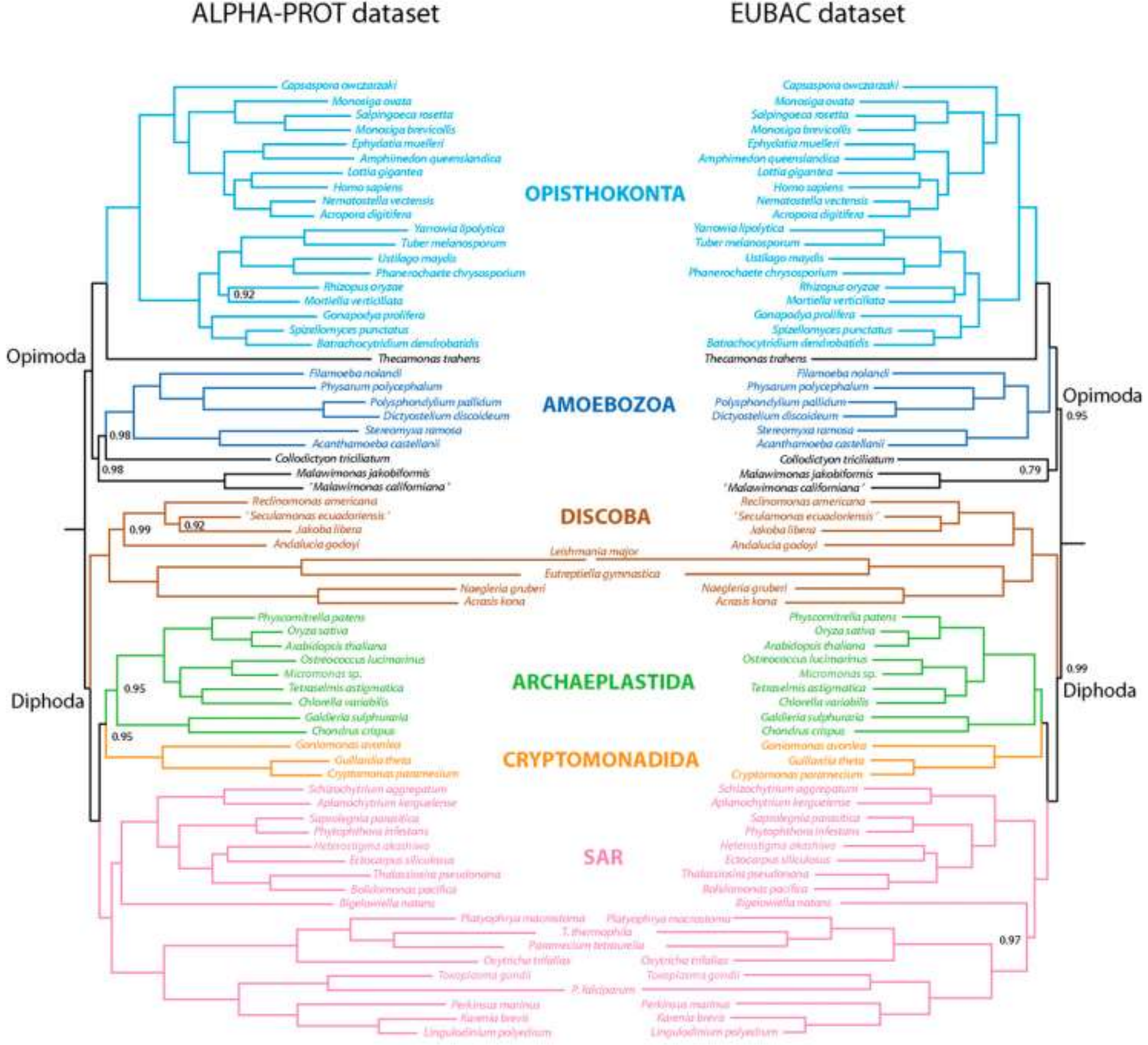
Why use of molecular phylogetics in zoology ?

- phylogeny of different groups of taxa
- definition of species boundary – use in taxonomy, cryptic species detection, character mapping and comparison
- studying biodiversity
- biogeography
- conservation biology
- disease prediction (Ebola, honey-bee pathogens, resistance etc.)

…

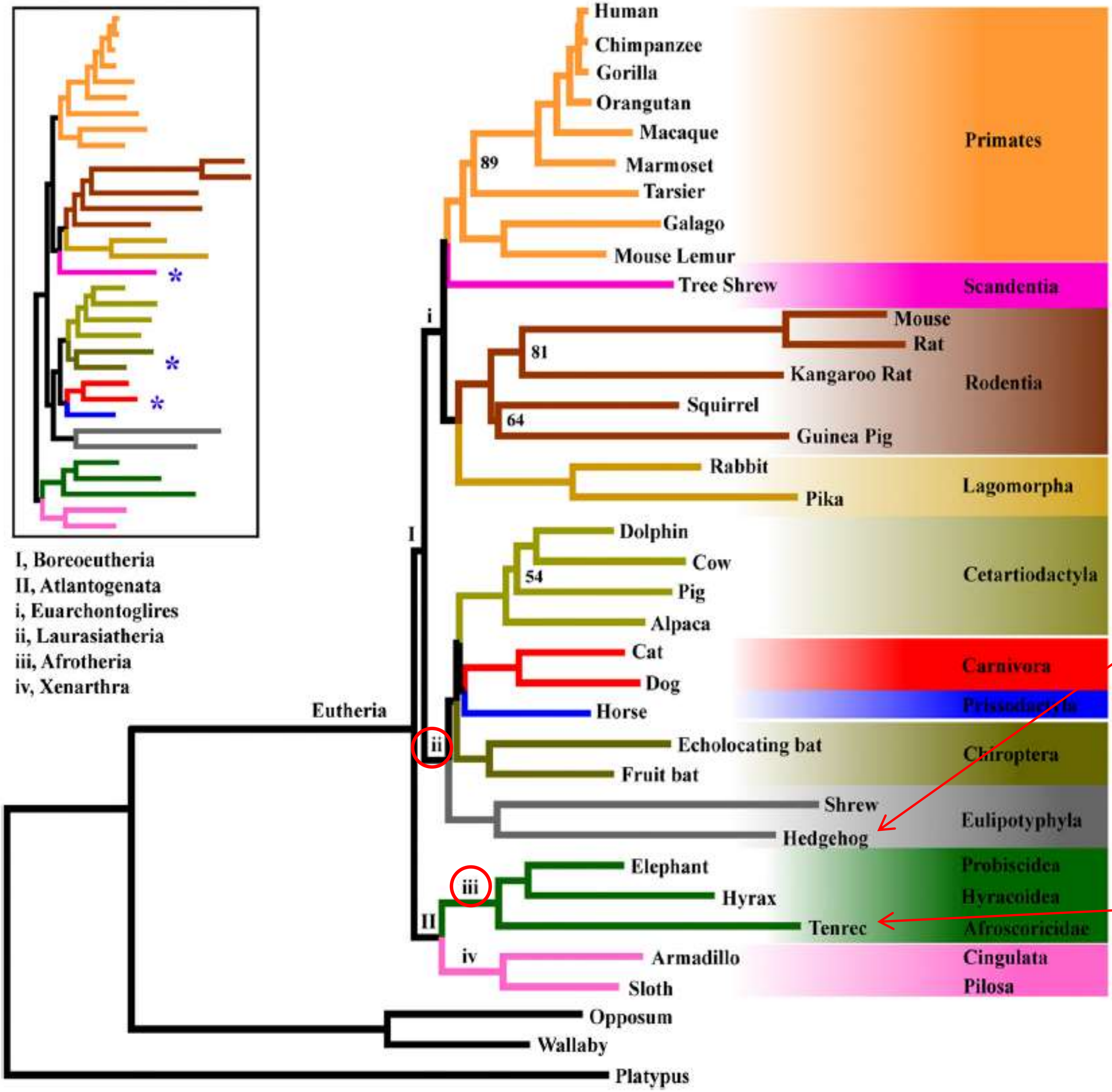
Phylogeny of Eucaryota (and Procaryota)


