Phylogenomic inference with IQ-TREE

Stephen Crotty

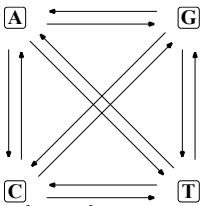
Cesky Krumlov, Czech Republic, 24 Jan 2019

Typical phylogenetic analysis

Sequence alignment

```
CACCTGTCGT   ----------   ----------   TCTGGTGCAG
CAGCTGTCGT   GCTCTTTCTG   TTGAGCCTGG   TCTGGTGCAG
CAGCTGCCGT   GTTTTCTCTG   TTGAGCCTGG   TCTGGTACAG
CAGCTGCCGC   GTTCTCTCCG   ----------   TCTGGTGCAA
CTCCTGCCGG   GTGCTCTCAG   ----------   ----------
CTCCTGCCGG   ----------   CTGAGCCGGG   TCTGGTGCAG
CTCTTGCCGG   ----------   CTGAGCCTTG   ----------
```

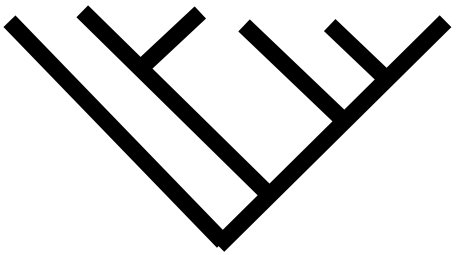
Model selection



Substitution models

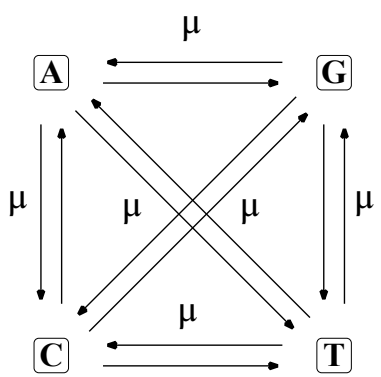
Tree search

Bootstrap

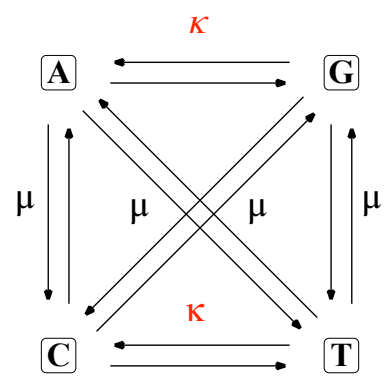


Phylogeny

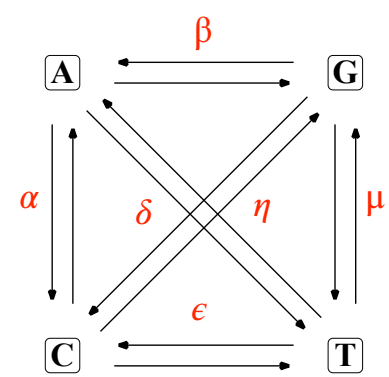
Models of sequence evolution



JC
(1969)



HKY
(1985)



GTR
(1986)

Rate heterogeneity: sites evolve at different rates - some slow, some fast.

Rate model	Explanation
+I	Some sites are *invariable* (zero rate), e.g. due to selective force.
+G	Site rates follow a *Gamma* distribution.
+I+G	Some sites are invariable, the rest follow a Gamma distribution.
+R	Sites fall into several categories from slow to fast rates. No assumption of rate distribution (free-rate model).

A model = substitution model + rate heterogeneity, e.g. "GTR+G"

Model selection

JC
JC+G
JC+I
JC+I+G
HKY
...
GTR+I+G

Which model is best?

Problem:
More complex models always have higher *likelihood* than simpler models!

