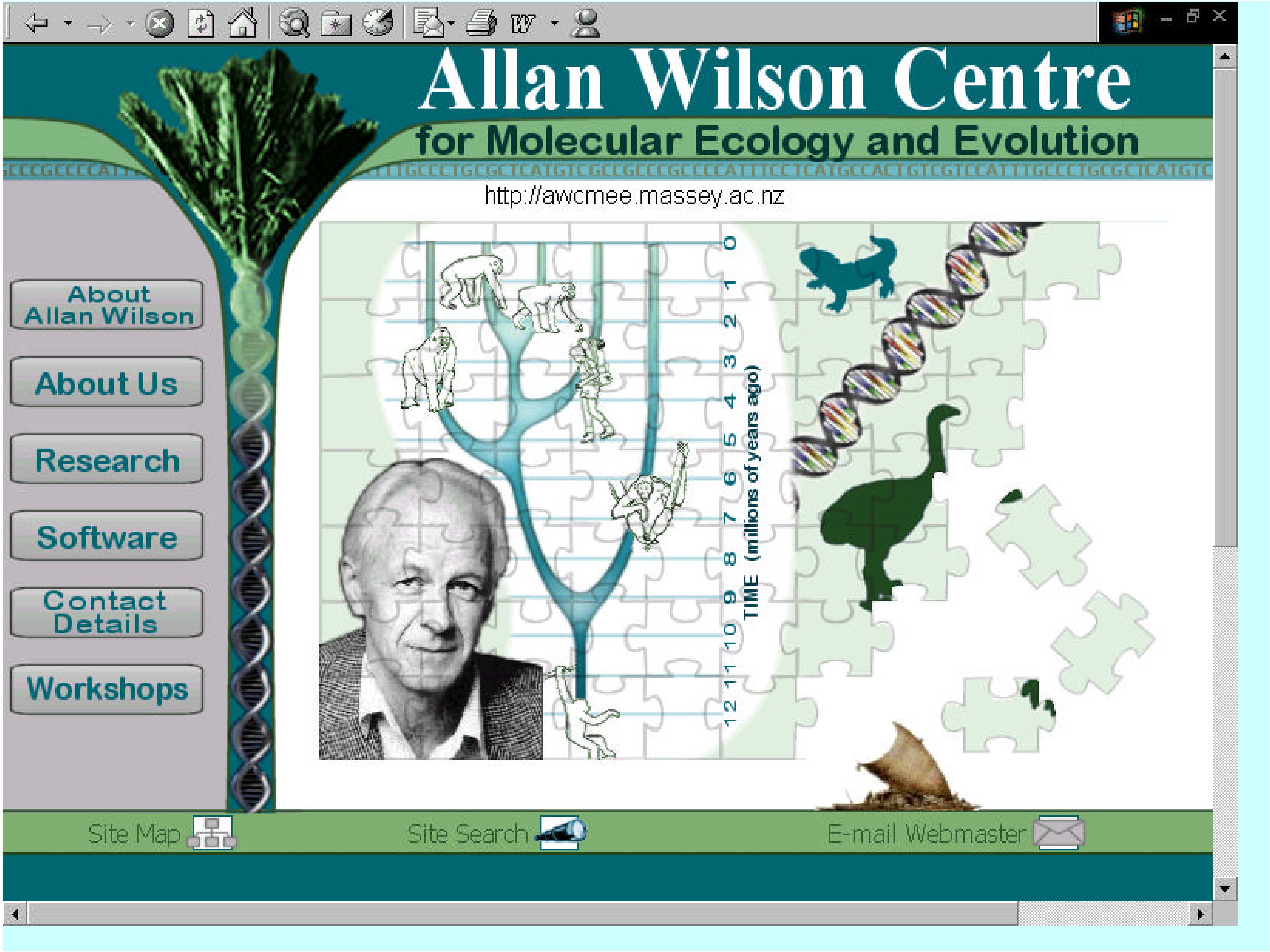


Allan Wilson Centre

for Molecular Ecology and Evolution

http://awcmee.massey.ac.nz

An introduction to trees and networks

David Penny

Allan Wilson Centre for
Molecular Ecology and Evolution

overview

trees and networks (with cycles)