PLANTAE
(Multicellular, eukaryotic)
ANIMALIA
(Multicellular, eukaryotic)
FUNGI
(Multicellular, eukaryotic)
PROTISTA
(Eukaryotic, unicellular and multicellular)
EUBACTERIA
(Unicellular, prokaryotic)
ARCHAEBACTERIA
(Unicellular, prokaryotic)



Derelle et al. PNAS 2015;112:E693-E699

Phylogeny of mammals

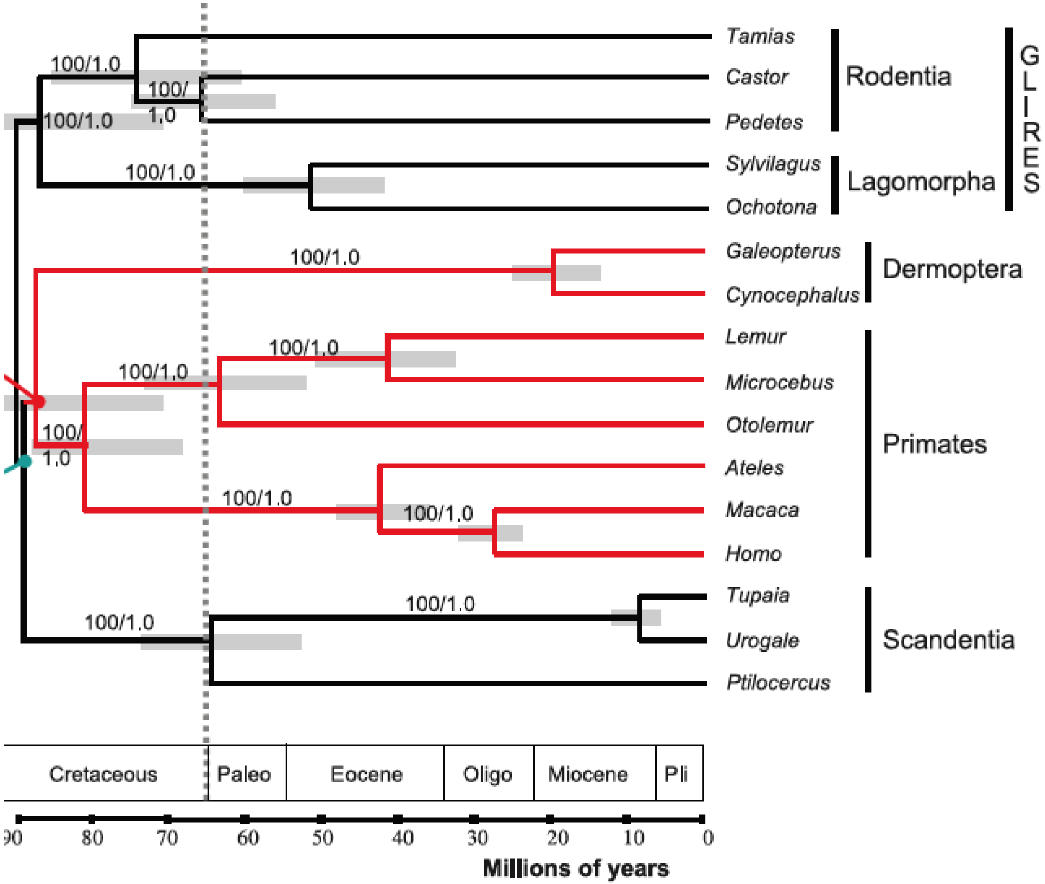


- 447 orthologous genes

Song et al. 2012 – PNAS

slide prepared by Petr Janšta

Closest relatives to primates

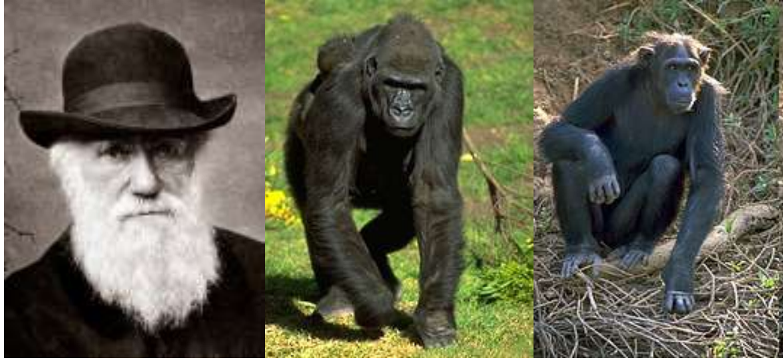


Janečka *et al.*, Science 2007

Colugos (Dermoptera) – sister group of primates,

- arboreal gliding mammals that are native to Southeast Asia
- also called flying lemurs

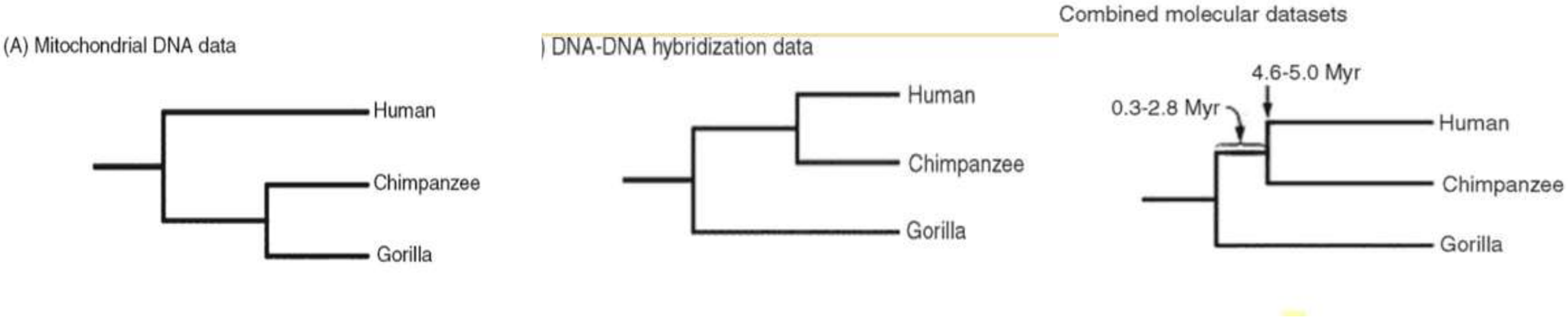
Phylogenetic analyses based on DNA data clarified the evolutionary relationships between humans and other primates



