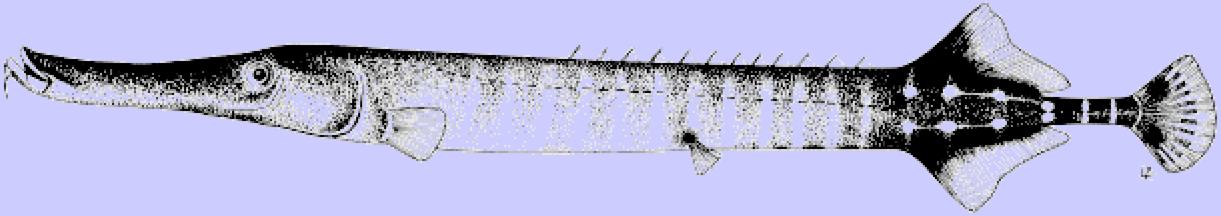
Towards a Reliability Index for Clades: an Application on Acantomorph Teleosts

Blaise Li and Guillaume Lecointre



UMR 7138
Département Systématique et Évolution
Muséum National d'Histoire Naturelle - Paris

ASIH annual meeting - 11 July, 2005

How to measure a result's quality ?

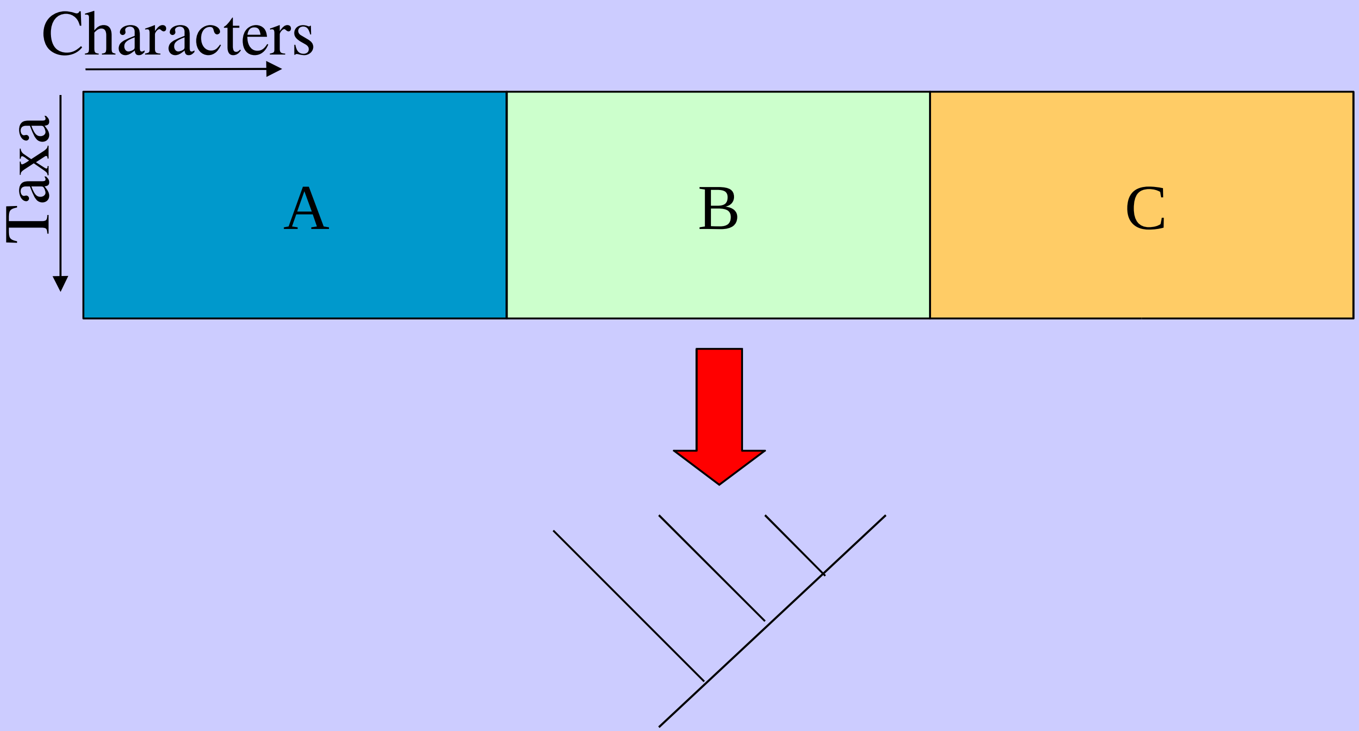
- Does it resist data perturbation (robustness) ?
- What is the credit given to a statement about relationships among species (**reliability**) ?

How to measure reliability ?

Use multiple data sets

Combine the data into a single matrix or keep it separate ?

Combined analysis (« Total Evidence »)



Characters

Taxa

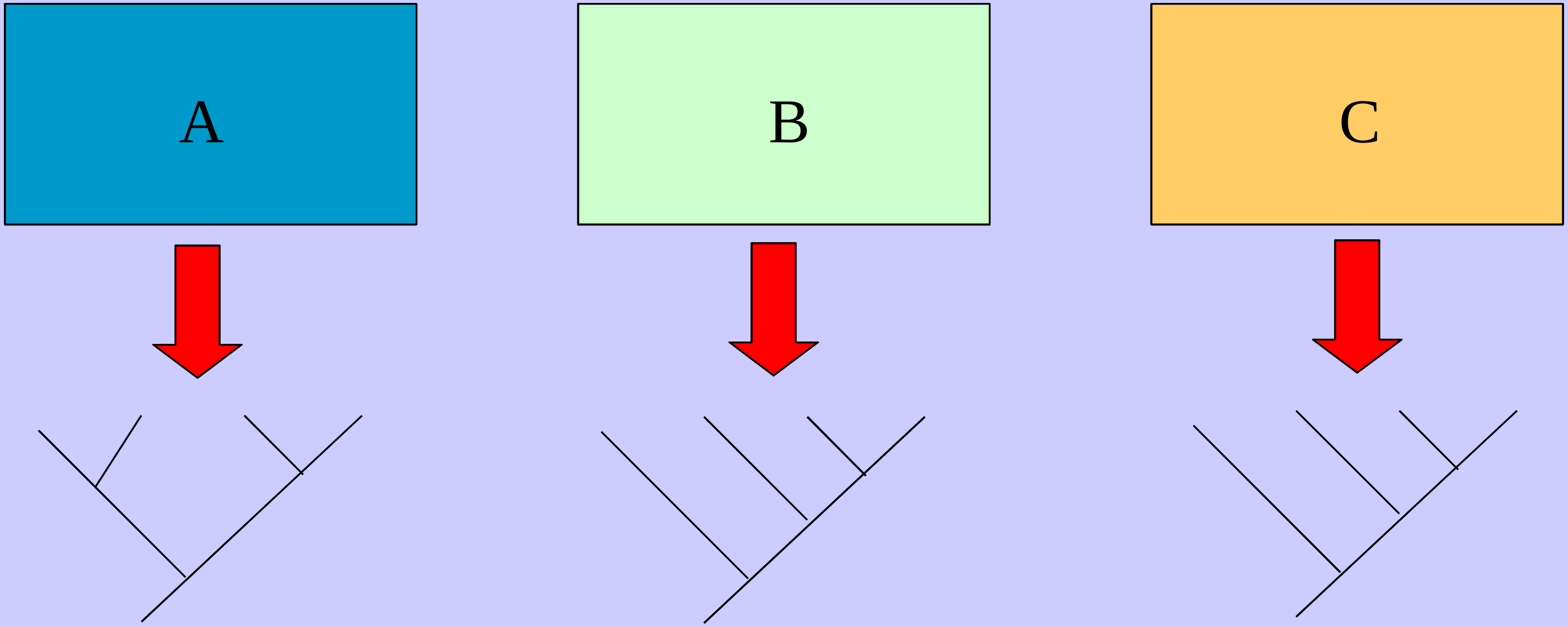
A

B

C

Drawback: a marker-specific bias can influence the inference from the whole data during the optimization process

Separate analysis



Biological background knowledge is needed to justify delineation and independence of the data sets

Drawbacks: higher stochastic effects
full expression of marker-specific biases

How to measure reliability ?