spanning and phylogenetic (Steiner) trees

exploratory data analysis, visualization

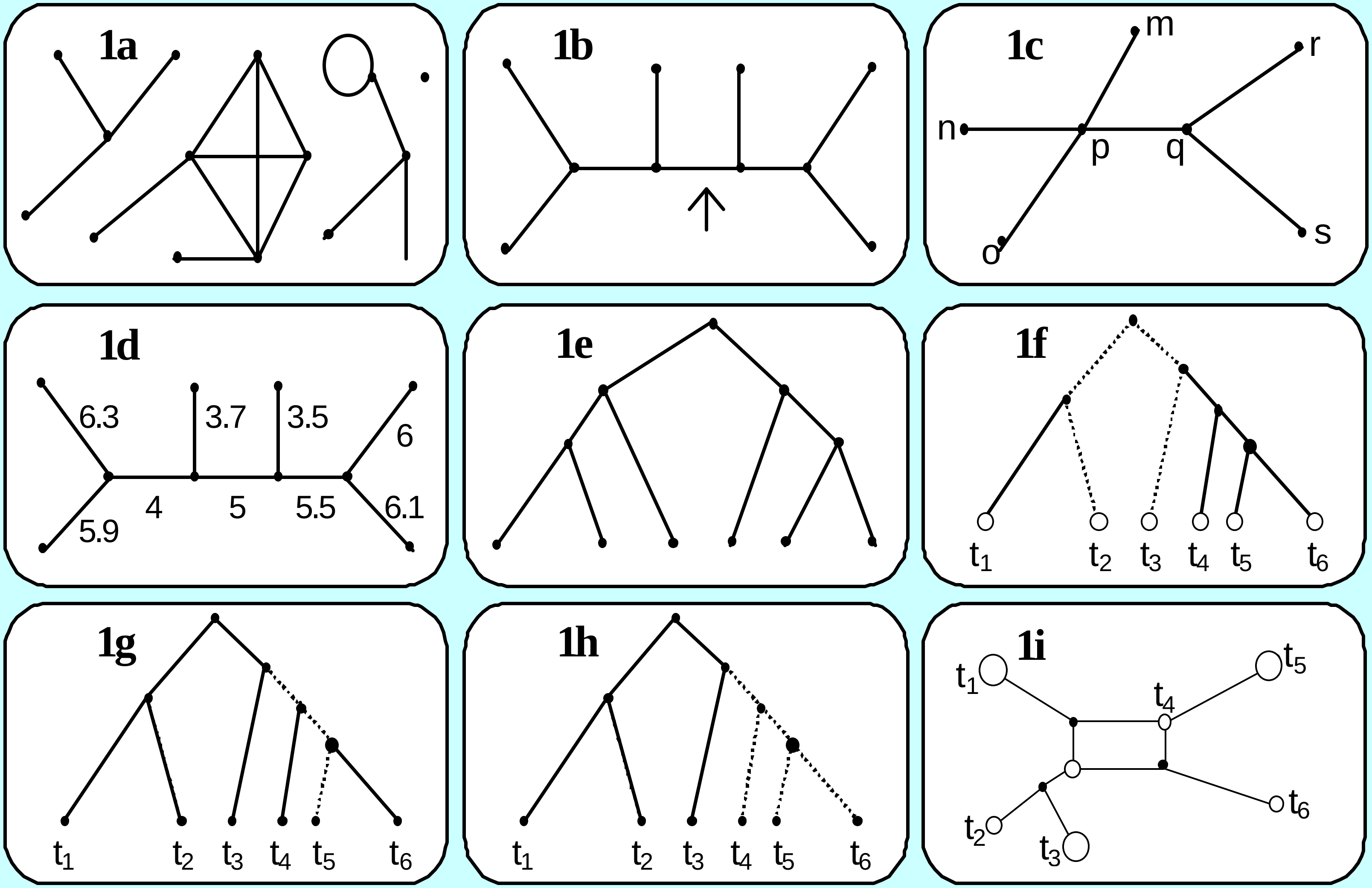
multiscale

splits

compatibility/incompatibility

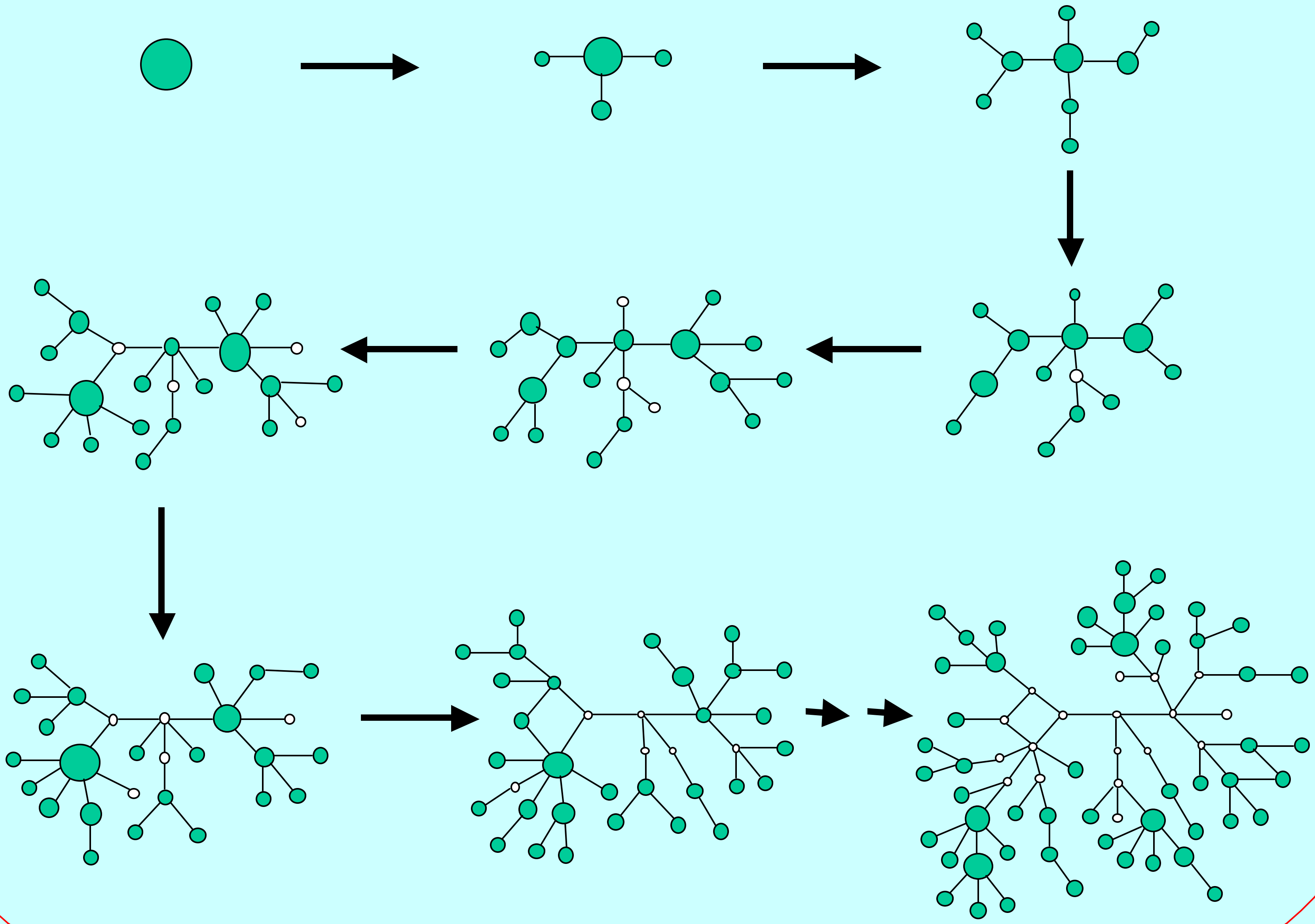
application - the MinMax squeeze

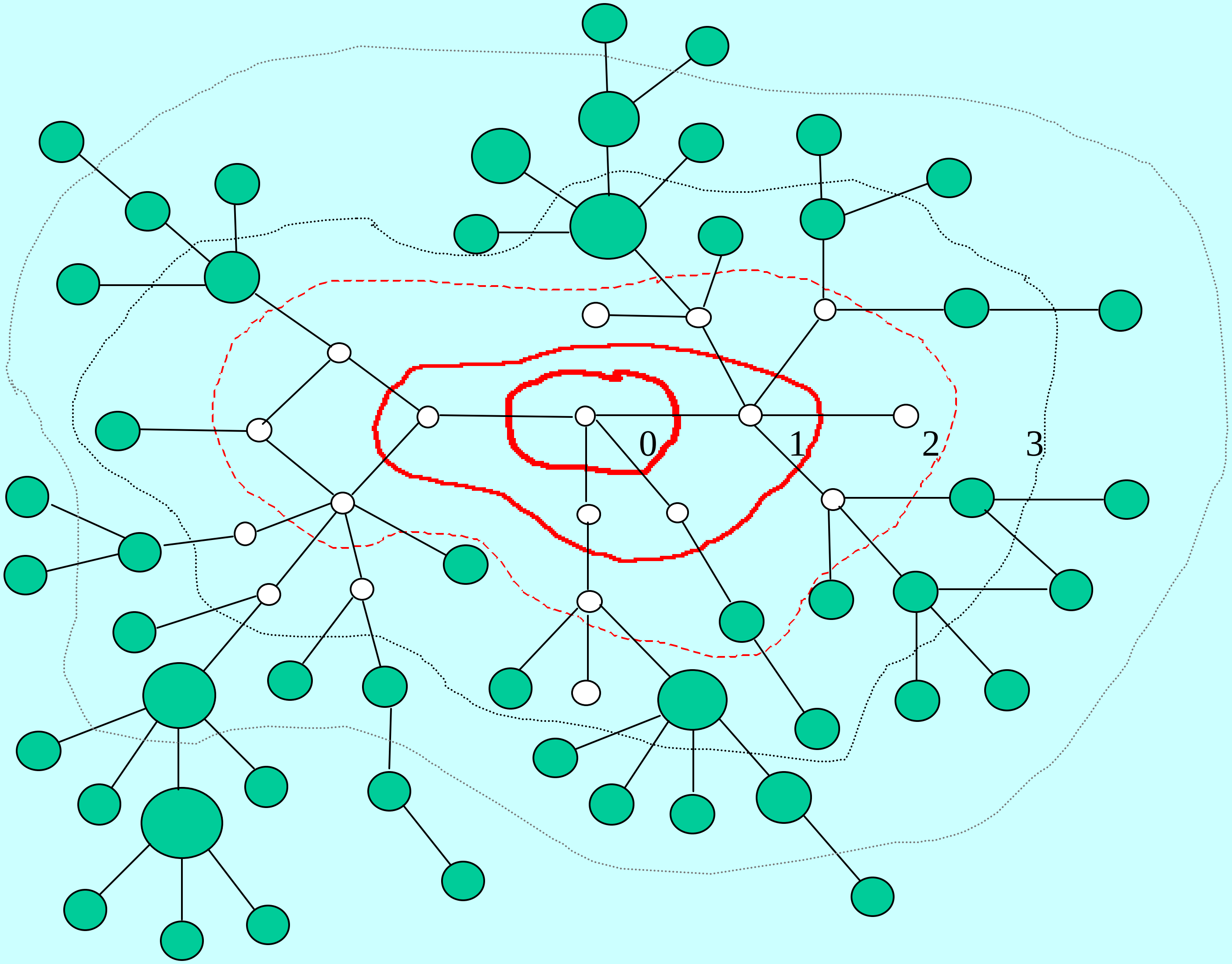
tree terminology



1c, spanning tree

1f, phylogenetic (Steiner) tree





lose nodes near the root

mixture of spanning and Steiner trees

minimum spanning subtrees

even using complete sequences

the question of multiscale

effects of time scale

sequences abundant ↔ sequences sparse

ancestral states still present ↔ speciation only at tips

multifurcations and cycles ↔ binary phylogenetic trees