- Darwin was the first to speculate on evolutionary relationships between humans and other primates

- in 1960 from fossils paleontologist concluded that chimps and gorillas are our closest relatives and that the split occurred 15 MYA

- different molecular data put this split as much more recent - around 5 MYA



(A) Mitochondrial DNA data

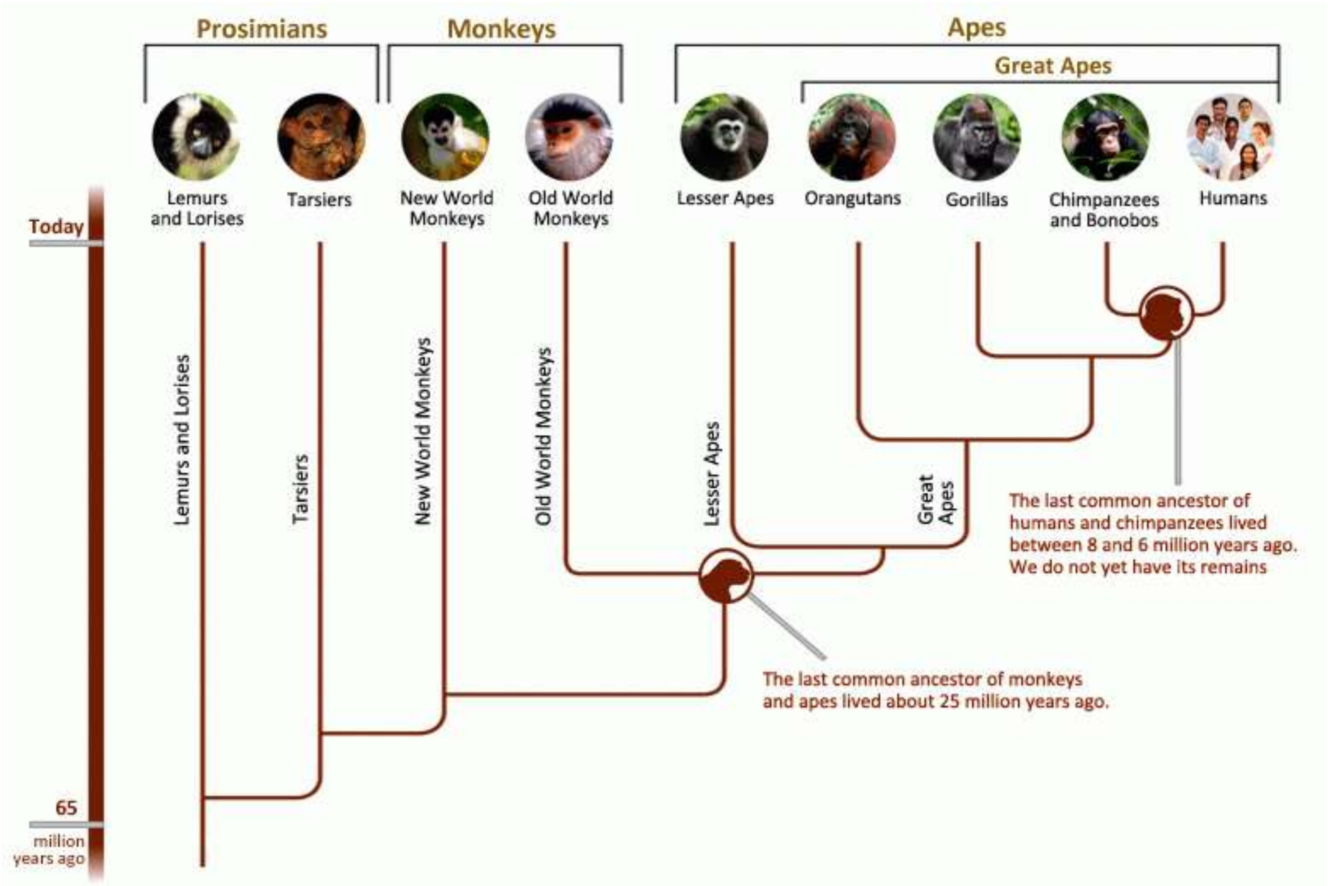
) DNA-DNA hybridization data

Combined molecular datasets

Humans, chimpanzees, and bonobos are more closely related to one another than either is to gorillas or any other primate.

Comparison of genomes: humans and chimpanzees shared a common ancestor ~5-7 MYA. The difference between the two genomes is ~4%—comprising ~35 million single nucleotide differences and ~90 Mb of insertions and deletions.

The 1.2% chimp-human distinction involves only substitutions in genes that chimpanzees and humans share.



The last common ancestor of humans and chimpanzees lived between 8 and 6 million years ago. We do not yet have its remains

The last common ancestor of monkeys and apes lived about 25 million years ago.

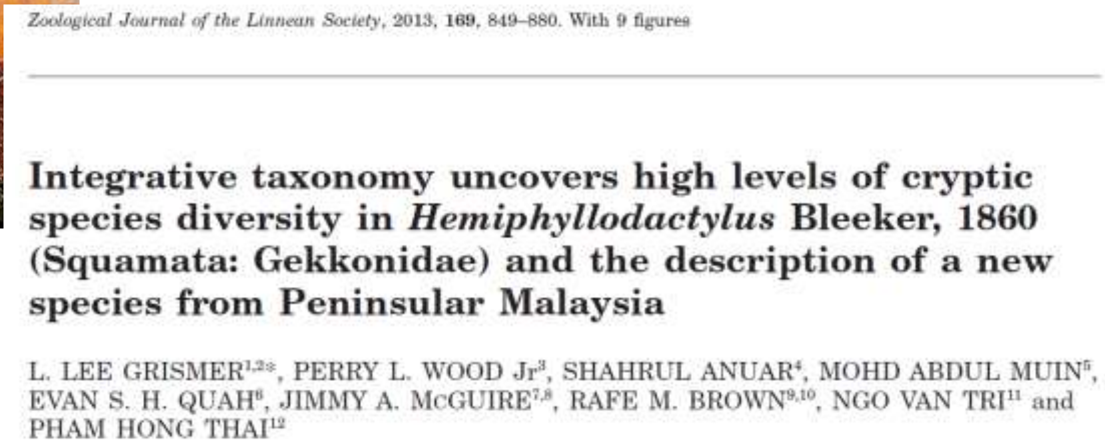
Integrative taxonomy

- only 14–75% of estimated planet´s biodiversity is described (Mora *et al*. 2011, Costello, May & Sork 2013)
- limitation of morphological x molecular taxonomy
- integrative taxonomy (at first molecules and then morphology)

slide prepared by Petr Janšta

Cryptic species diversity in *Hemiphyllodactylus* geckos

- Previously known only 8 species and some subspecies, same appearance, loss of good diagnostic characters