Consider corroboration between **independent parts** of the data (with **partial data combination**)

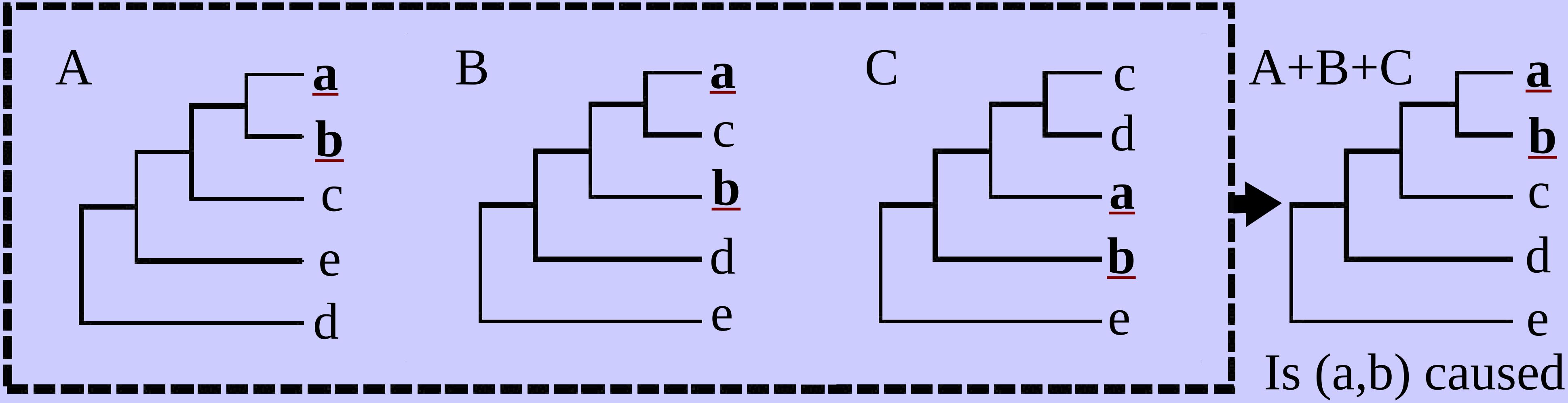
About independent data

Choose basic markers you postulate to evolve independently

Thus, if a marker is subject to some bias, you would expect that bias not to exist for the other markers also

It increases the chance that **repeated results** are caused by a shared feature of the data sets: the history of the taxa represented by the markers

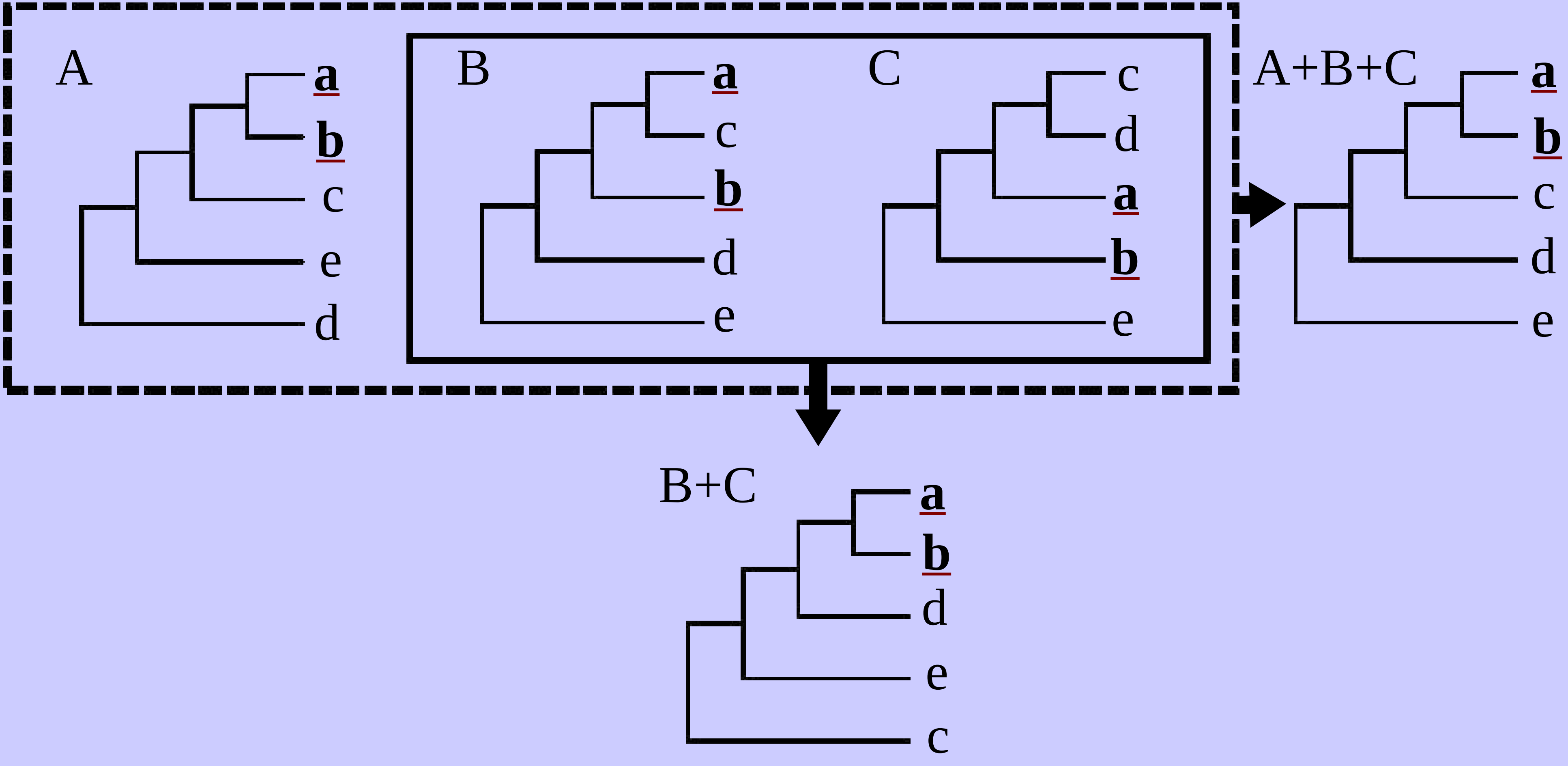
Partial combinations...



Is (a,b) caused by A only ?

A bears signal for (a,b). Is it true historical signal ?

... to reduce stochastic effects

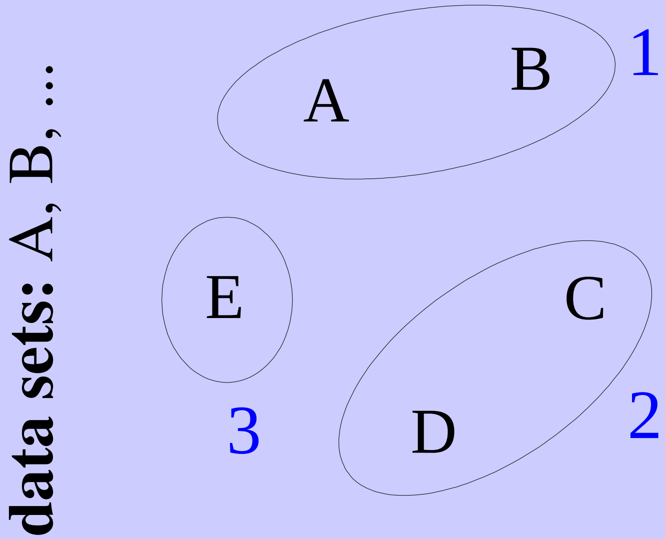


B+C corroborates (a,b). Partial combination overcame some stochastic effects that prevented (a,b) from appearing.

A repeatability index for clades: a way to formalize reliability

The more a clade is found by the analysis of independent data, the more reliable it is

- Separate the data into non-overlapping **parts** and analyze each part with the same method



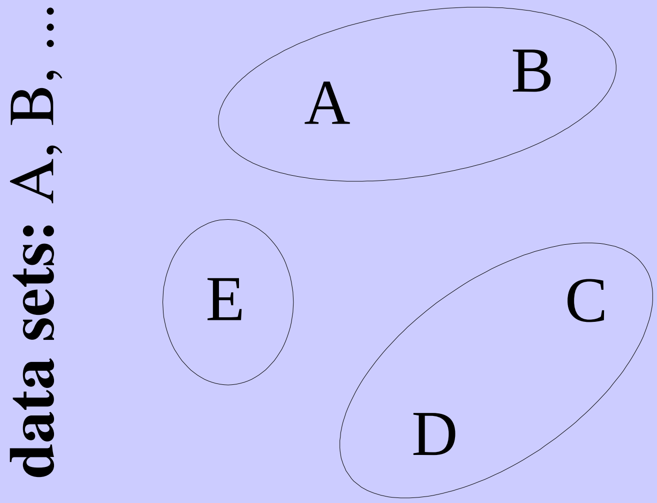
One of the possible partitioning schemes with 3 parts (1, 2, 3)

A repeatability index for clades: a way to formalize reliability

The more a clade is found by the analysis of independent data, the more reliable it is