sites in same rate class (gamma) ↔ sites on or off, covarion model

neutral plus deleterious ↔ hypervariable sites ↔ average over all sites

TIME SCALE →

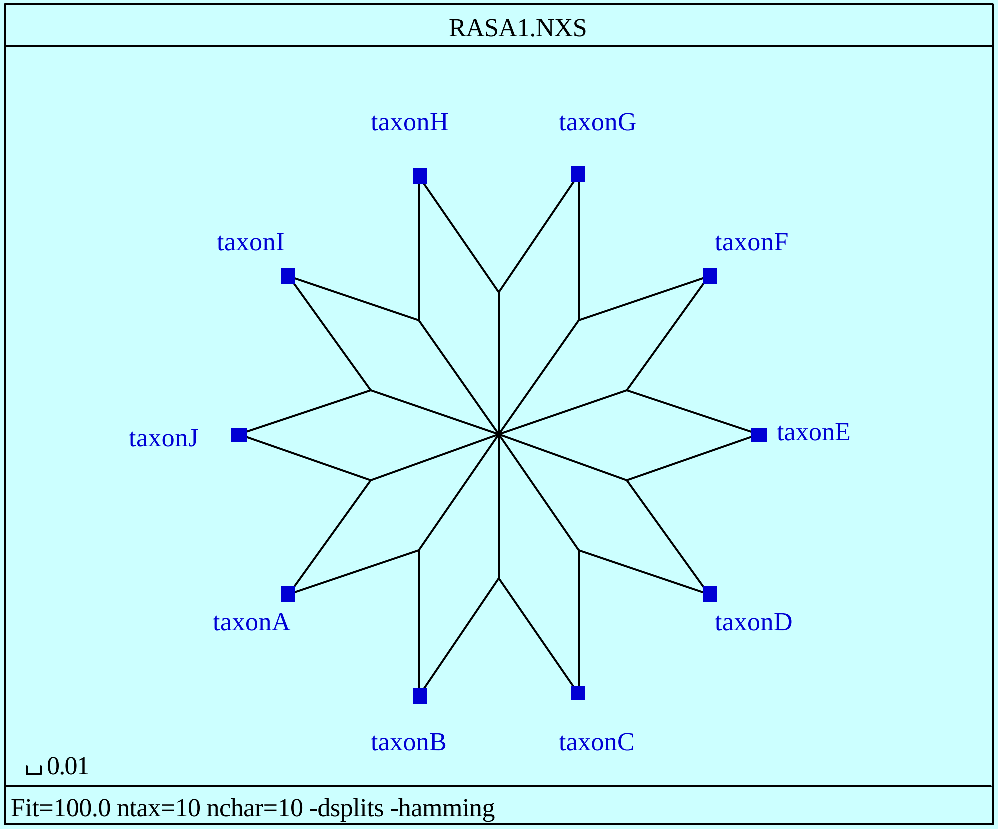
generations populations species families orders classes phyla

microevolution ↔ macroevolution

dataset used to argue poor performance of RASA

Mol. Biol. Evol. 19:14–23. 2002

taxon A	1000010000
taxon B	1000001000
taxon C	0100001000
taxon D	0100000100
taxon E	0010000100
taxon F	0010000010
taxon G	0001000010
taxon H	0001000001
taxon I	0000100001
taxon J	0000110000



Fit=100.0 ntax=10 nchar=10 -dsplits -hamming

some reasons for cycles?

repeated mutations – random (insufficient information)

- systematic [e.g. GC bias]