Solution: Penalize a model *M* by the number of its parameters (*k*)

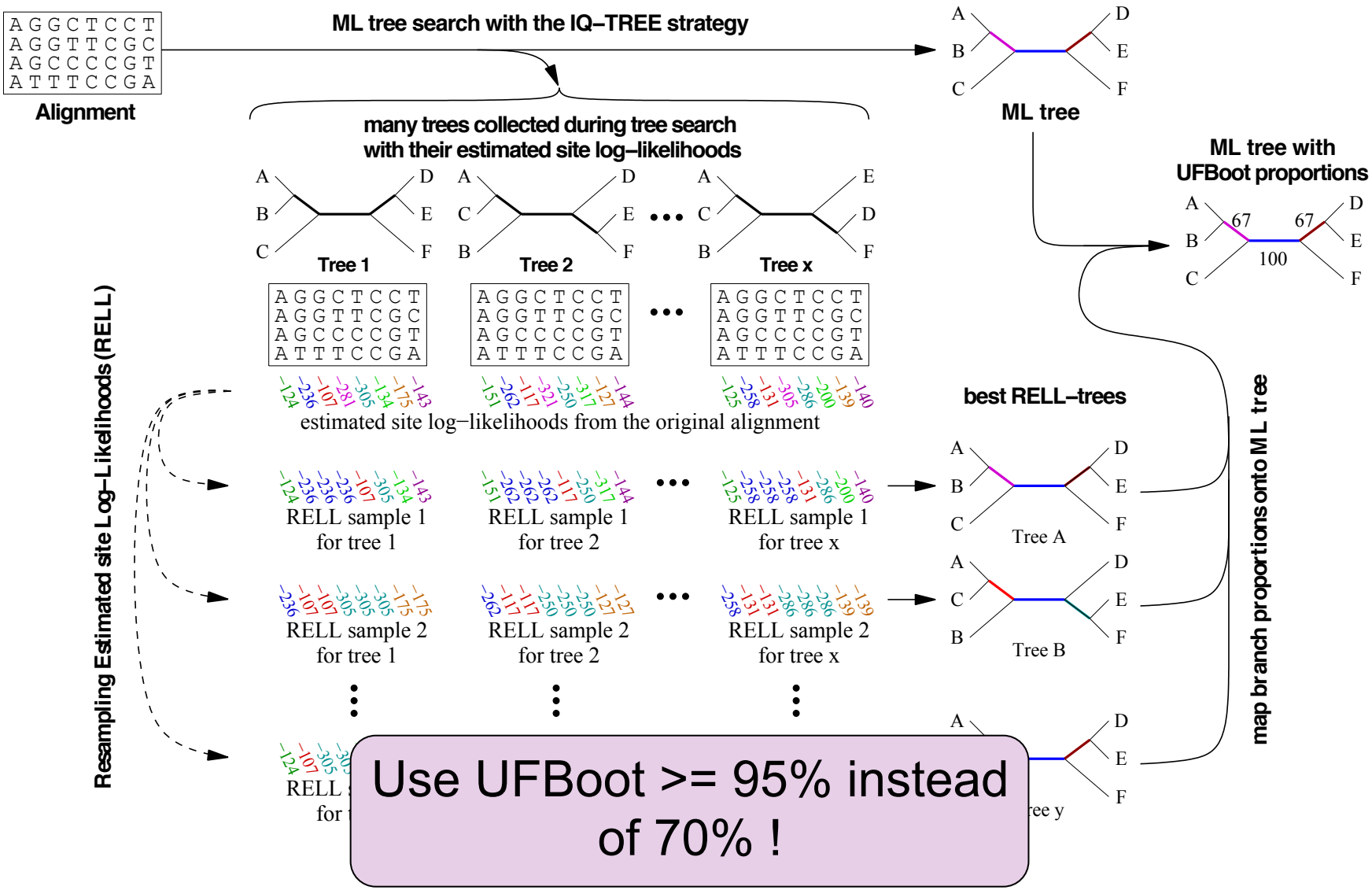
1. Akaike information criterion (AIC): $AIC = 2k - 2ln(L(D|T,M))\,.$

2. Bayesian information criterion (BIC): $BIC = ln(n)k - 2ln(L(D|T,M)),$

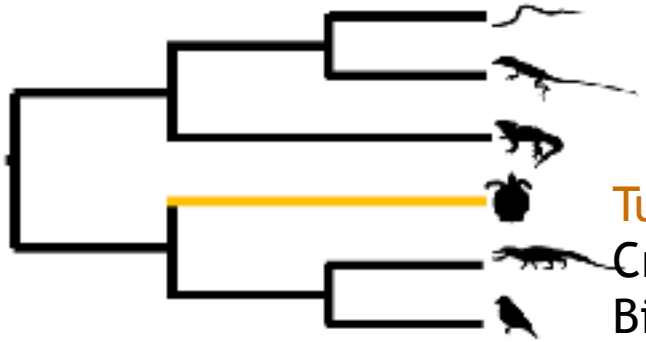
   where *n* is the number of alignment sites.