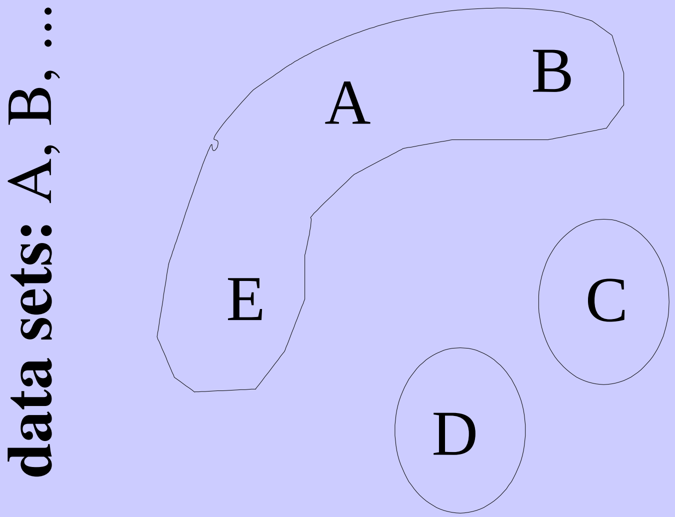
- Separate the data into non-overlapping **parts** and analyze each part with the same method
- Count the occurrences of the clades among non-overlapping parts
- Repeat the process with each possible **partitioning scheme**
- For each clade, retain the highest number of occurrences over all the partitioning schemes

A partitioning scheme with 3 parts



clade	occurrences
α:	**3** occurrences
β:	**3** occurrences
γ:	**2** occurrences
δ:	**2** occurrences
ϵ:	1 occurrence
ζ:	1 occurrence
...	

Another partitioning scheme with 3 parts



clade	occurrences
α:	2 occurrences
β:	**3** occurrences
γ:	1 occurrence
δ:	**2** occurrences
ϵ:	**3** occurrences
ζ:	**2** occurrences
...	

Taking into account probabilities for clades to appear at random

- Clades appear also from data containing no signal
- Therefore one must measure the probability of a clade to appear purely at random for a given program:
  - give random data to the program
  - analyze that data
  - repeat the process and compute the frequency of each clade

data sets

	part 1		part 2		part 3
	A B		C D		E
α (3 occurrences):	$1\text{-}P_1(\alpha)$	+	$1\text{-}P_2(\alpha)$	+	$1\text{-}P_3(\alpha)$
β (3 occurrences):	$1\text{-}P_1(\beta)$	+	$1\text{-}P_2(\beta)$	+	$1\text{-}P_3(\beta)$
γ (2 occurrences):	$1\text{-}P_1(\gamma)$	+	0	+	$1\text{-}P_3(\gamma)$
δ (2 occurrences):	0	+	$1\text{-}P_2(\delta)$	+	$1\text{-}P_3(\delta)$
ϵ (1 occurrence):	$1\text{-}P_1(\epsilon)$	+	0	+	0
ζ (1 occurrence):	0	+	$1\text{-}P_2(\zeta)$	+	0
...	...		...		...

Taking into account contradiction among clades

- First order reliability of clade α:

$$R_1(\alpha)=\mathrm{Max}_{\text{partitioning schemes}}(\sum_{\text{parts}}(\text{Occurrences-P}))$$

- Second order reliability of clade α:

$$R_2(\alpha)=R_1(\alpha)-R_1(\beta_1)$$

β_1 being the highest R_1 α contradictor

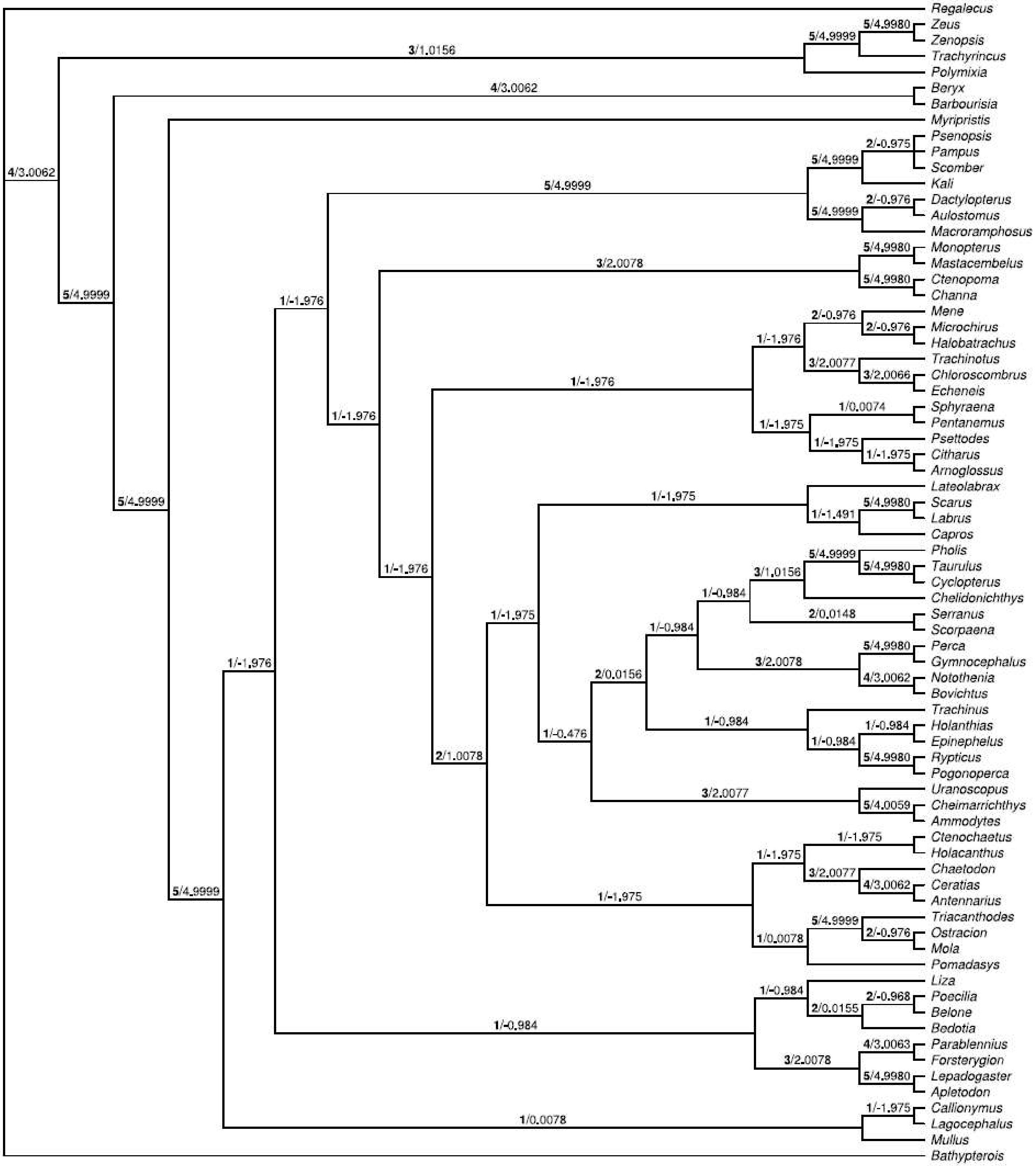
- Etc...