- positive selection

hybridization

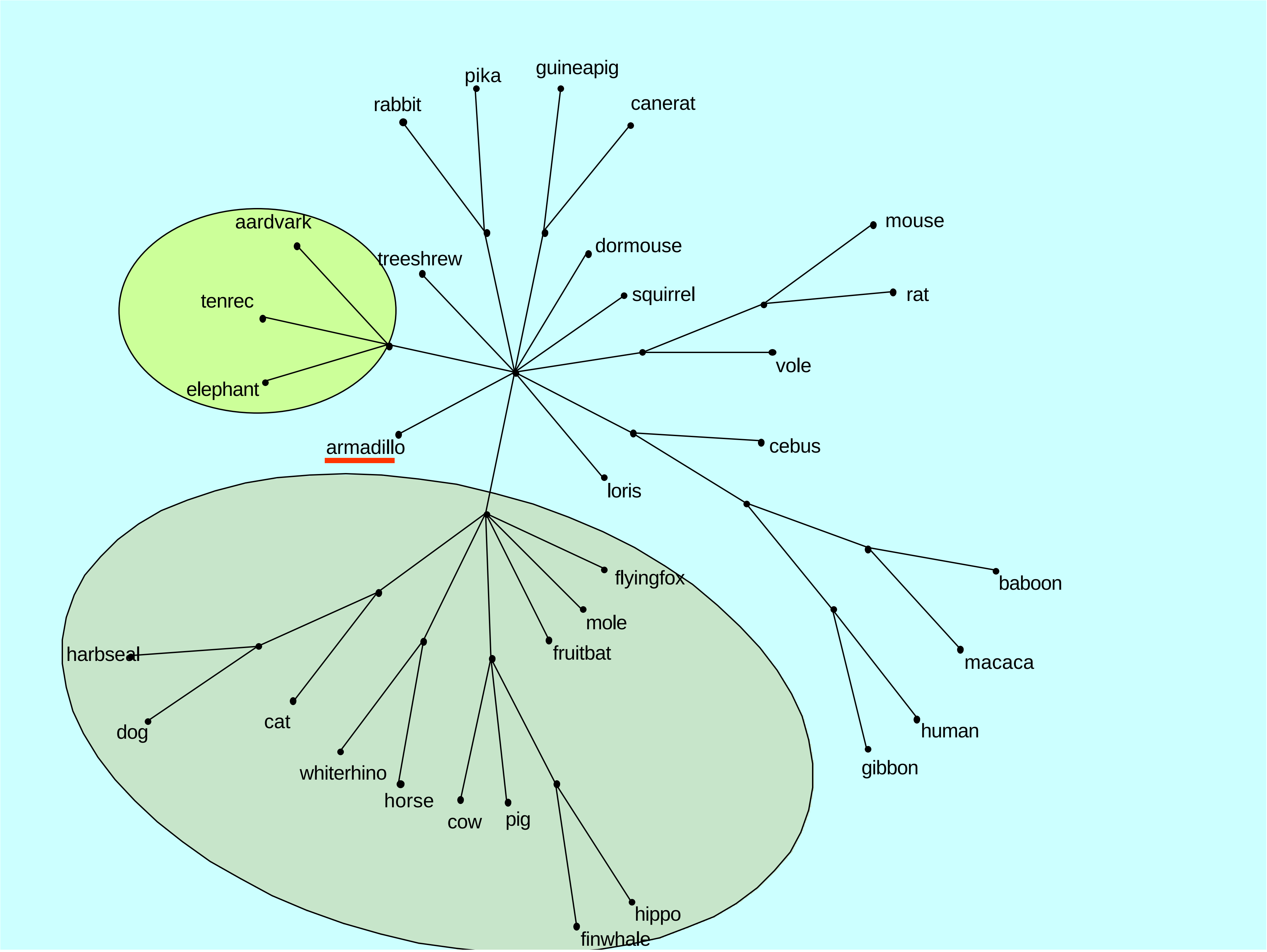
gene conversion/concerted evolution

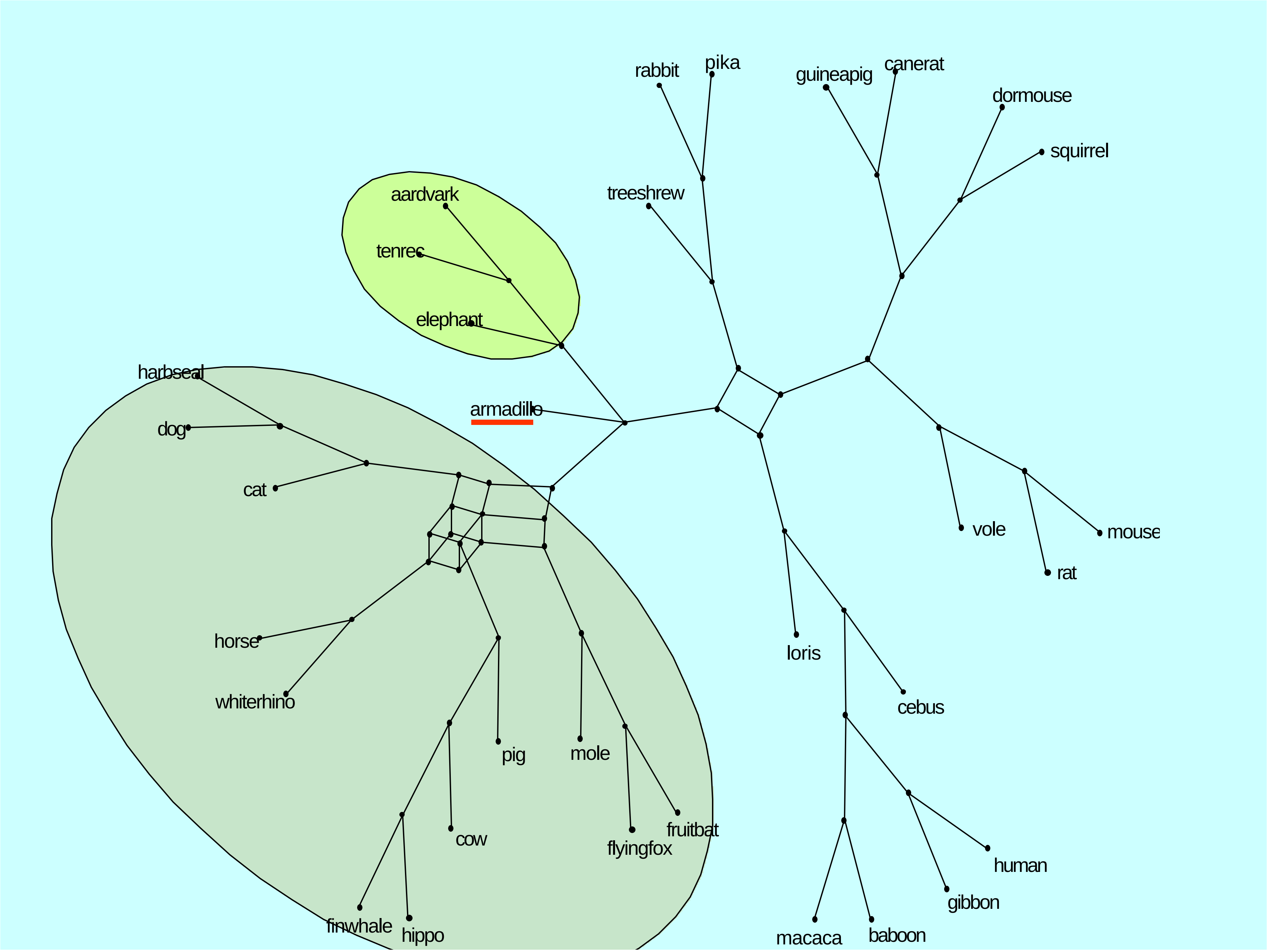
lateral transfer (incl. endosymbiosis)