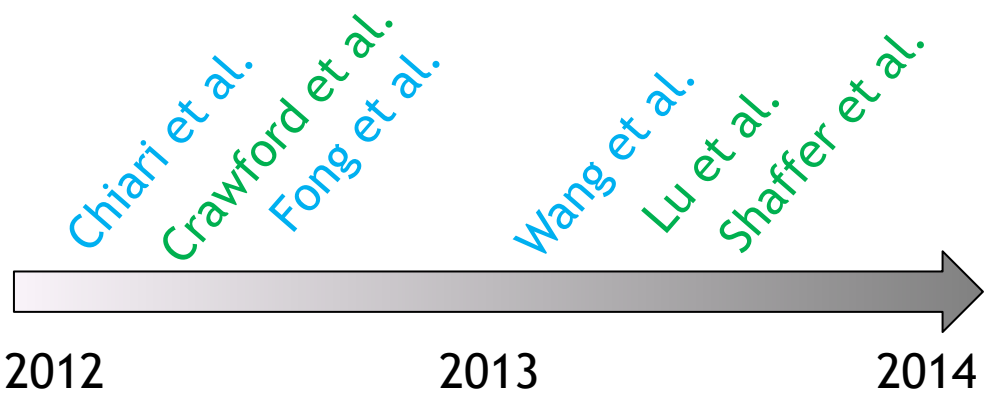
Select the model with smallest AIC or BIC score.

UFBoot: Ultrafast bootstrap approximation



Practical: Where is Turtle in the tree?





Different studies led to different trees!

Dataset: 16 species, 29 genes, 20,820 bp (a subset of Chiari et al. 2012)

Thanks Jeremy Brown

Exercises

1. Input data
2. Inferring the first phylogeny
3. Applying partition model
4. Choosing the best partitioning scheme
5. Applying a mixture model - GHOST
6. Tree topology tests
7. Concordance factors
8. Identifying most influential genes