$$R_n(\alpha)=R_1(\alpha)-R_1(\beta_{n-1})$$

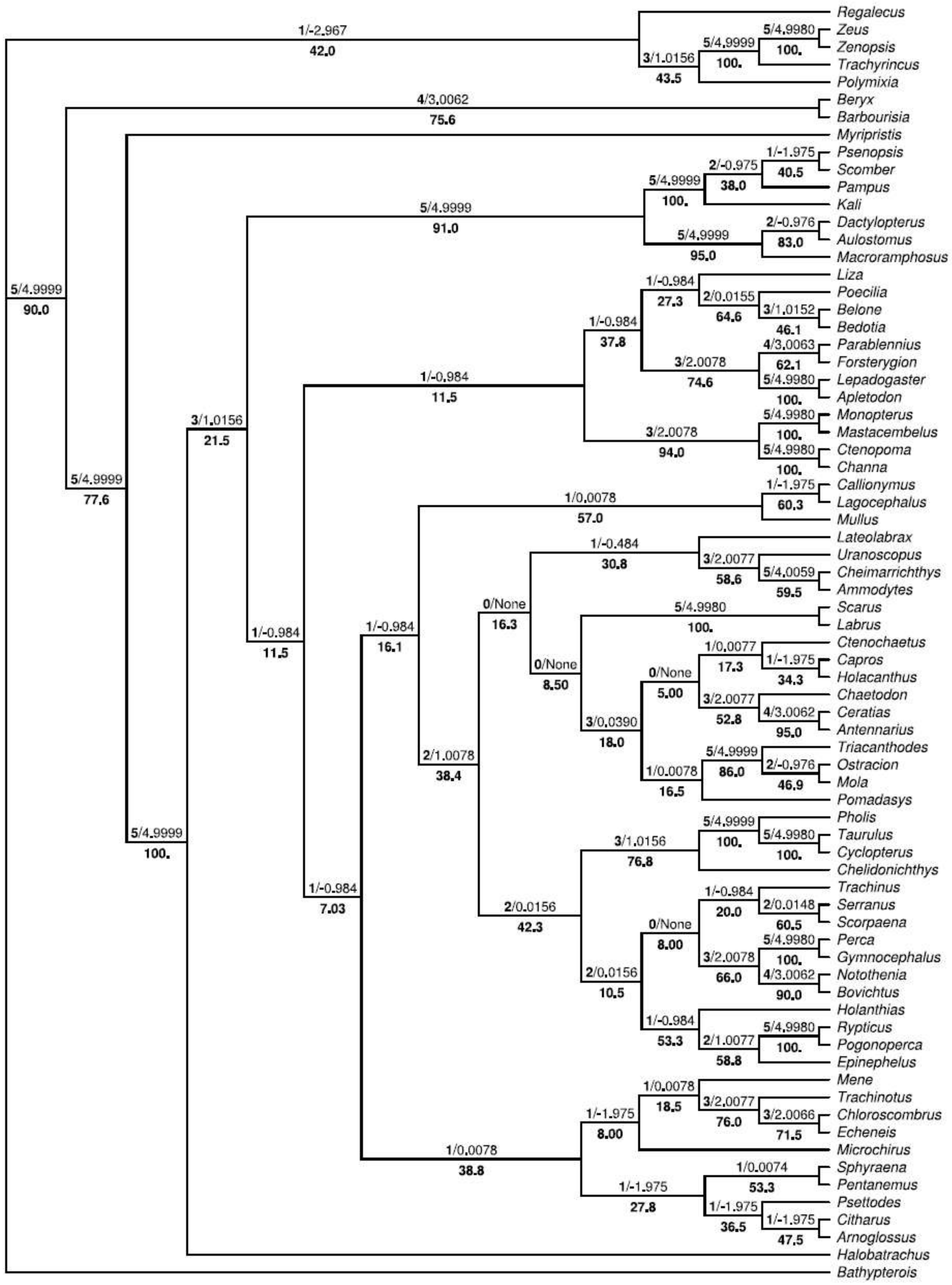
An application on Acanthomorpha

- 5 basic data sets:
  - mitochondrial markers: partial 12S+16S (828 bp)
  - nuclear markers: partial 28S (801 bp), partial Rhodopsin (759 bp), partial MLL (552 bp), partial IRBP (713 bp)
- 73 taxa shared by all data sets
- 31 ways to combine the 5 independent data sets (allowing to compose 51 partitioning schemes), each analyzed with Paup 4 under Maximum Parsimony

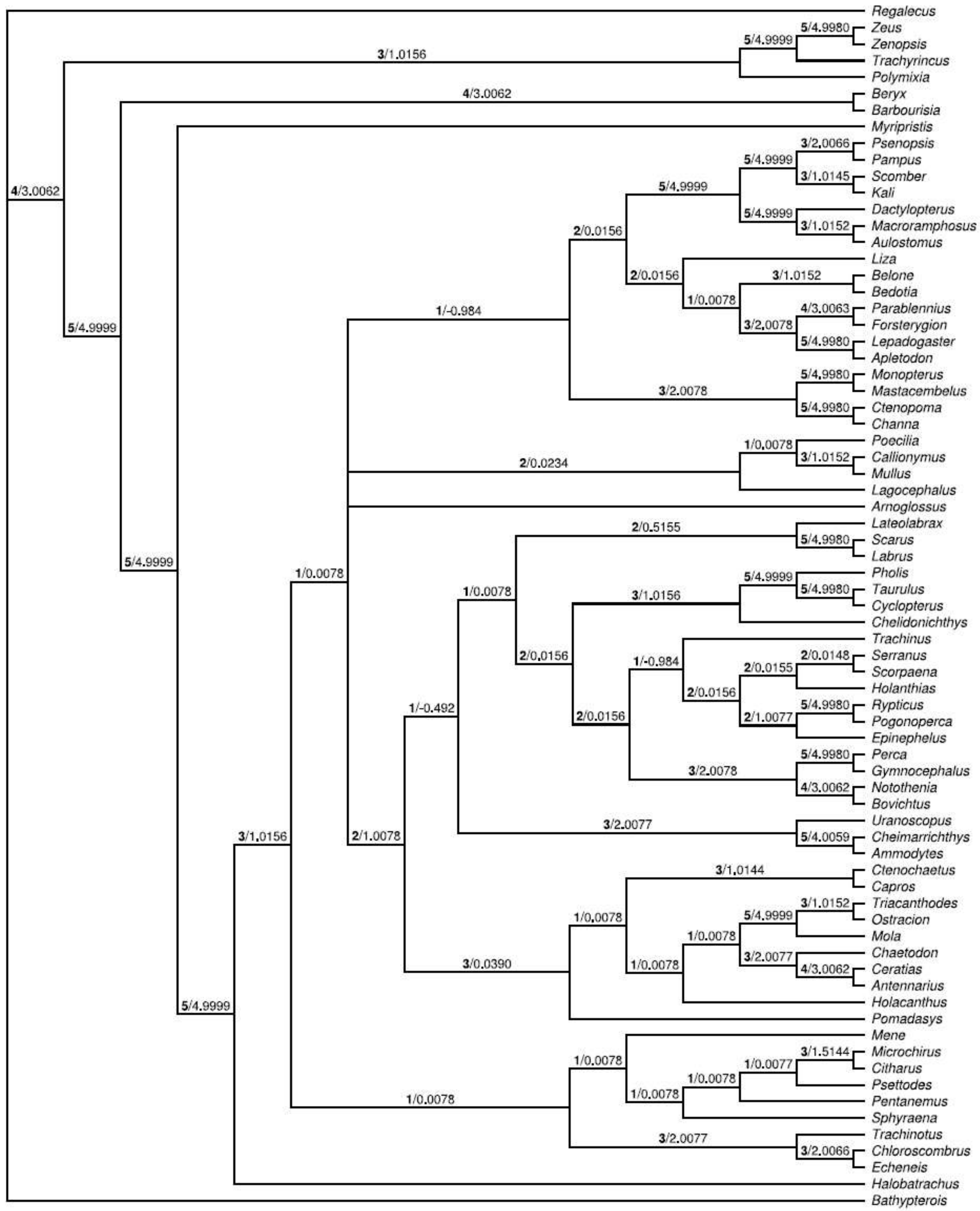
Total evidence majority rule consensus tree



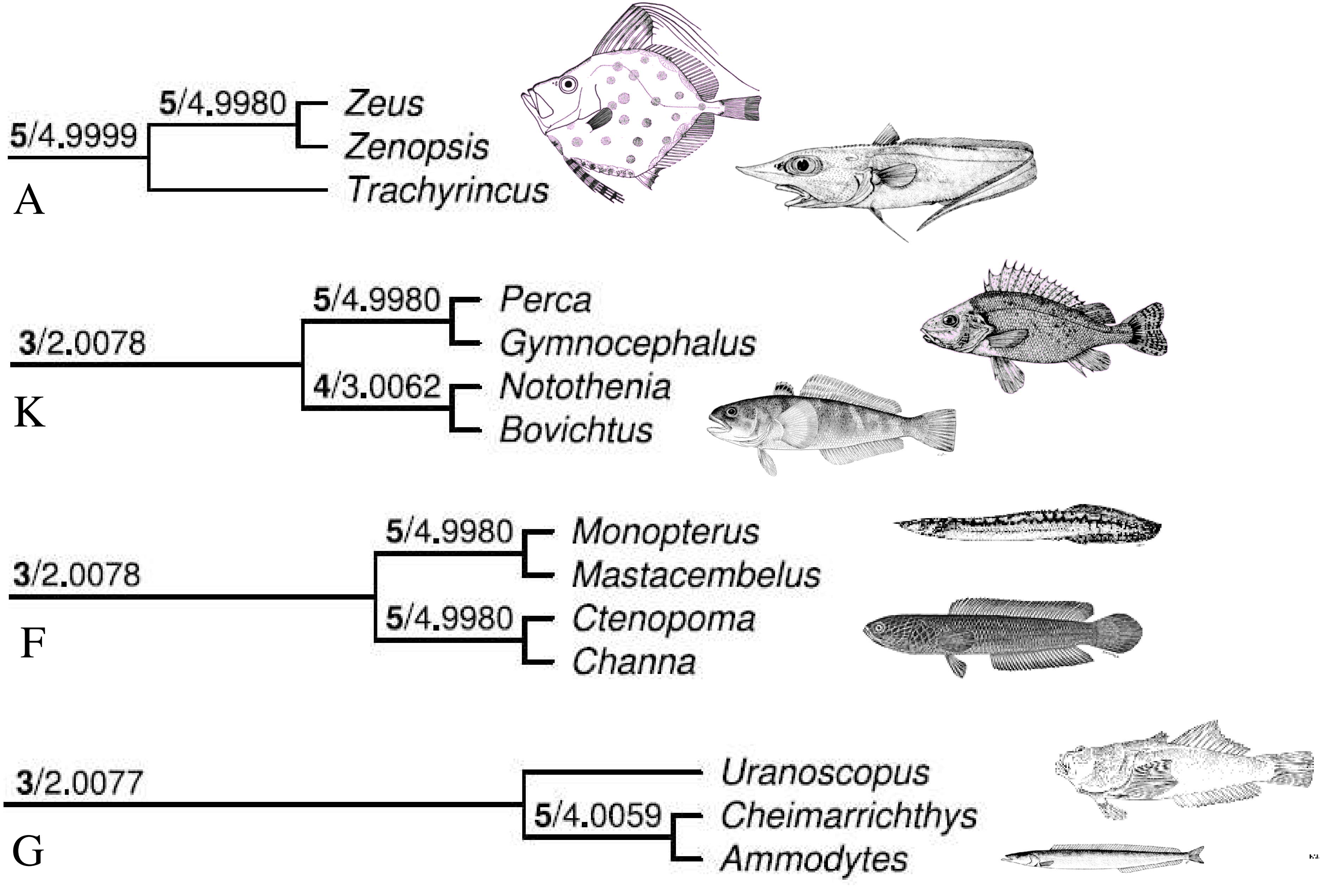
Majority-rule bootstrap tree for the total data combination