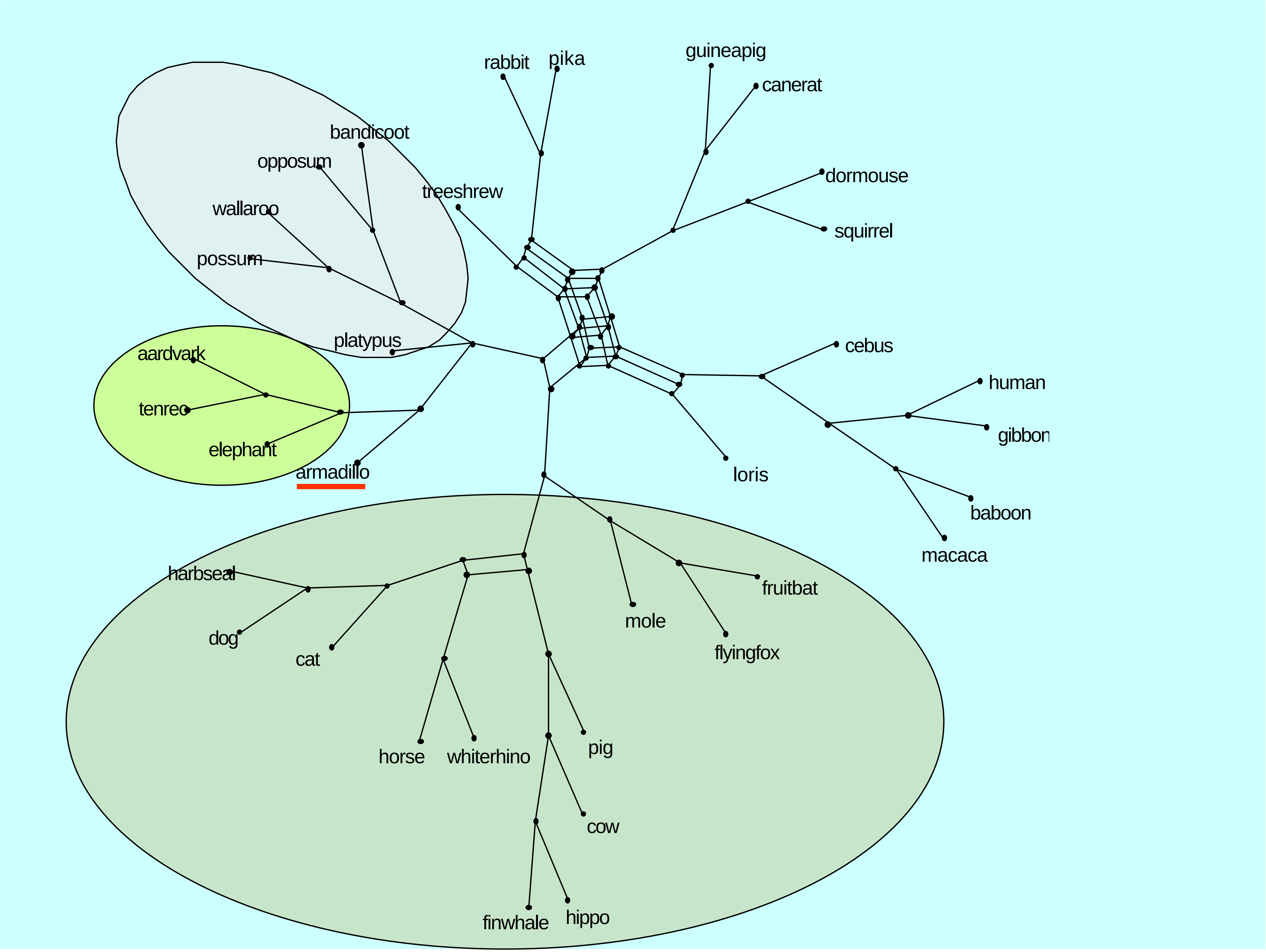
alternative (non-Steiner tree) models

(incorrect assumptions about the mechanism)

exploratory data analysis - EDA


pika
guineapig
rabbit
canerat
aardvark
mouse
dormouse
treeshrew
tenrec
squirrel
rat
vole
elephant
armadillo
cebus
loris
flyingfox
baboon
mole
harbseal
fruitbat
macaca
cat
dog
human
whiterhino
gibbon
horse
cow
pig
hippo
finwhale

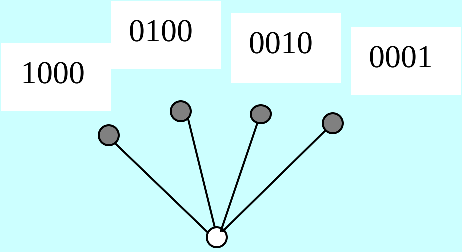

rabbit
pika
guineapig
canerat
dormouse
squirrel
treeshrew
aardvark
tenrec
elephant
harbseal
armadillo
dog
cat
vole
mouse
rat
loris
cebus
horse
whiterhino
pig
mole
cow
fruitbat
flyingfox
human
gibbon
finwhale
hippo
macaca
baboon

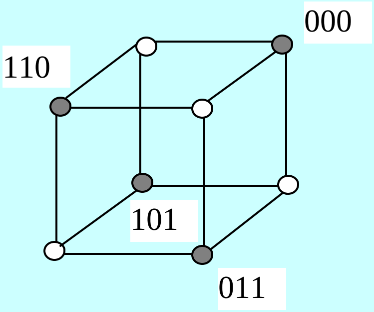

rabbit
pika
guineapig
canerat
dormouse
squirrel
bandicoot
opposum
treeshrew
wallaroo
possum
platypus
aardvark
tenrec
elephant
armadillo
cebus
human
gibbon
loris
baboon
macaca
harbseal
dog
cat
mole
fruitbat
flyingfox
horse
whiterhino
pig
cow
finwhale
hippo

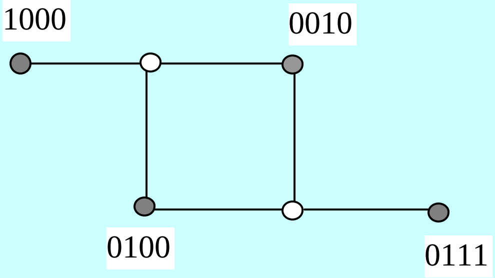
consensus methods lose information

	1	**2**	**3**
taxon A	1000	110	1000
taxon B	0100	101	0111
taxon C	0010	011	0100
taxon D	0001	000	0010

Three data sets from Bandelt et al (1995). The consensus of each would be a polytomy as for (a). However, the median network of each shows that each has a different data structure



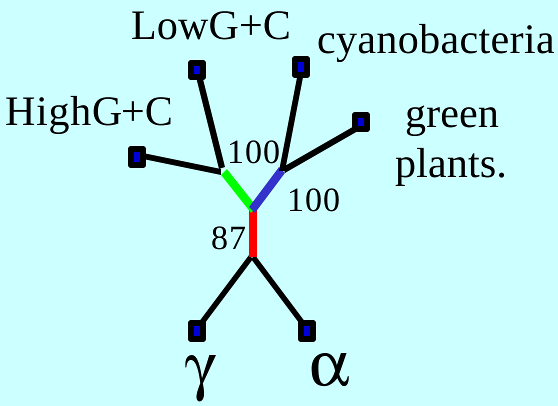




are sites always in the same rate class?