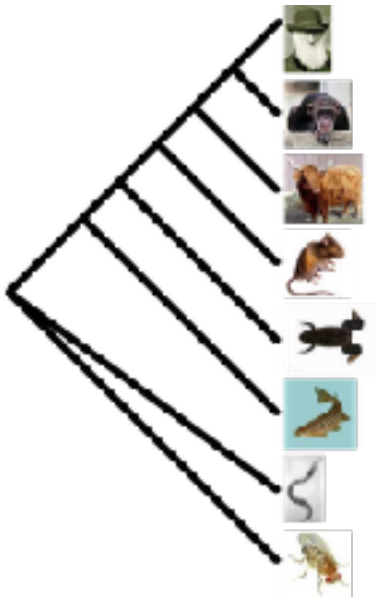
http://www.iqtree.org/workshop/ck2019

Genome-scale data

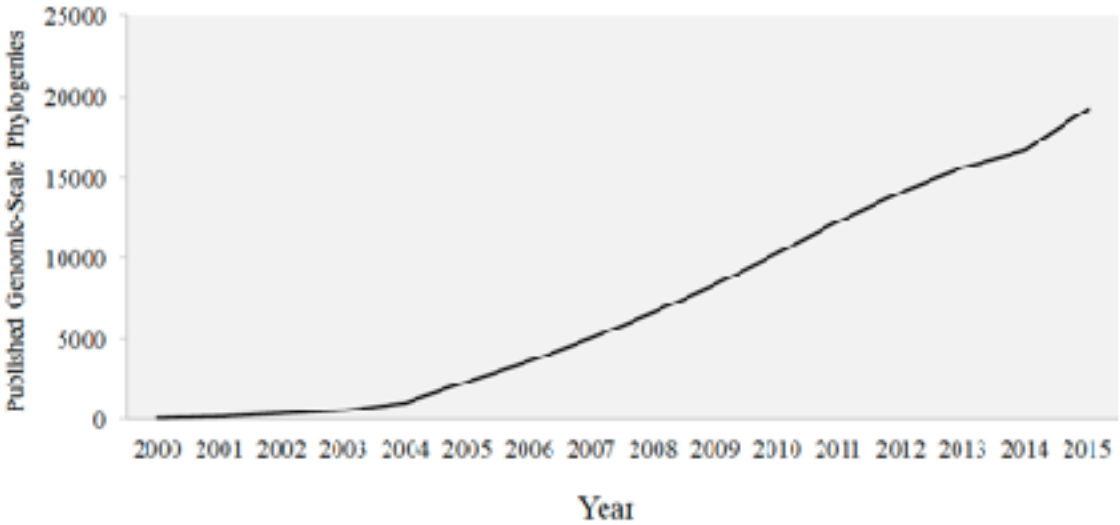
Supermatrix

Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

Phylogenomic Inference



Species tree of life



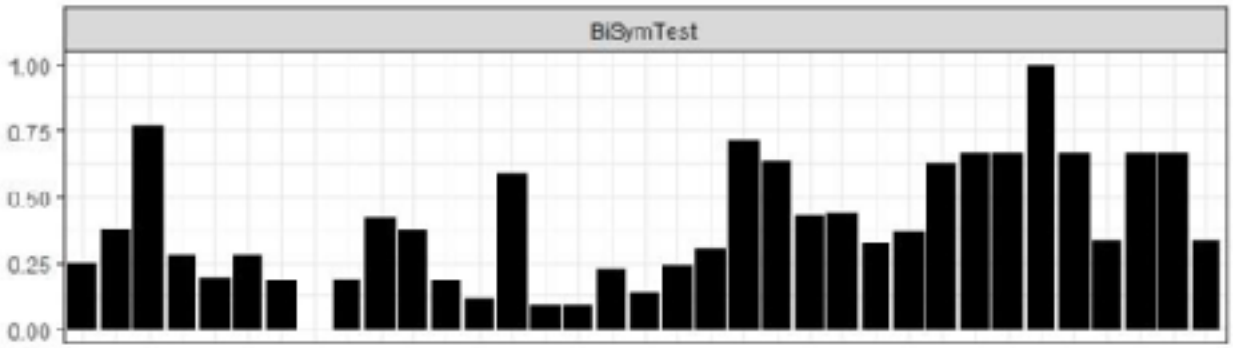
30 days of computation and 280 GB RAM for an insect data set (CSIRO)!

Exercises

1. Input data
2. Inferring the first phylogeny
3. Applying partition model
4. Choosing the best partitioning scheme
5. Applying a mixture model - GHOST
6. Tree topology tests
7. Concordance factors
8. Identifying most influential genes

http://www.iqtree.org/workshop/ck2019

“Data-model gap” is increasing!



Level of model violations in 35 phylogenomic datasets (https://doi.org/10.1101/460121)

1. Resulting trees tend to be biased towards the genes that violated model assumptions.
2. Bootstrap supports tend to 100% as #genes increases.