Tree constructed to include the highest reliability inter-compatible clades



A few reliable clades



5/4.9999

5/4.9980

Zeus

Zenopsis

Trachyrincus

A

3/2.0078

5/4.9980

Perca

Gymnocephalus

4/3.0062

Notothenia

Bovichtus

K

3/2.0078

5/4.9980

Monopterus

Mastacembelus

5/4.9980

Ctenopoma

Channa

F

3/2.0077

Uranoscopus

5/4.0059

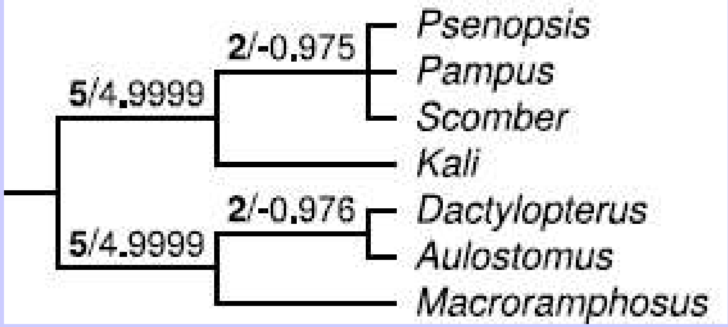
Cheimarrichthys

Ammodytes

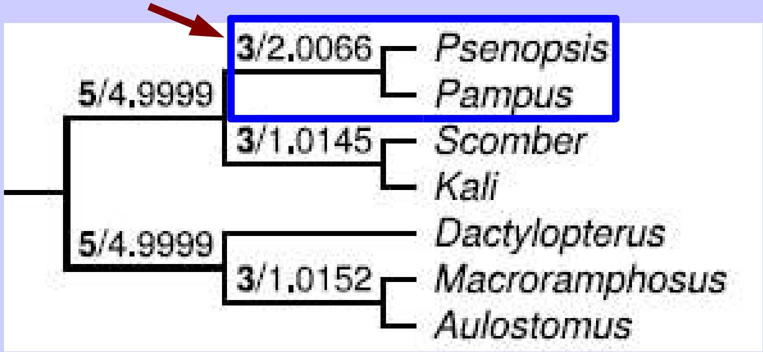
G

Reliability is not robustness

high reliability, neither present in the total evidence tree, nor in the bootstrap consensus