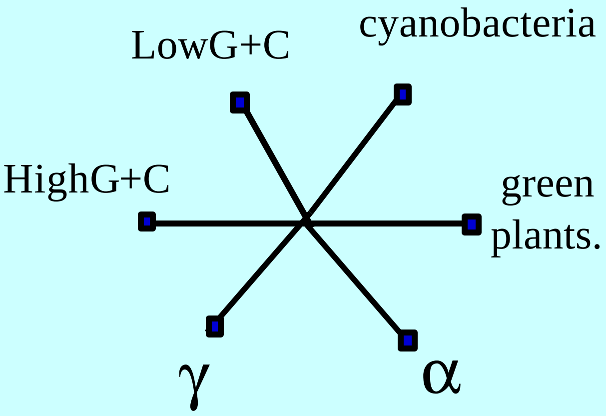
observed

NJ



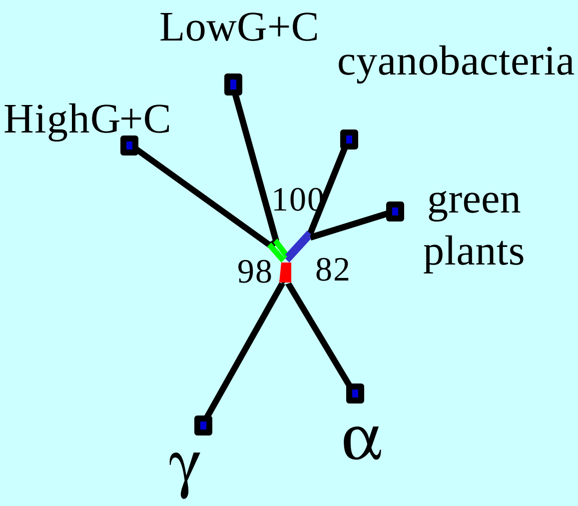
expected under i.i.d

Split graph/d_{cov}



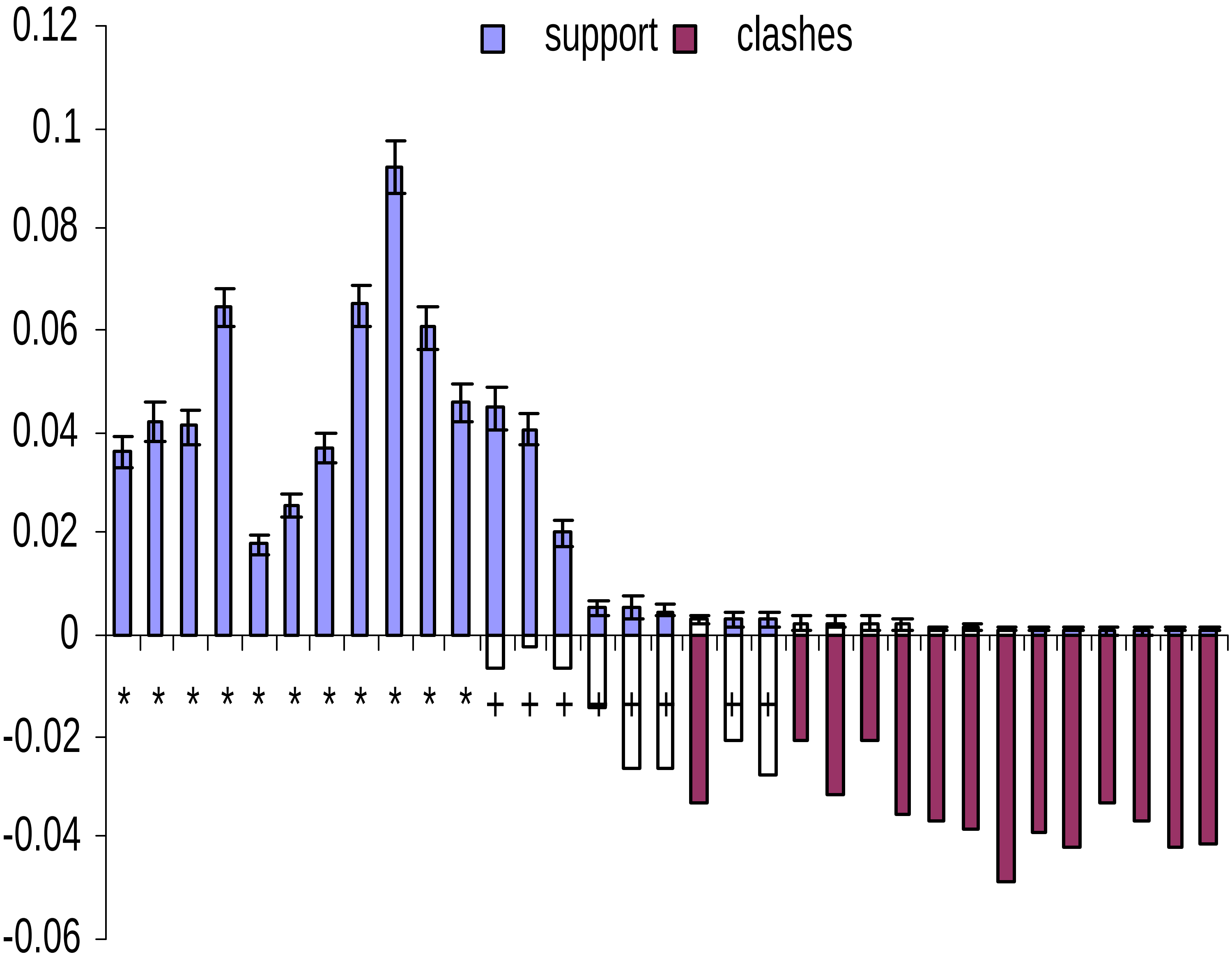
observed

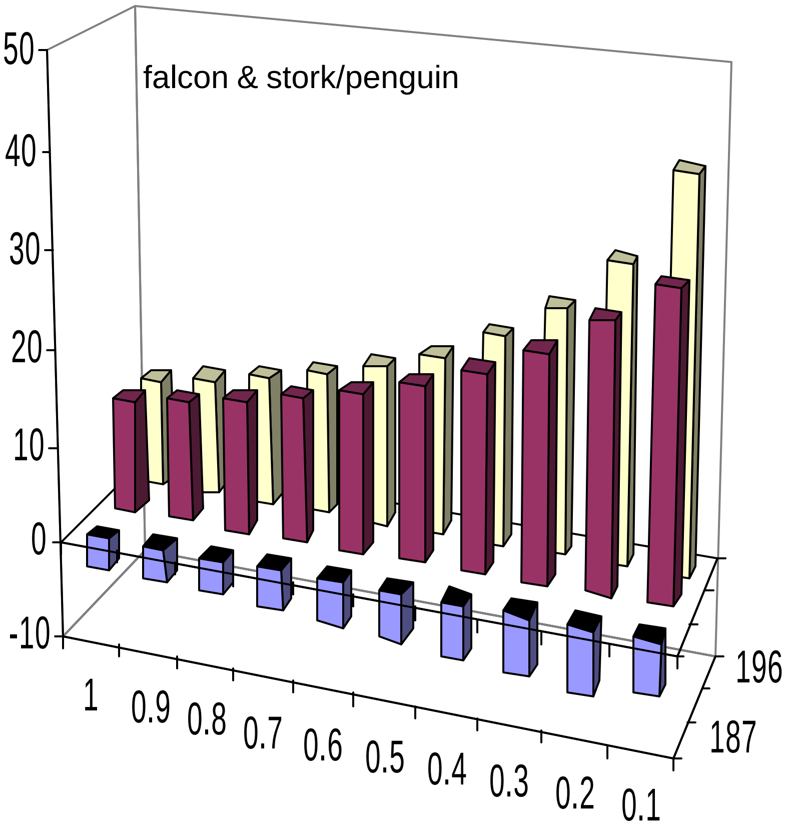
Split graph/d_{cov}

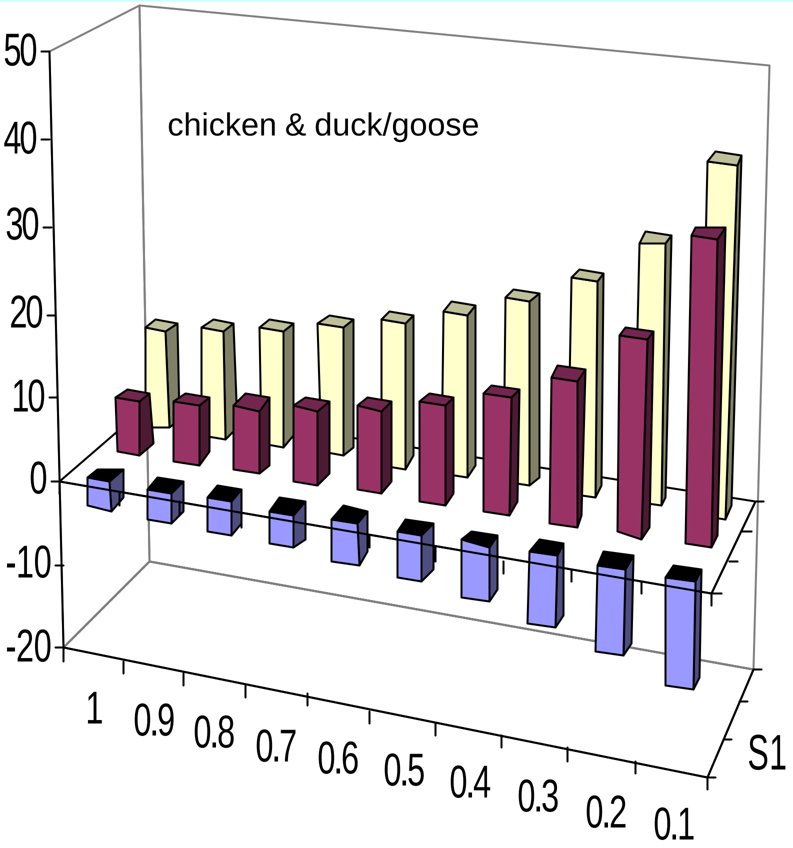


AtpB

Similar results for RecA, 16SrDNA, Hsp60, FtsZ
(2000) *MBE* 17, 835-838


support
clashes
0.12
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
* * * * * * * * * * *
+ + + + + + + +


falcon & stork/penguin
50
40
30
20
10
0
-10
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
196
187


chicken & duck/goose
50
40
30
20
10
0
-10
-20
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
S1

overview (contd)