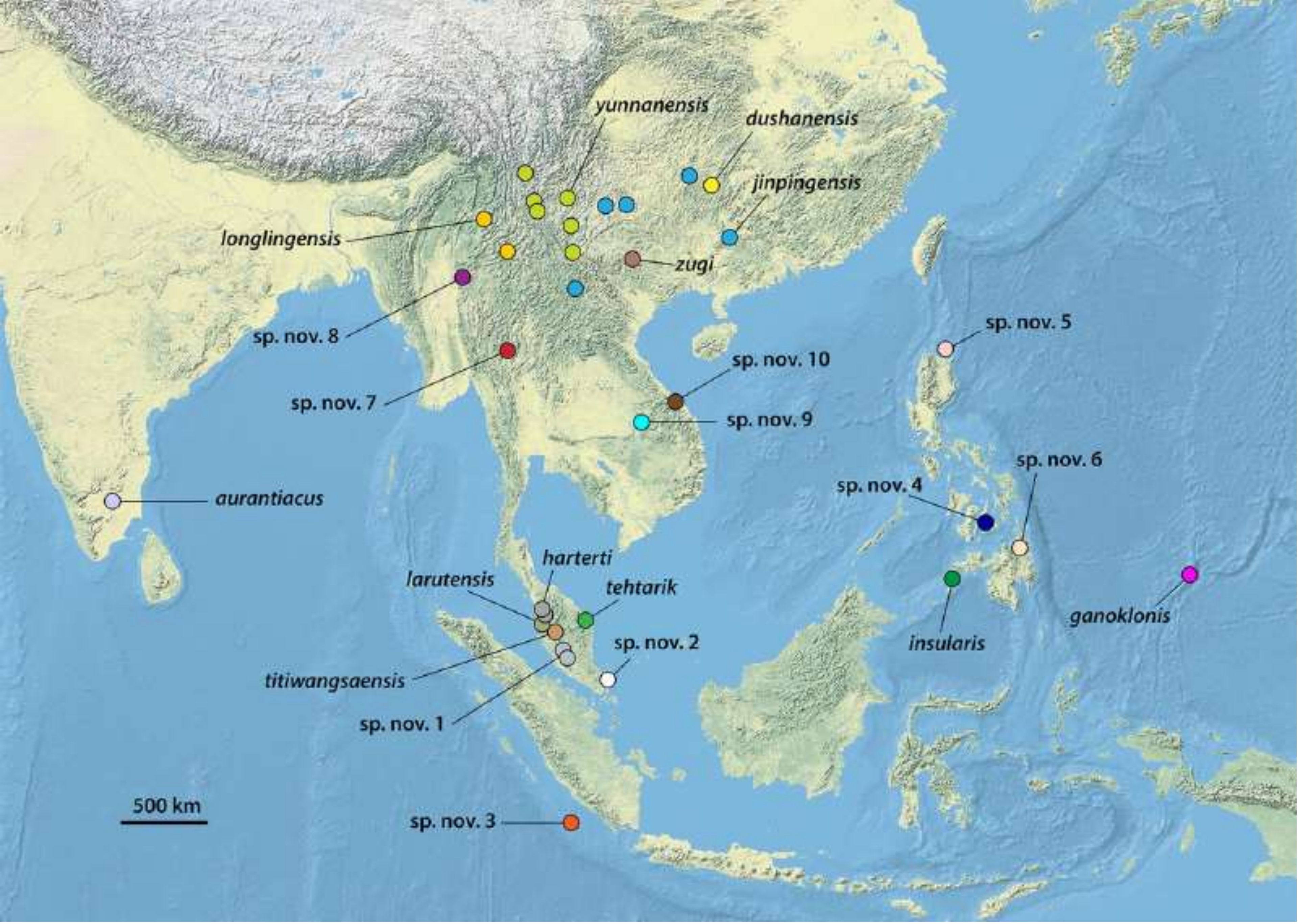


Zoological Journal of the Linnean Society, 2013, 169, 849–880. With 9 figures

Integrative taxonomy uncovers high levels of cryptic species diversity in *Hemiphyllodactylus* Bleeker, 1860 (Squamata: Gekkonidae) and the description of a new species from Peninsular Malaysia

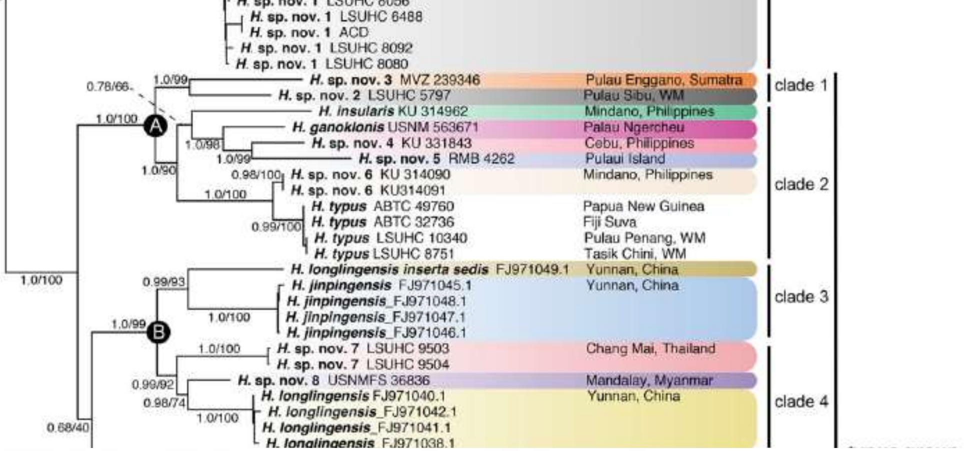
L. LEE GRISMER[1,2*], PERRY L. WOOD Jr[3], SHAHRUL ANUAR[4], MOHD ABDUL MUIN[5], EVAN S. H. QUAH[6], JIMMY A. McGUIRE[7,8], RAFE M. BROWN[9,10], NGO VAN TRI[11] and PHAM HONG THAI[12]