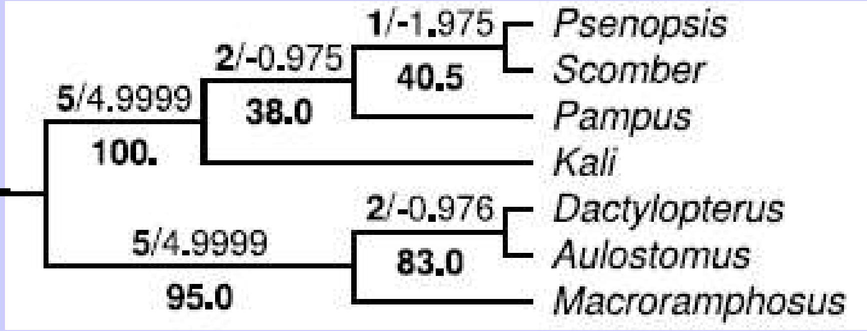
Total evidence



Best repeatability scores

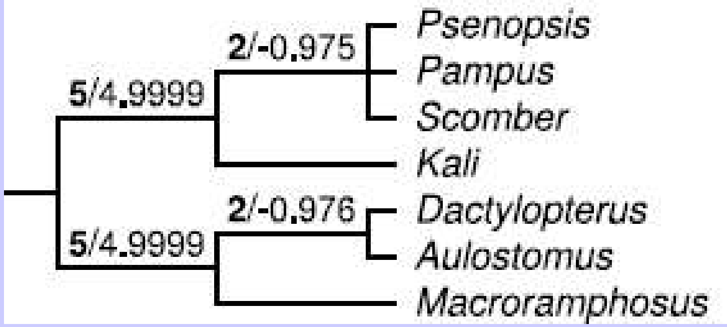
max occurrences/repeatability score

bootstrap value

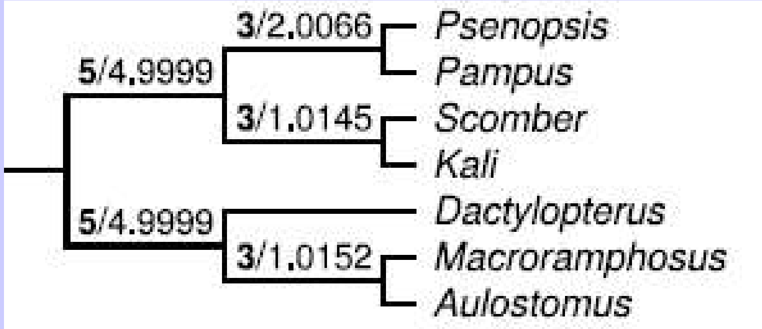


Bootstrap

Reliability is not robustness



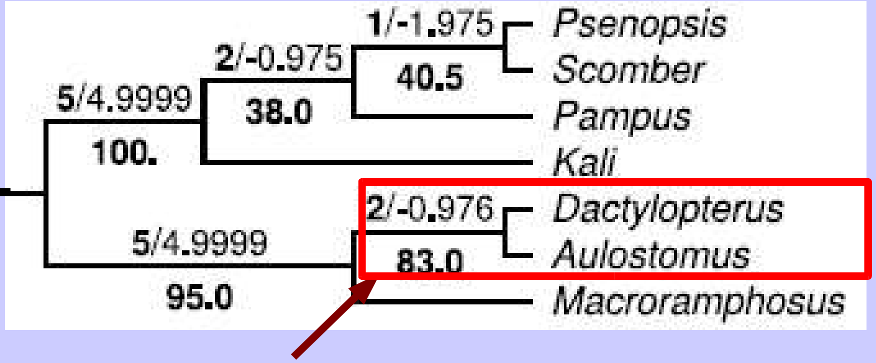
Total evidence



Best repeatability scores

max occurrences/repeatability score

bootstrap value



high robustness,
low reliability

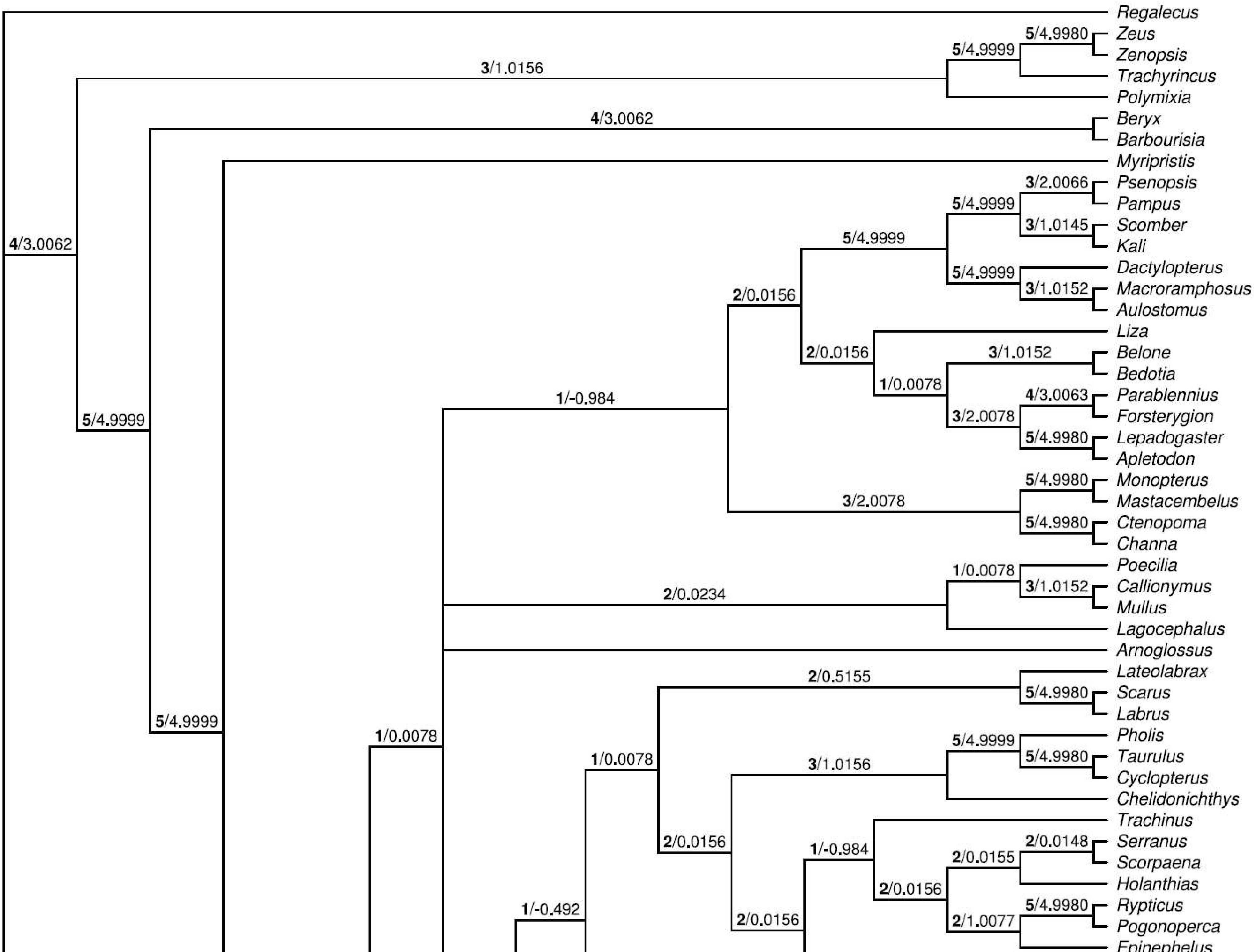
Bootstrap

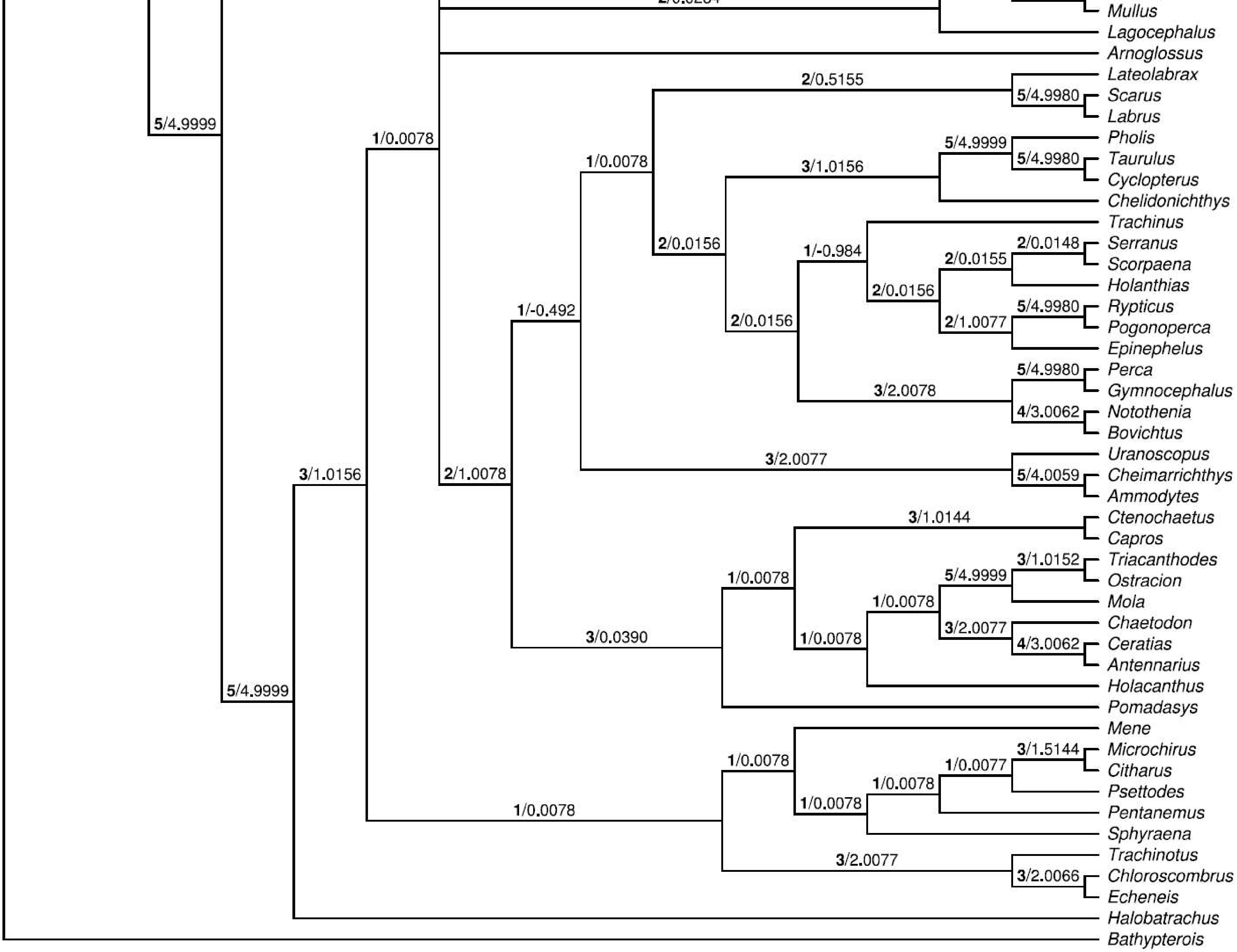
A short conclusion because it's time to finish

- Some non-robust clades have been identified as reliable by our method (Psenopsis, Pampus)
- Some clades with high boostrap support are not considered reliable by our method (Dactylopterus, Aulostomus)
- Works with completely shared taxonomic samplings

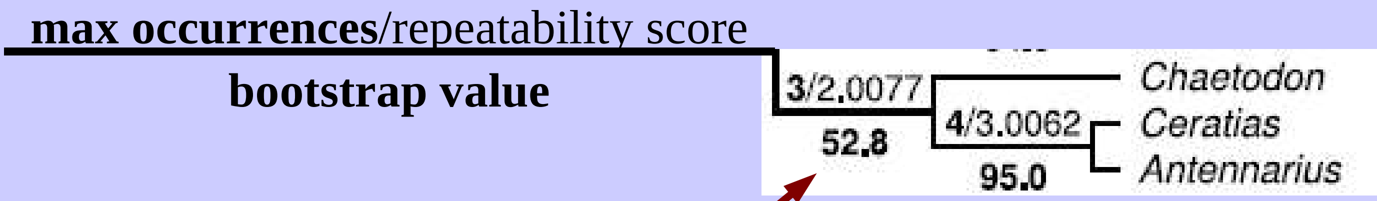
Acknowledgements

- G. Lecointre, W.-J. Chen and A. Dettaï
- Service de Systématique Moléculaire (MNHN)
- École Normale Supérieure and French Ministry of Research
- Free software programers