Model violation → Systematic bias

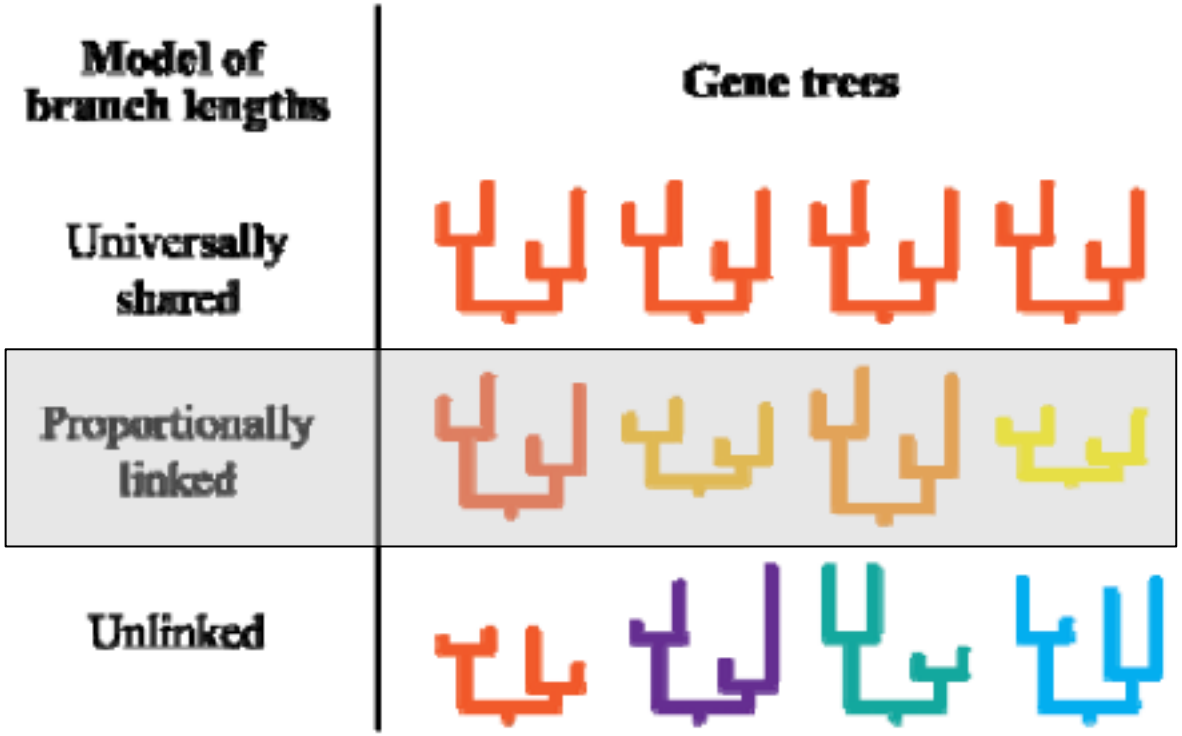
1. Remove “bad” loci
2. Use more realistic models

Partition model

Supermatrix

Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

Substitution models: JC HKY+G ..... GTR+G



Recommended for typical analysis, confirmed by Dunchene et al. (2018) https://doi.org/10.1101/467449

Exercises

1. Input data
2. Inferring the first phylogeny
3. Applying partition model
4. Choosing the best partitioning scheme
5. Applying a mixture model - GHOST
6. Tree topology tests
7. Concordance factors
8. Identifying most influential genes

http://www.iqtree.org/workshop/ck2019

How to reduce potential model overfitting?

Supermatrix

Gene 1	Gene 2	……	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

PartitionFinder

Gene 1+2	……	Gene 200+1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

PartitionFinder algorithm (Lanfear et al. 2012):

1. Evaluate to merge all pairs of genes.
2. Choose the pair with the best score.
3. If score improves, merge two genes and repeat steps 1-3.
4. Otherwise, stop.

Relaxed clustering algorithm (Lanfear et al. 2014):

In step 1: only examine the top k% of most "promising" pairs.

Substitution models: HKY …… GTR+G

Exercises

1. Input data
2. Inferring the first phylogeny
3. Applying partition model
4. Choosing the best partitioning scheme
5. Applying a mixture model - GHOST
6. Tree topology tests
7. Concordance factors
8. Identifying most influential genes

http://www.iqtree.org/workshop/ck2019

Mixture model - GHOST

Partition Model

Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

$$\mathscr{L}(s_{ik} \mid \mathscr{M}) = \sum_{j=1}^{m} \delta_{kj} \mathscr{L}_{ij}(s_{ik} \mid M_j, T, \lambda_j)$$

Mixture Model

Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------

$$\mathscr{L}(s_i \mid \mathscr{M}) = \sum_{j=1}^{m} w_j \mathscr{L}_{ij}(s_i \mid M_j, T, \lambda_j)$$

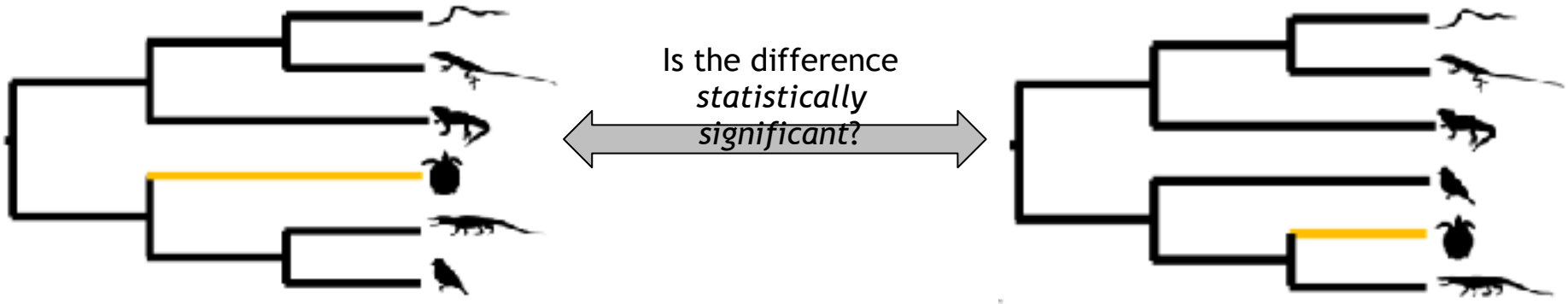
For full details of the GHOST model see Crotty et al. (2019) https://doi.org/10.1101/174789