trees and networks (with cycles)

spanning and phylogenetic (Steiner) trees

exploratory data analysis, visualization

multiscale

splits

compatibility/incompatibility

application - the MinMax squeeze

when MP and ML are equivalent

- for a single column of data (this leads to),
- there is no common mechanism shared by the characters,
- maximum evolutionary pathway likelihood(ML_{ep}) is equivalent to maximum parsimony,
- there are effectively infinite character states (the same mutation never occurs twice)

- abundant sequences?

overview

trees and networks (with cycles)

spanning and phylogenetic (Steiner) trees

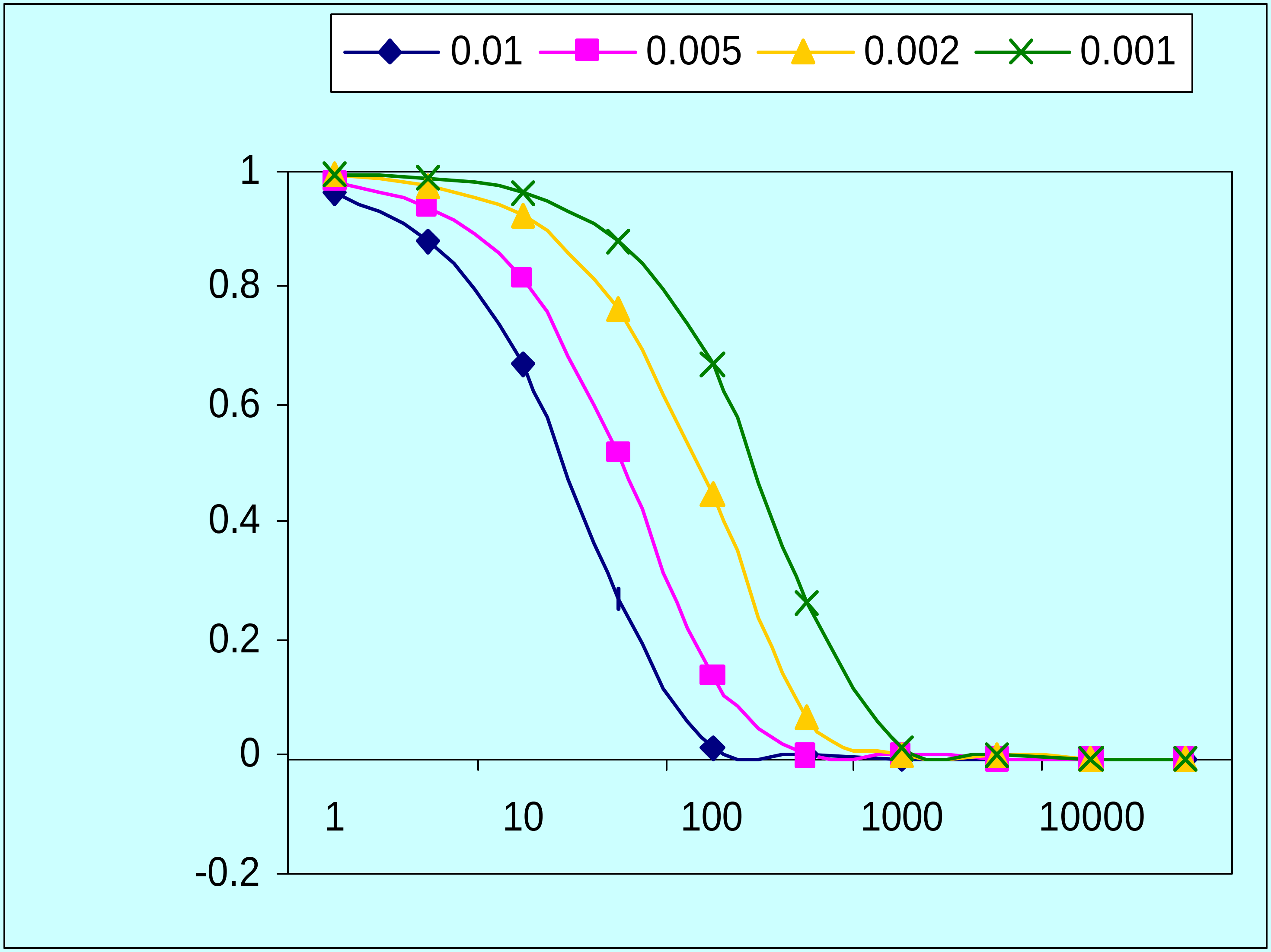
exploratory data analysis, visualization