Regalecus
5/4.9980
Zeus
5/4.9999
Zenopsis
3/1.0156
Trachyrincus
Polymixia
4/3.0062
Beryx
Barbourisia
Myripristis
3/2.0066
Psenopsis
5/4.9999
Pampus
3/1.0145
Scomber
4/3.0062
Kali
5/4.9999
Dactylopterus
5/4.9999
Macroramphosus
3/1.0152
2/0.0156
Aulostomus
Liza
2/0.0156
3/1.0152
Belone
Bedotia
1/0.0078
4/3.0063
Parablennius
1/-0.984
3/2.0078
Forsterygion
5/4.9999
5/4.9980
Lepadogaster
Apletodon
5/4.9980
Monopterus
3/2.0078
Mastacembelus
5/4.9980
Ctenopoma
Channa
1/0.0078
Poecilia
3/1.0152
Callionymus
2/0.0234
Mullus
Lagocephalus
Arnoglossus
2/0.5155
Lateolabrax
5/4.9980
Scarus
5/4.9999
Labrus
1/0.0078
Pholis
5/4.9999
5/4.9980
Taurulus
1/0.0078
3/1.0156
Cyclopterus
Chelidonichthys
Trachinus
2/0.0148
Serranus
2/0.0156
1/-0.984
Scorpaena
2/0.0155
Holanthias
2/0.0156
5/4.9980
Rypticus
1/-0.492
2/0.0156
2/1.0077
Pogonoperca
Epinephelus


Mullus
Lagocephalus
Arnoglossus
2/0.5155
Lateolabrax
5/4.9980
Scarus
Labrus
5/4.9999
1/0.0078
Pholis
5/4.9999
1/0.0078
5/4.9980
Taurulus
3/1.0156
Cyclopterus
Chelidonichthys
Trachinus
2/0.0148
Serranus
1/-0.984
2/0.0156
2/0.0155
Scorpaena
Holanthias
2/0.0156
1/-0.492
5/4.9980
Rypticus
2/0.0156
2/1.0077
Pogonoperca
Epinephelus
5/4.9980
Perca
3/2.0078
Gymnocephalus
4/3.0062
Notothenia
Bovichtus
3/2.0077
Uranoscopus
5/4.0059
Cheimarrichthys
3/1.0156
2/1.0078
Ammodytes
3/1.0144
Ctenochaetus
Capros
3/1.0152
Triacanthodes
1/0.0078
5/4.9999
Ostracion
Mola
1/0.0078
Chaetodon
3/0.0390
1/0.0078
3/2.0077
4/3.0062
Ceratias
Antennarius
5/4.9999
Holacanthus
Pomadasys
Mene
3/1.5144
Microchirus
1/0.0078
1/0.0077
Citharus
1/0.0078
Psettodes
1/0.0078
1/0.0078
Pentanemus
Sphyraena
Trachinotus
3/2.0077
3/2.0066
Chloroscombrus
Echeneis
Halobatrachus
Bathypterois

Reliability is not robustness (2)



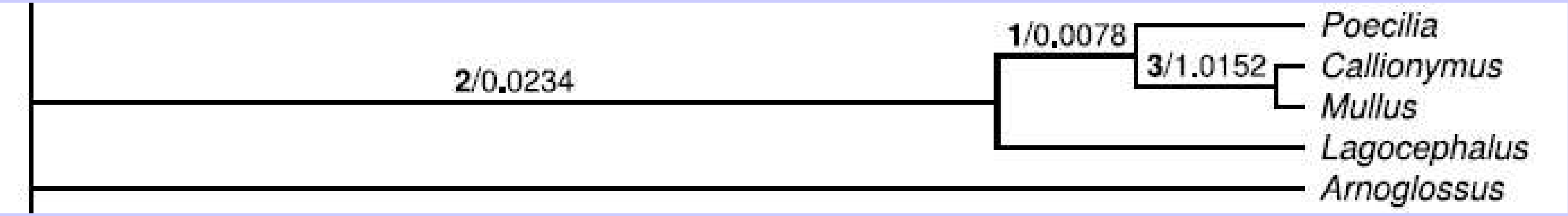
max occurrences/repeatability score

bootstrap value

Not very robust
but rather reliable

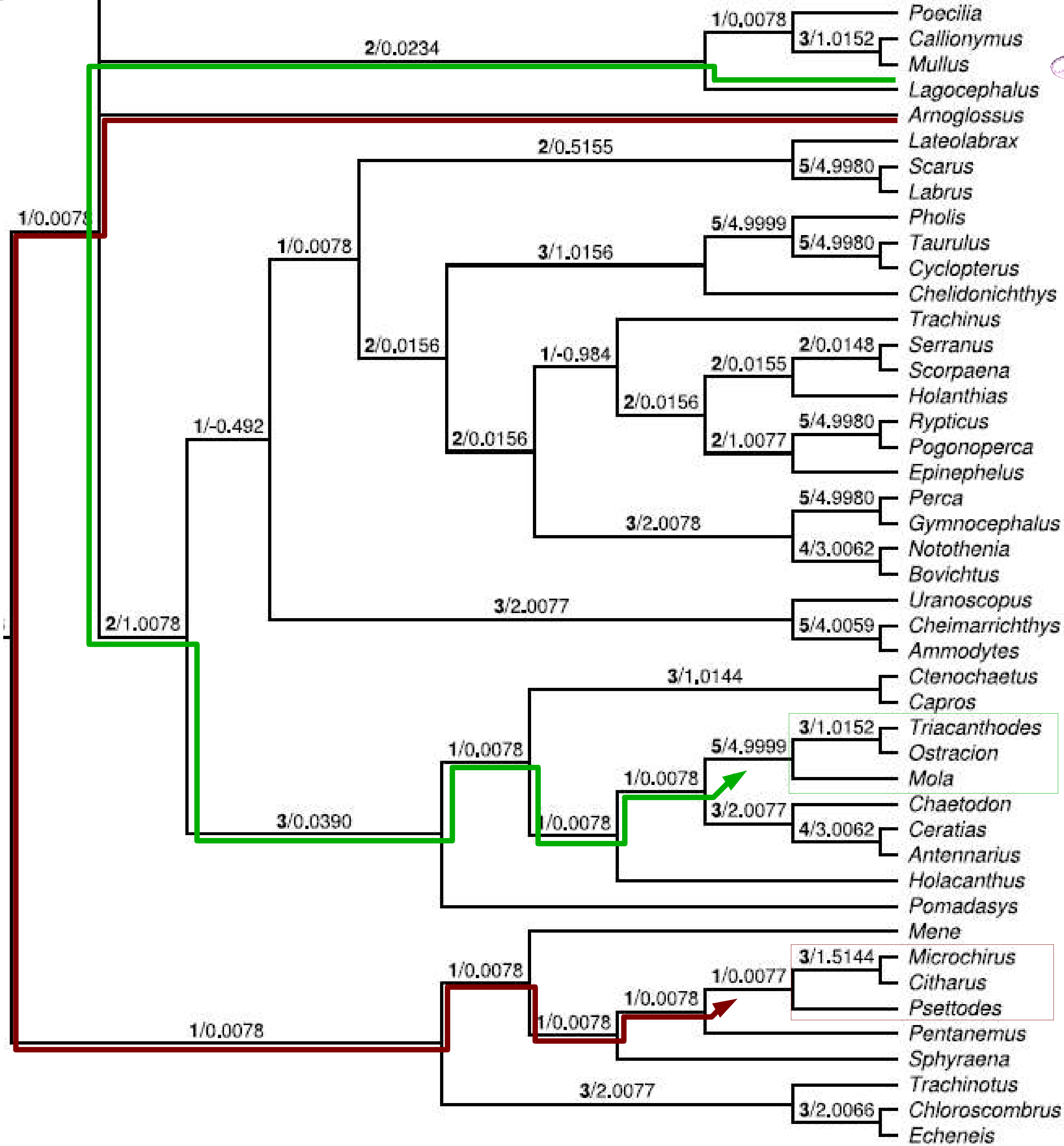
Bootstrap

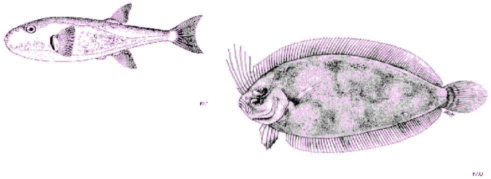
Limits of the method



General long-branch
attraction is still a problem

Best repeatability scores





Long branches to swim back, following the low-reliability track !