slide prepared by Petr Janšta



slide prepared by Petr Janšta

Divergency plot – usually 18-30% in ND gene



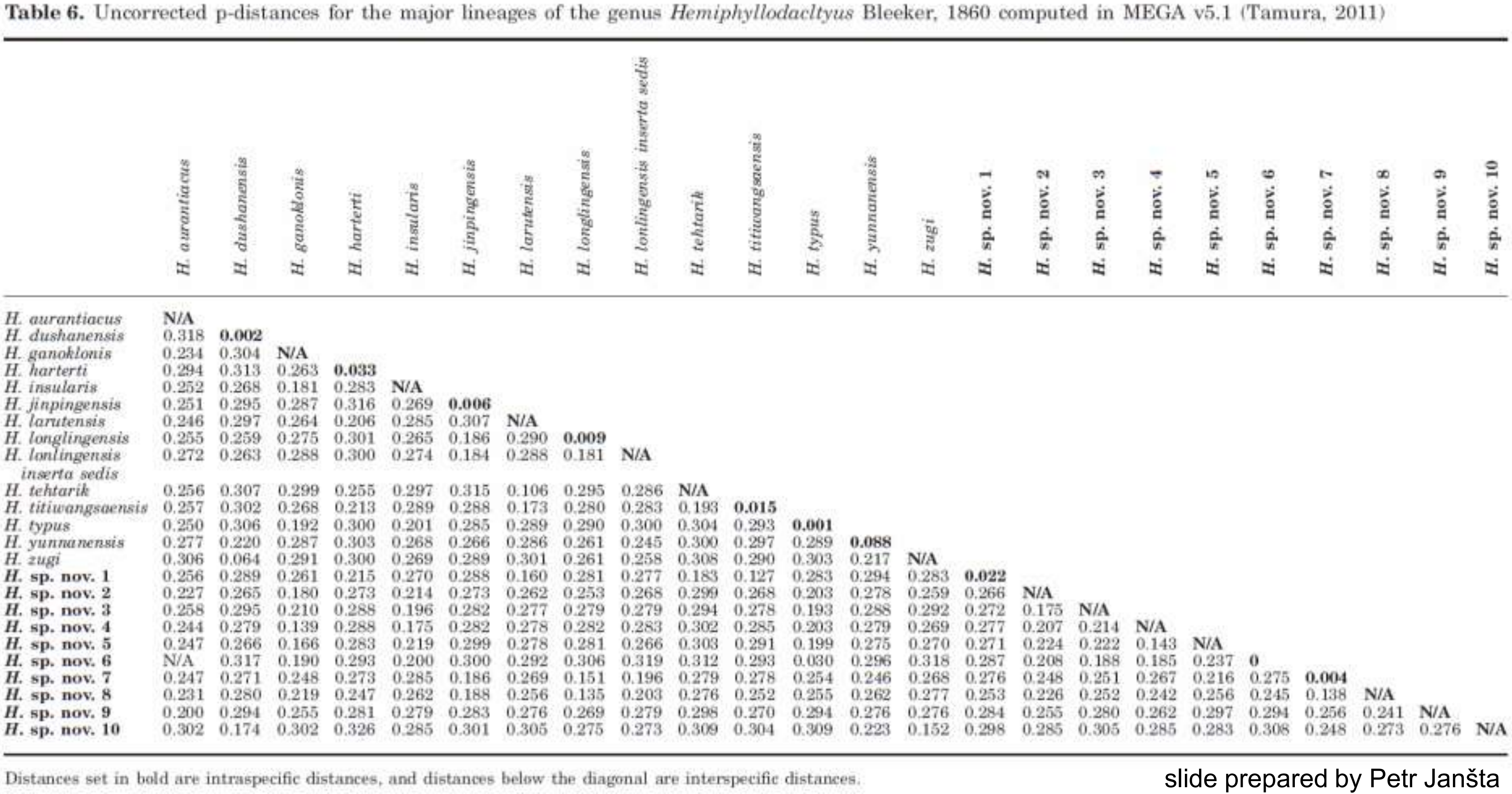


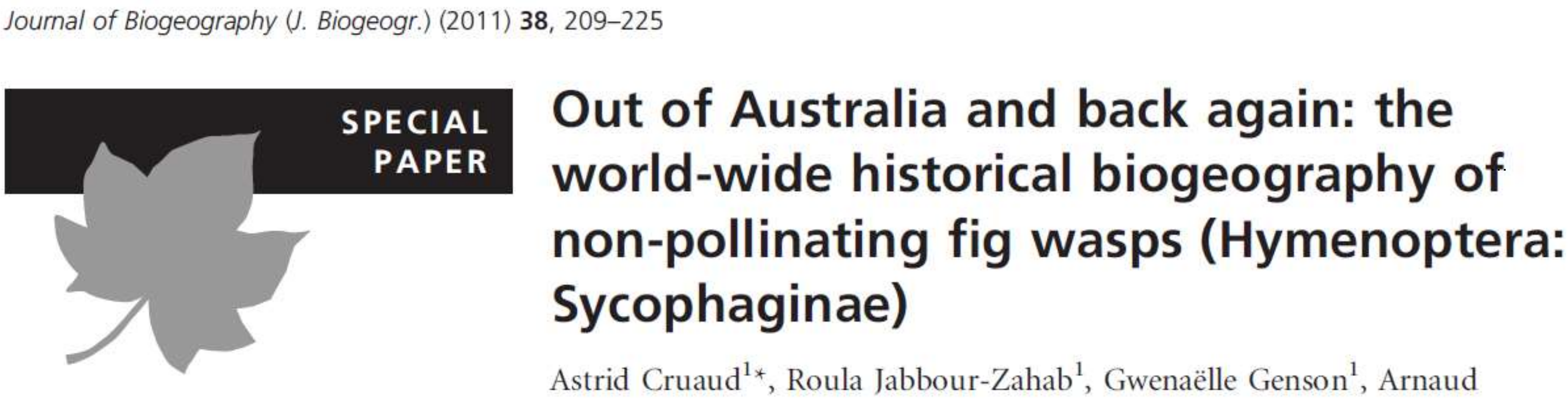
Table 6. Uncorrected p-distances for the major lineages of the genus *Hemiphyllodacltyus* Bleeker, 1860 computed in MEGA v5.1 (Tamura, 2011)