Exercises

1. Input data
2. Inferring the first phylogeny
3. Applying partition model
4. Choosing the best partitioning scheme
5. Applying a mixture model - GHOST
6. Tree topology tests
7. Concordance factors
8. Identifying most influential genes

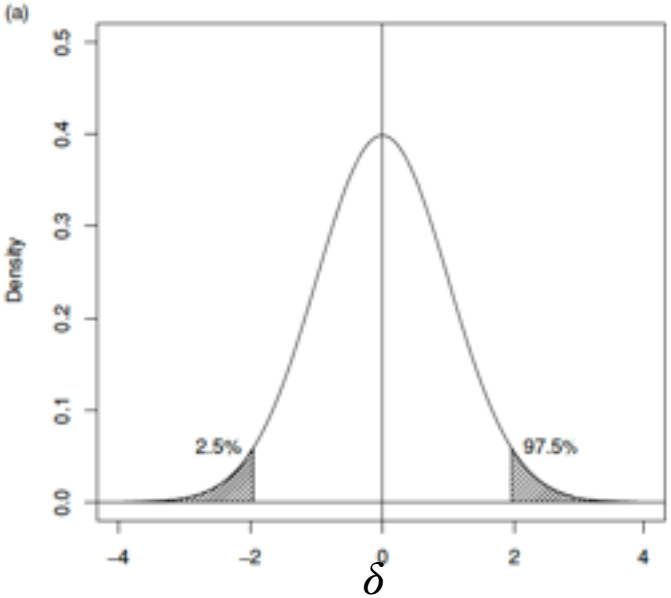
http://www.iqtree.org/workshop/ck2019

Tree topology tests



Testing two trees (Kishino & Hasegawa, 1989):

1. Statistic: $\Delta = ln(L(D)\,|\,T_1)) - ln(L(D\,|\,T_2))$
2. Generate distribution of from many "random" data (e.g. by 1000 bootstrap resampling).
3. Compare the statistic between original and random data to obtain *p-value*.
4. If p-value < 0.05: YES! two trees are significantly different.
5. If p-value >= 0.05: NO! they are not.



Exercises

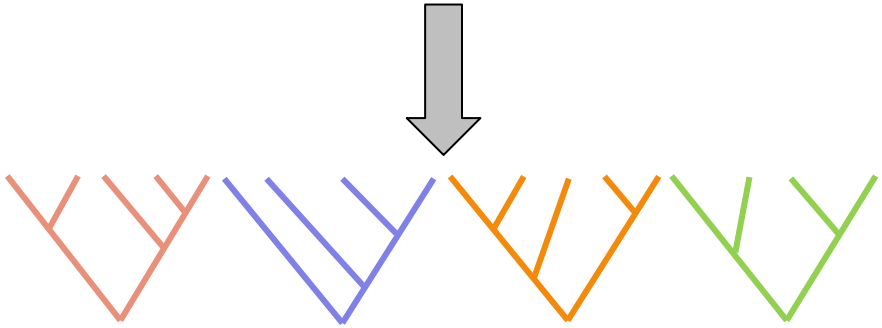
1. Input data
2. Inferring the first phylogeny
3. Applying partition model
4. Choosing the best partitioning scheme
5. Applying a mixture model - GHOST
6. Tree topology tests
7. Concordance factors
8. Identifying most influential genes

http://www.iqtree.org/workshop/ck2019

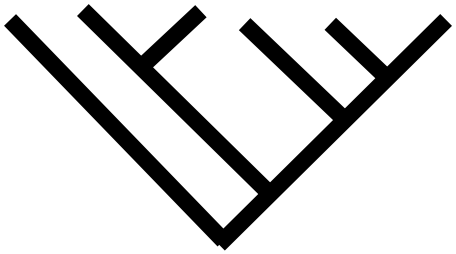
Inferring species tree from gene trees

Supermatrix

Gene 1	Gene 2	.....	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG	----------	CTGAGCCTTG	----------