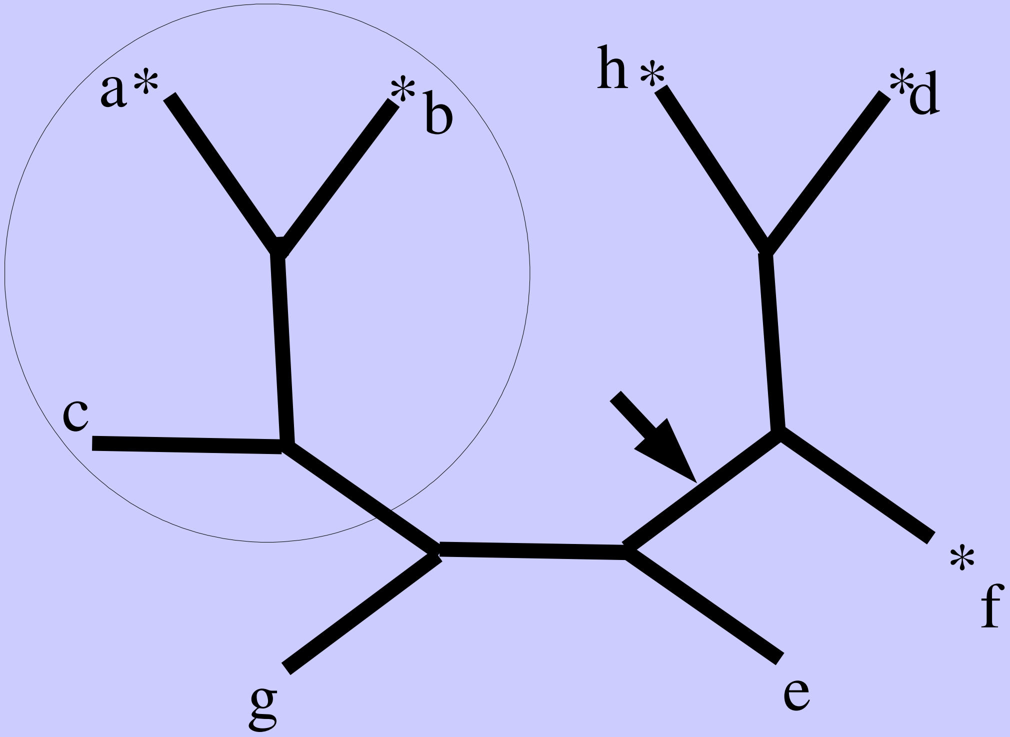
Taking into account probabilities of clades to appear by chance

- One must measure them for a given tree reconstruction program:
  - produce data without signal (randomly chosen character states)
  - analyze it and count the clades
  - reproduce the experiment many times
- Problem: it takes a lot of time

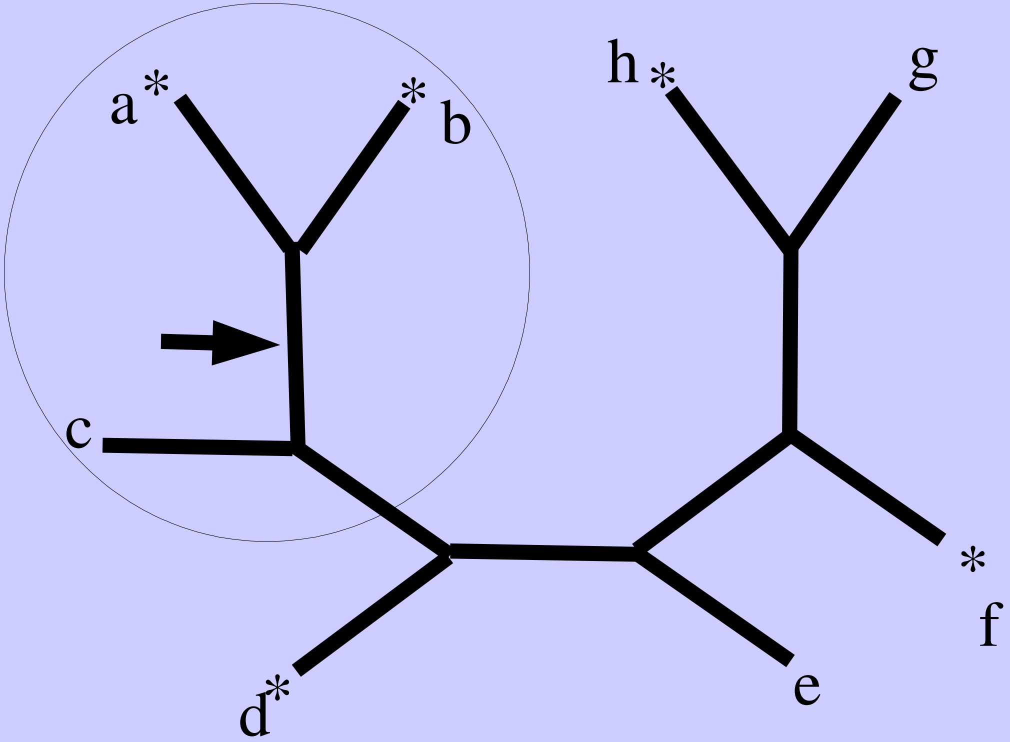
Estimating probabilities of clades knowing their sizes

- All groups of the same size are equivalent
  - random data
  - random addition sequence
  - work on unrooted trees
- Shift to rooted trees
  - clades that appear in rooted trees were there before rooting
  - but clades present in unrooted trees may disappear at rooting if there are 4 outgroups or more -> one overestimates probabilities of clades containing outgroup taxa

* : outgroup taxon



The clade is in the unrooted tree and in the rooted one

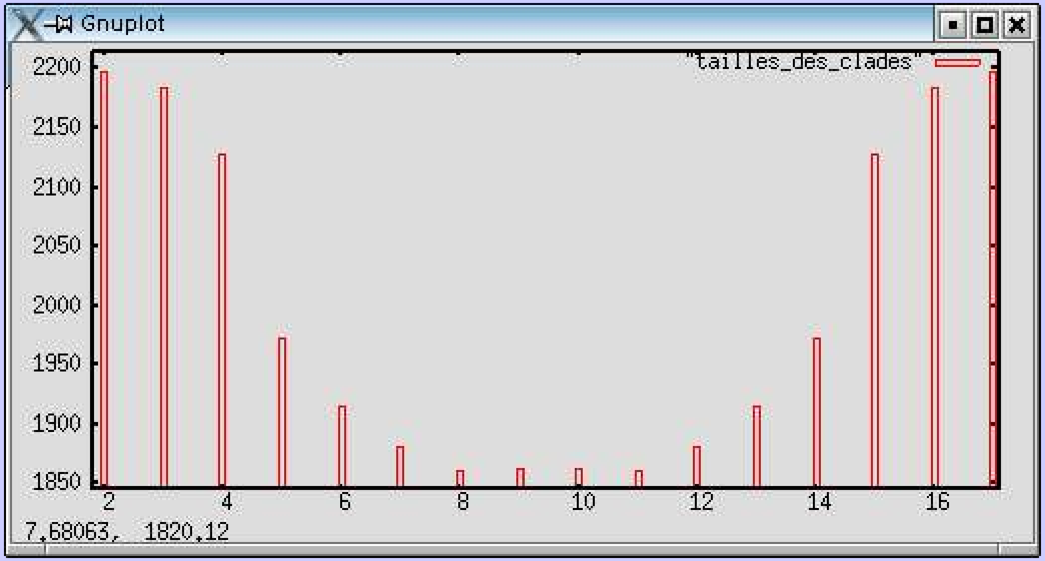


The clade is in the unrooted tree but not in the rooted one

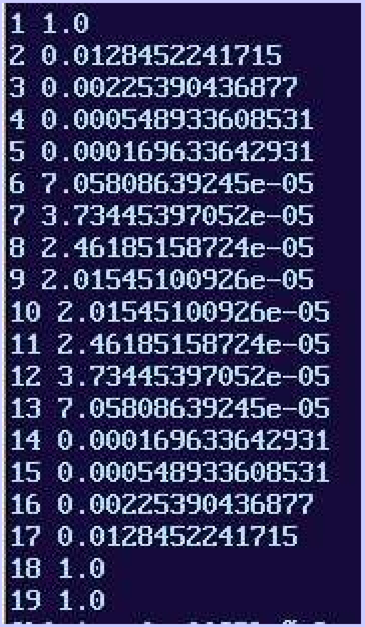
taxon 1 ATCCGTGGCAATCGG...
taxon 2 CCTAGGTGGCGAAAT...
taxon 3 AACATTGCGAACCTC...
...

Phylogenetic analysis program

This is a black box



Frequencies of the clades:



```
1 1.0
2 0.0128452241715
3 0.00225390436877
4 0.000548933608531
5 0.000169633642931
6 7.05808639245e-05
7 3.73445397052e-05
8 2.46185158724e-05
9 2.01545100926e-05
10 2.01545100926e-05
11 2.46185158724e-05
12 3.73445397052e-05
13 7.05808639245e-05
14 0.000169633642931
15 0.000548933608531
16 0.00225390436877
17 0.0128452241715
18 1.0
19 1.0
```

From counts to probabilities

$N(t)$: number of clades including t taxa

T: number of trees examined

$N(t)/T$ = mean number of clades including t taxa per tree

n: number of taxa in the trees

$n!/t!(n-t)!$: number of possible clades including t taxa when there are n taxa in the study

$P(t) = (N(t)/T)/(n!/t!(n-t)!)$