multiscale

splits

compatibility/incompatibility

application - the MinMax squeeze

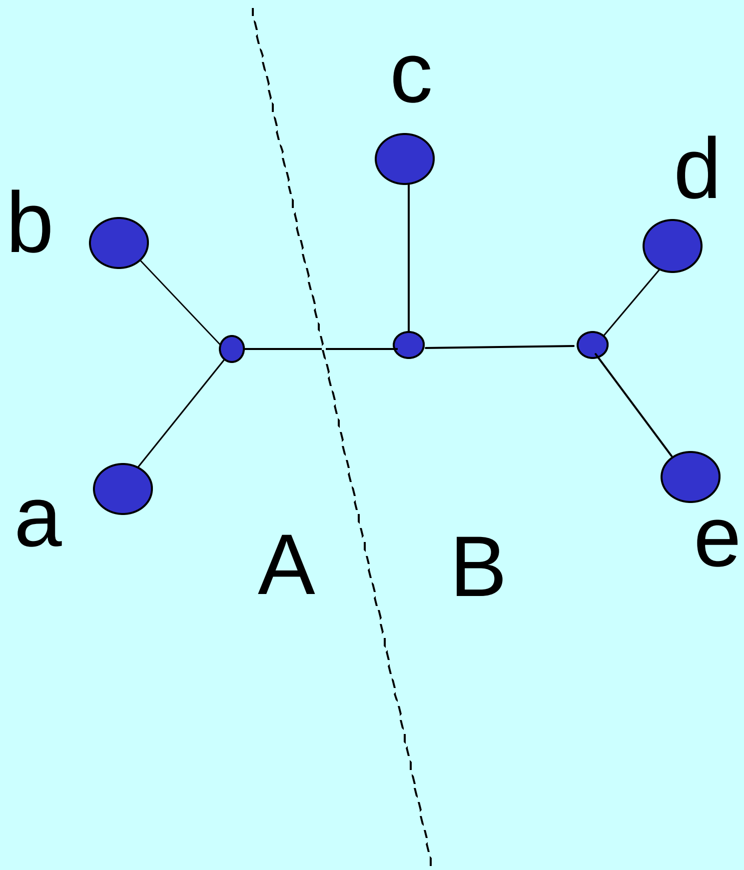


◆ 0.01
■ 0.005
▲ 0.002
✕ 0.001
1
0.8
0.6
0.4
0.2
0
-0.2
1
10
100
1000
10000

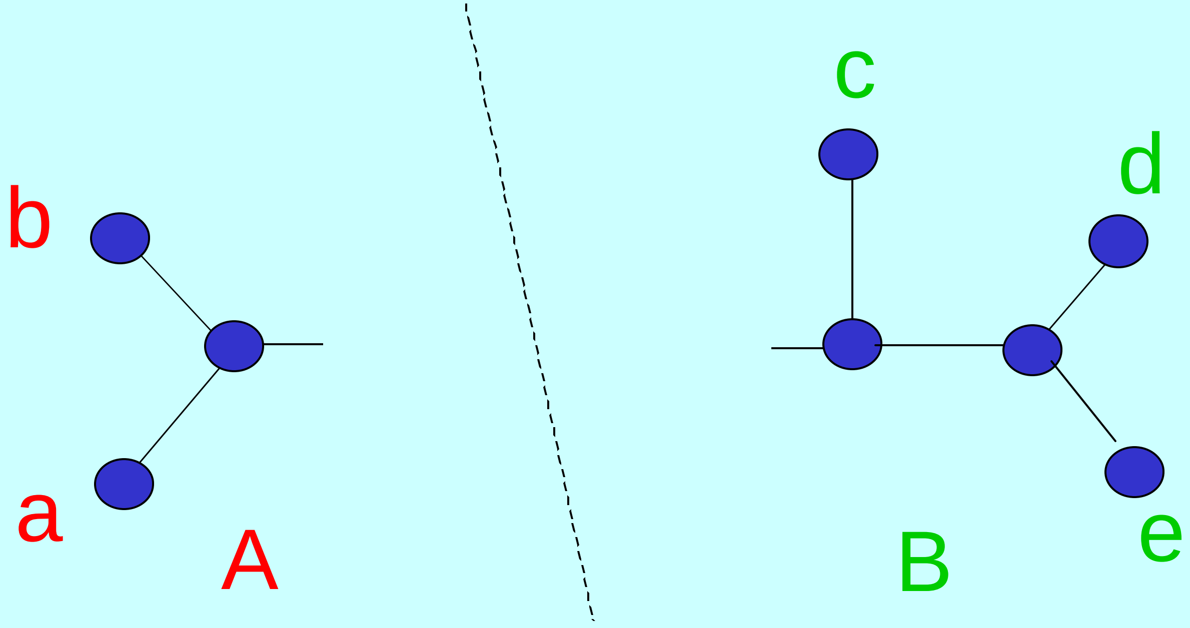
splits

- from trees and networks
- from data - sequences (character states)
- from data - distances

what is a split - from trees



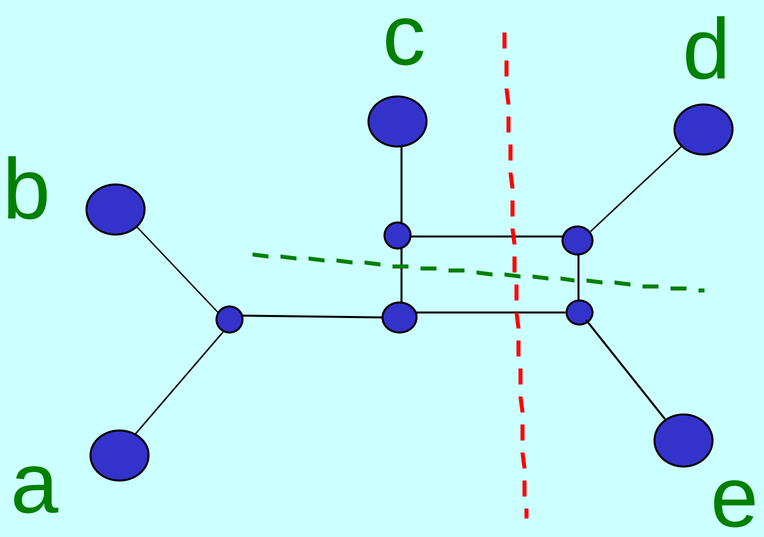
what is a split - from trees



A = {a,b} B = {c, d, e}

Denote by A | B.

splits in a network



{ a,b,c} {d,e}

{ a,b,e} {c,d}

summary of introduction

trees and networks (with cycles)

spanning and phylogenetic (Steiner) trees

exploratory data analysis, visualization

multiscale – generations to macroevolution

splits – from trees, characters or distances

compatibility/incompatibility

application - the MinMax squeeze