Gene tree 1 Gene tree 2 ..... Gene tree 1,000

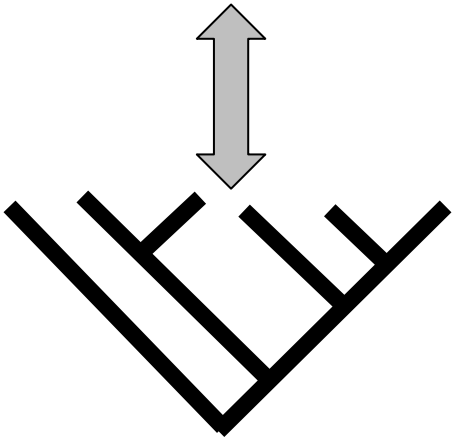


Species tree

Concordance factor

https://doi.org/10.1101/487801

Gene tree 1 Gene tree 2 ..... Gene tree 1,000



Species tree

Gene concordance factor (gCF): How often each branch of species tree is found among the gene trees?

Site concordance factor (sCF): How often each branch of species tree is supported by the alignment sites?

```
CACCTGTCGT   ----------   ----------   TCTGGTGCAG
CAGCTGTCGT   GCTCTTTCTG   TTGAGCCTGG   TCTGGTGCAG
CAGCTGCCGT   GTTTTCTCTG   TTGAGCCTGG   TCTGGTACAG
CAGCTGCCGC   GTTCTCTCCG   ----------   TCTGGTGCAA
CTCCTGCCGG   GTGCTCTCAG   ----------   ----------
CTCCTGCCGG   ----------   CTGAGCCGGG   TCTGGTGCAG
CTCTTGCCGG   ----------   CTGAGCCTTG   ----------
```

Sequence alignment

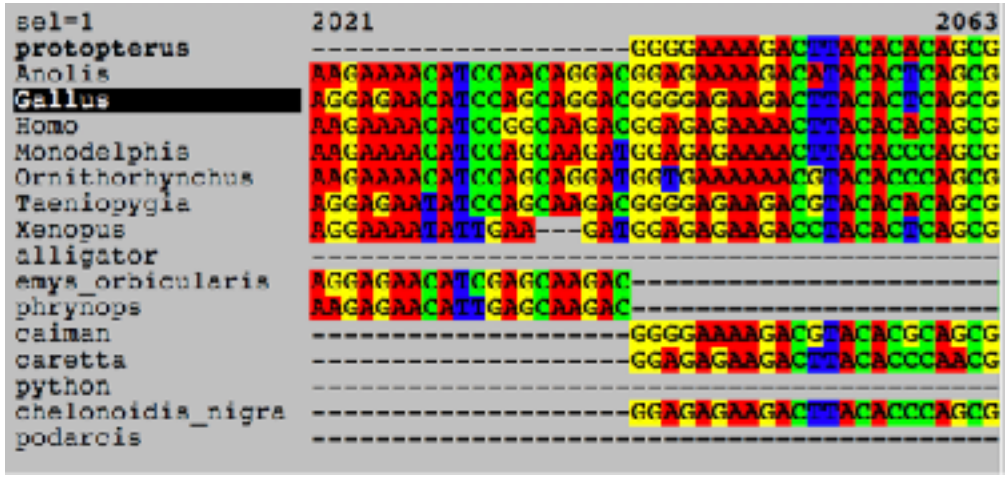
- gCF and sCF captures dis(agreement) in the data, whereas boostrap doesn't!
- gCF and sCF are useful when bootstrap supports reach 100%.

Exercises

1. Input data
2. Inferring the first phylogeny
3. Applying partition model
4. Choosing the best partitioning scheme
5. Applying a mixture model - GHOST
6. Tree topology tests
7. Concordance factors
8. Identifying most influential genes

http://www.iqtree.org/workshop/ck2019

Solutions - Input data



Gene boundaries are obvious within the dataset, as most genes are not present for all species.