	H. aurantiacus	*H. dushanensis*	*H. ganoklonis*	*H. harterti*	*H. insularis*	*H. jinpingensis*	*H. larutensis*	*H. longlingensis*	*H. lonlingensis inserta sedis*	*H. tehtarik*	*H. titiwangsaensis*	*H. typus*	*H. yunnanensis*	*H. zugi*	***H.* sp. nov. 1**	***H.* sp. nov. 2**	***H.* sp. nov. 3**	***H.* sp. nov. 4**	***H.* sp. nov. 5**	***H.* sp. nov. 6**	***H.* sp. nov. 7**	***H.* sp. nov. 8**	***H.* sp. nov. 9**	***H.* sp. nov. 10**
H. aurantiacus	**N/A**																							
H. dushanensis	0.318	**0.002**																						
H. ganoklonis	0.234	0.304	**N/A**																					
H. harterti	0.294	0.313	0.263	**0.033**																				
H. insularis	0.252	0.268	0.181	0.283	**N/A**																			
H. jinpingensis	0.251	0.295	0.287	0.316	0.269	**0.006**																		
H. larutensis	0.246	0.297	0.264	0.206	0.285	0.307	**N/A**																	
H. longlingensis	0.255	0.259	0.275	0.301	0.265	0.186	0.290	**0.009**																
H. lonlingensis inserta sedis	0.272	0.263	0.288	0.300	0.274	0.184	0.288	0.181	**N/A**															
H. tehtarik	0.256	0.307	0.299	0.255	0.297	0.315	0.106	0.295	0.286	**N/A**														
H. titiwangsaensis	0.257	0.302	0.268	0.213	0.289	0.288	0.173	0.280	0.283	0.193	**0.015**													
H. typus	0.250	0.306	0.192	0.300	0.201	0.285	0.289	0.290	0.300	0.304	0.293	**0.001**												
H. yunnanensis	0.277	0.220	0.287	0.303	0.268	0.266	0.286	0.261	0.245	0.300	0.297	0.289	**0.088**											
H. zugi	0.306	0.064	0.291	0.300	0.269	0.289	0.301	0.261	0.258	0.308	0.290	0.303	0.217	**N/A**										
***H.* sp. nov. 1**	0.256	0.289	0.261	0.215	0.270	0.288	0.160	0.281	0.277	0.183	0.127	0.283	0.294	0.283	**0.022**									
***H.* sp. nov. 2**	0.227	0.265	0.180	0.273	0.214	0.273	0.262	0.253	0.268	0.299	0.268	0.203	0.278	0.259	0.266	**N/A**								
***H.* sp. nov. 3**	0.258	0.295	0.210	0.288	0.196	0.282	0.277	0.279	0.279	0.294	0.278	0.193	0.288	0.292	0.272	0.175	**N/A**							
***H.* sp. nov. 4**	0.244	0.279	0.139	0.288	0.175	0.282	0.278	0.282	0.283	0.302	0.285	0.203	0.279	0.269	0.277	0.207	0.214	**N/A**						
***H.* sp. nov. 5**	0.247	0.266	0.166	0.283	0.219	0.299	0.278	0.281	0.266	0.303	0.291	0.199	0.275	0.270	0.271	0.224	0.222	0.143	**N/A**					
***H.* sp. nov. 6**	N/A	0.317	0.190	0.293	0.200	0.300	0.292	0.306	0.319	0.312	0.293	0.030	0.296	0.318	0.287	0.208	0.188	0.185	0.237	**0**				
***H.* sp. nov. 7**	0.247	0.271	0.248	0.273	0.285	0.186	0.269	0.151	0.196	0.279	0.278	0.254	0.246	0.268	0.276	0.248	0.251	0.267	0.216	0.275	**0.004**			
***H.* sp. nov. 8**	0.231	0.280	0.219	0.247	0.262	0.188	0.256	0.135	0.203	0.276	0.252	0.255	0.262	0.277	0.253	0.226	0.252	0.242	0.256	0.245	0.138	**N/A**		
***H.* sp. nov. 9**	0.200	0.294	0.255	0.281	0.279	0.283	0.276	0.269	0.279	0.298	0.270	0.294	0.276	0.276	0.284	0.255	0.280	0.262	0.297	0.294	0.256	0.241	**N/A**	
***H.* sp. nov. 10**	0.302	0.174	0.302	0.326	0.285	0.301	0.305	0.275	0.273	0.309	0.304	0.309	0.223	0.152	0.298	0.285	0.305	0.285	0.283	0.308	0.248	0.273	0.276	**N/A**

Distances set in bold are intraspecific distances, and distances below the diagonal are interspecific distances.

slide prepared by Petr Janšta

Biogeography



Journal of Biogeography (J. Biogeogr.) (2011) **38**, 209–225

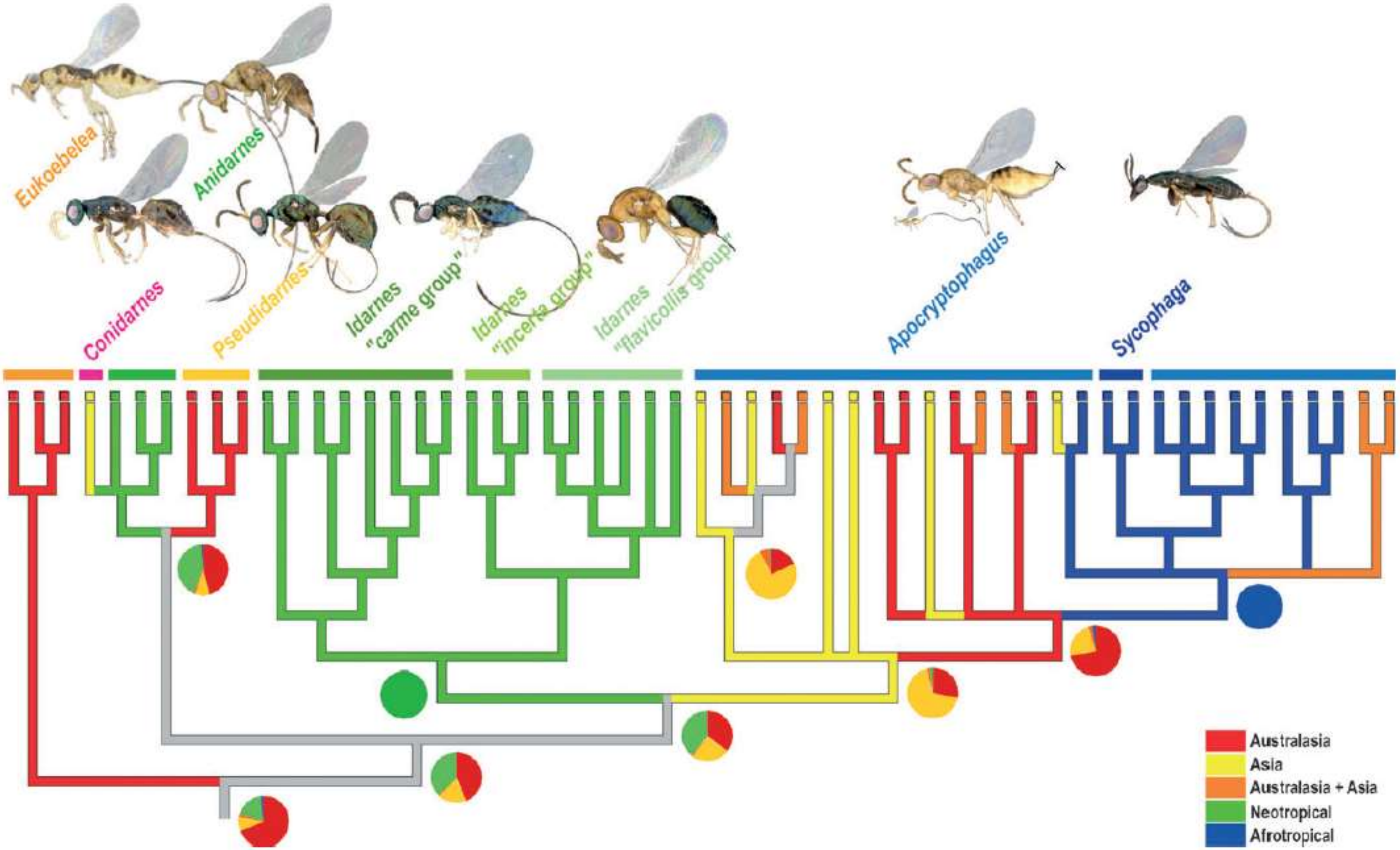
SPECIAL PAPER

Out of Australia and back again: the world-wide historical biogeography of non-pollinating fig wasps (Hymenoptera: Sycophaginae)

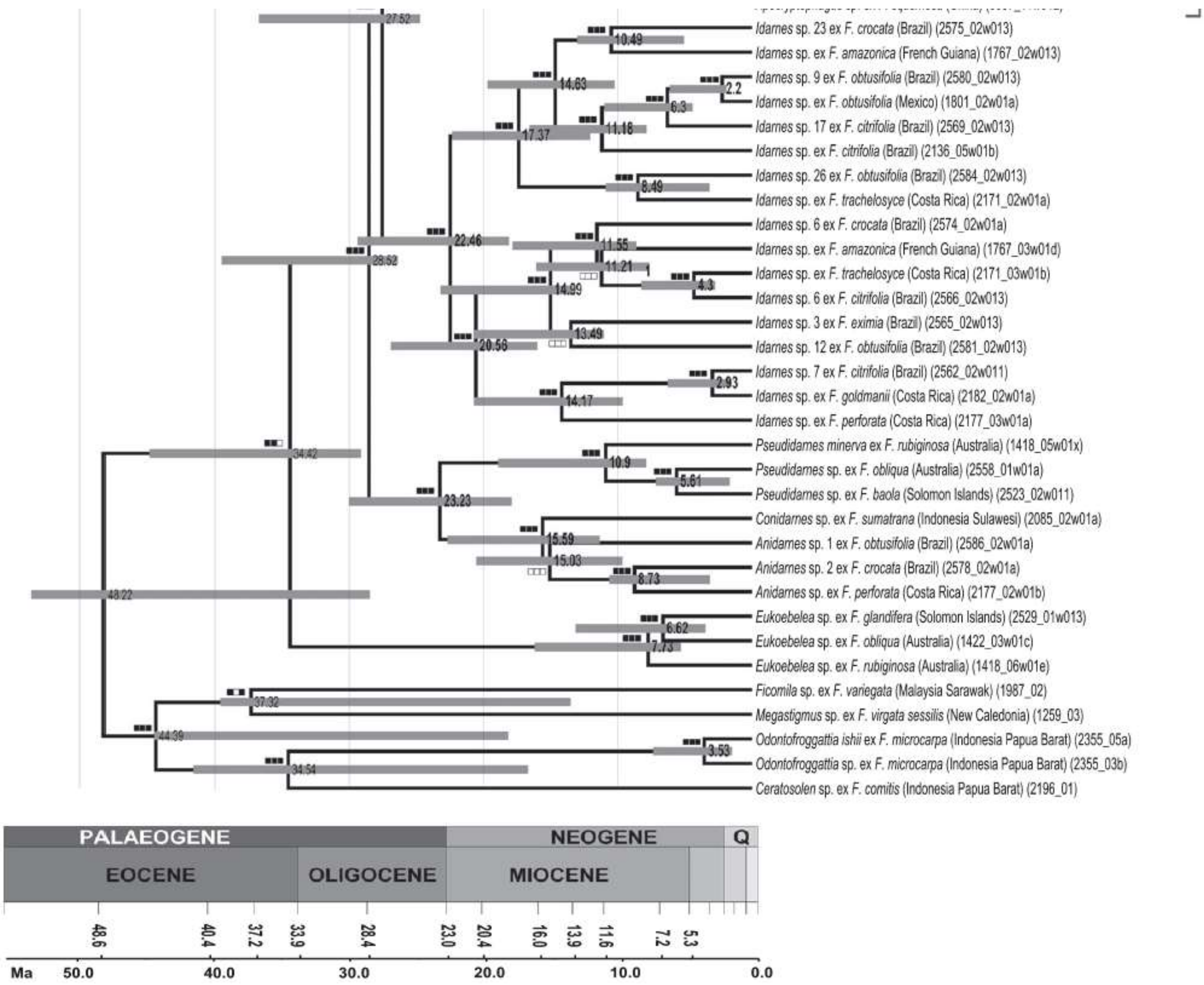
Astrid Cruaud[1]*, Roula Jabbour-Zahab[1], Gwenaëlle Genson[1], Arnaud

- Fossil dating (*Idarnes* from Dominican Amber – 30-15My; endemic taxa to Mauritius (8My) and Solomon Islands (11-12My)
- Several genes, very well resolved topology
- *Ficus* origin 100-60My, *Ficus* pollinator origin 70-15My
- Sycophaginae origin (48-35My)

slide prepared by Petr Janšta

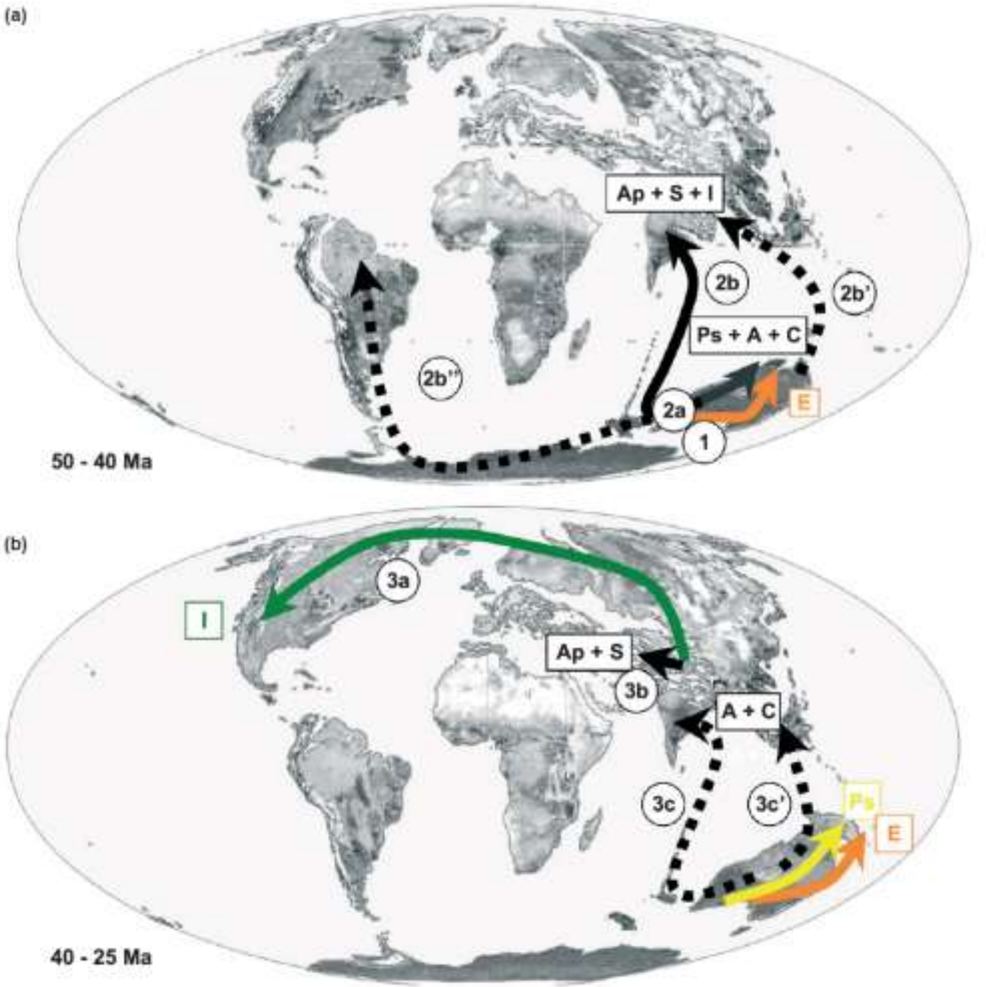


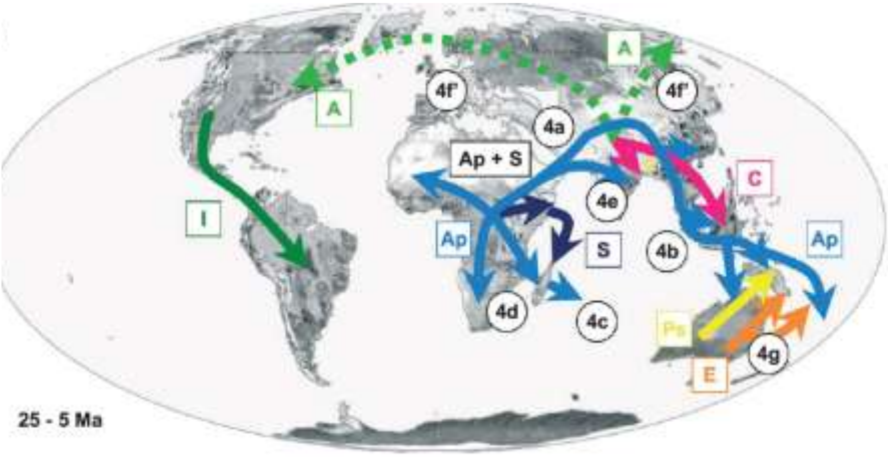
slide prepared by Petr Janšta

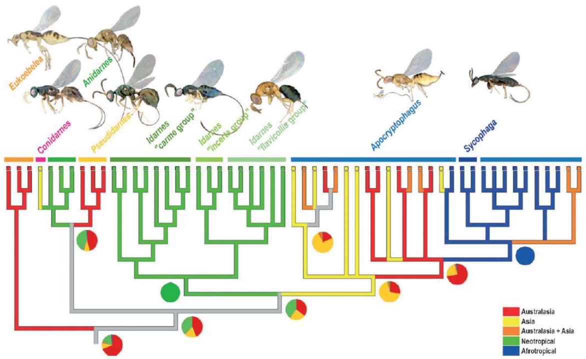

27.52
Idarnes sp. 23 ex F. crocata (Brazil) (2575_02w013)
Idarnes sp. ex F. amazonica (French Guiana) (1767_02w013)
Idarnes sp. 9 ex F. obtusifolia (Brazil) (2580_02w013)
Idarnes sp. ex F. obtusifolia (Mexico) (1801_02w01a)
Idarnes sp. 17 ex F. citrifolia (Brazil) (2569_02w013)
Idarnes sp. ex F. citrifolia (Brazil) (2136_05w01b)
Idarnes sp. 26 ex F. obtusifolia (Brazil) (2584_02w013)
Idarnes sp. ex F. trachelosyce (Costa Rica) (2171_02w01a)
Idarnes sp. 6 ex F. crocata (Brazil) (2574_02w01a)
Idarnes sp. ex F. amazonica (French Guiana) (1767_03w01d)
Idarnes sp. ex F. trachelosyce (Costa Rica) (2171_03w01b)
Idarnes sp. 6 ex F. citrifolia (Brazil) (2566_02w013)
Idarnes sp. 3 ex F. eximia (Brazil) (2565_02w013)
Idarnes sp. 12 ex F. obtusifolia (Brazil) (2581_02w013)
Idarnes sp. 7 ex F. citrifolia (Brazil) (2562_02w011)
Idarnes sp. ex F. goldmanii (Costa Rica) (2182_02w01a)
Idarnes sp. ex F. perforata (Costa Rica) (2177_03w01a)
Pseudidarnes minerva ex F. rubiginosa (Australia) (1418_05w01x)
Pseudidarnes sp. ex F. obliqua (Australia) (2558_01w01a)
Pseudidarnes sp. ex F. baola (Solomon Islands) (2523_02w011)
Conidarnes sp. ex F. sumatrana (Indonesia Sulawesi) (2085_02w01a)
Anidarnes sp. 1 ex F. obtusifolia (Brazil) (2586_02w01a)
Anidarnes sp. 2 ex F. crocata (Brazil) (2578_02w01a)
Anidarnes sp. ex F. perforata (Costa Rica) (2177_02w01b)
Eukoebelea sp. ex F. glandifera (Solomon Islands) (2529_01w013)
Eukoebelea sp. ex F. obliqua (Australia) (1422_03w01c)
Eukoebelea sp. ex F. rubiginosa (Australia) (1418_06w01e)
Ficomila sp. ex F. variegata (Malaysia Sarawak) (1987_02)
Megastigmus sp. ex F. virgata sessilis (New Caledonia) (1259_03)
Odontofroggattia ishii ex F. microcarpa (Indonesia Papua Barat) (2355_05a)
Odontofroggattia sp. ex F. microcarpa (Indonesia Papua Barat) (2355_03b)
Ceratosolen sp. ex F. comitis (Indonesia Papua Barat) (2196_01)
10.49
14.63
2.2
6.3
17.37
11.18
8.49
22.46
28.52
11.55
11.21
4.3
14.99
13.49
20.56
2.93
14.17
34.42
10.9
5.61
23.23
15.59
15.03
8.73
48.22
6.62
7.73
37.32
44.39
3.53
34.54
PALAEOGENE
NEOGENE
Q
EOCENE
OLIGOCENE
MIOCENE
48.6
40.4
37.2
33.9
28.4
23.0
20.4
16.0
13.9
11.6
7.2
5.3
Ma
50.0
40.0
30.0
20.0
10.0
0.0

slide prepared by Petr Janšta

Out of Australia and back again!







slide prepared by Petr Janšta