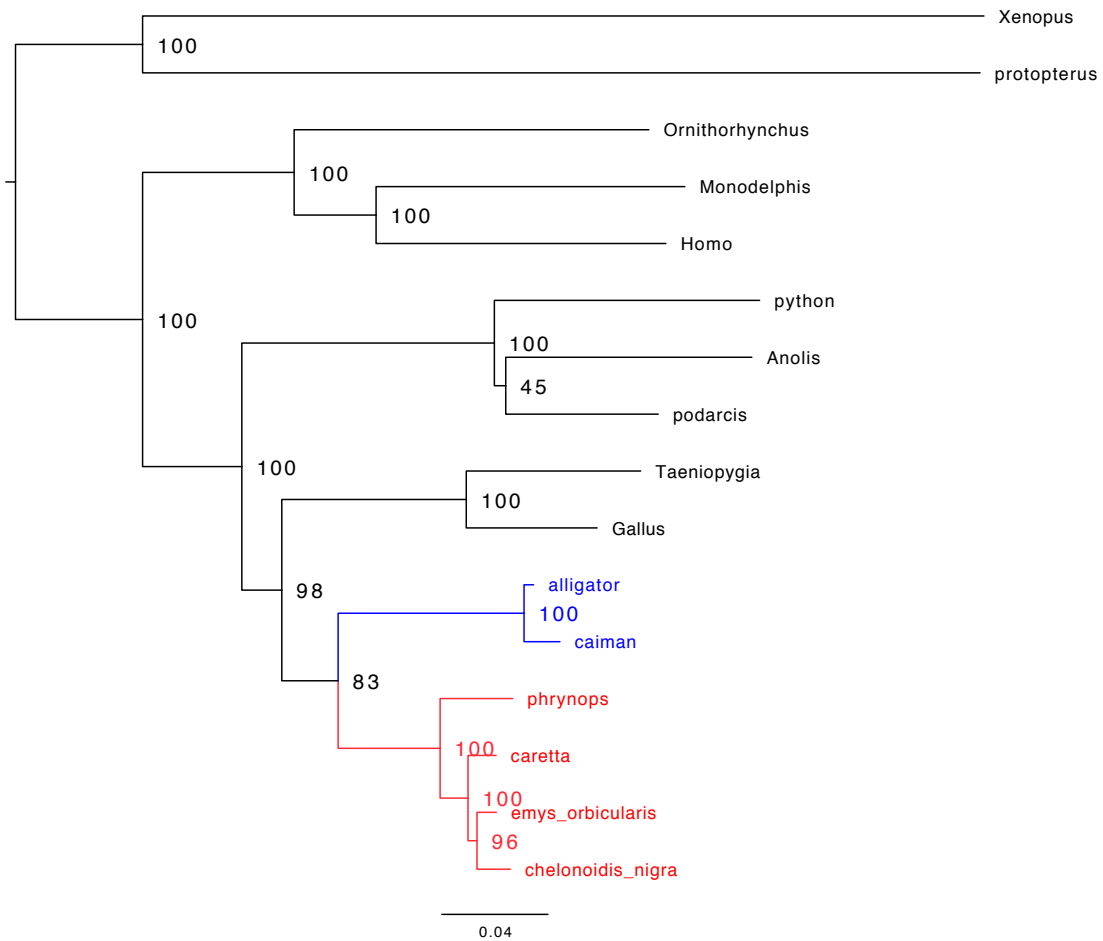
The four turtle (emys, phrynops, caretta, chelonodis)and two crocodile (caiman, alligator) species have much more missing data than most other species in the alignment. This might make their position in the tree more difficult to resolve.

Solutions - Inferring the first phylogeny

The best-fit model found by ModelFinder was GTR+F+R3.
This means:

- the GTR model of sequence evolution
- base frequencies calculated empirically from the alignment (as opposed to inferred under ML)
- three categories of rate heterogeneity, with rates and weights inferred by ML, not constrained to the Gamma distribution.

Solutions - Inferring the first phylogeny



Tree inferred under GTR+F+R3 without partitioning the alignment. Turtles (red) are sister to crocodiles (blue), in contradiction with the published tree, in which turtles are sister to archosaurs (crocodiles and birds).

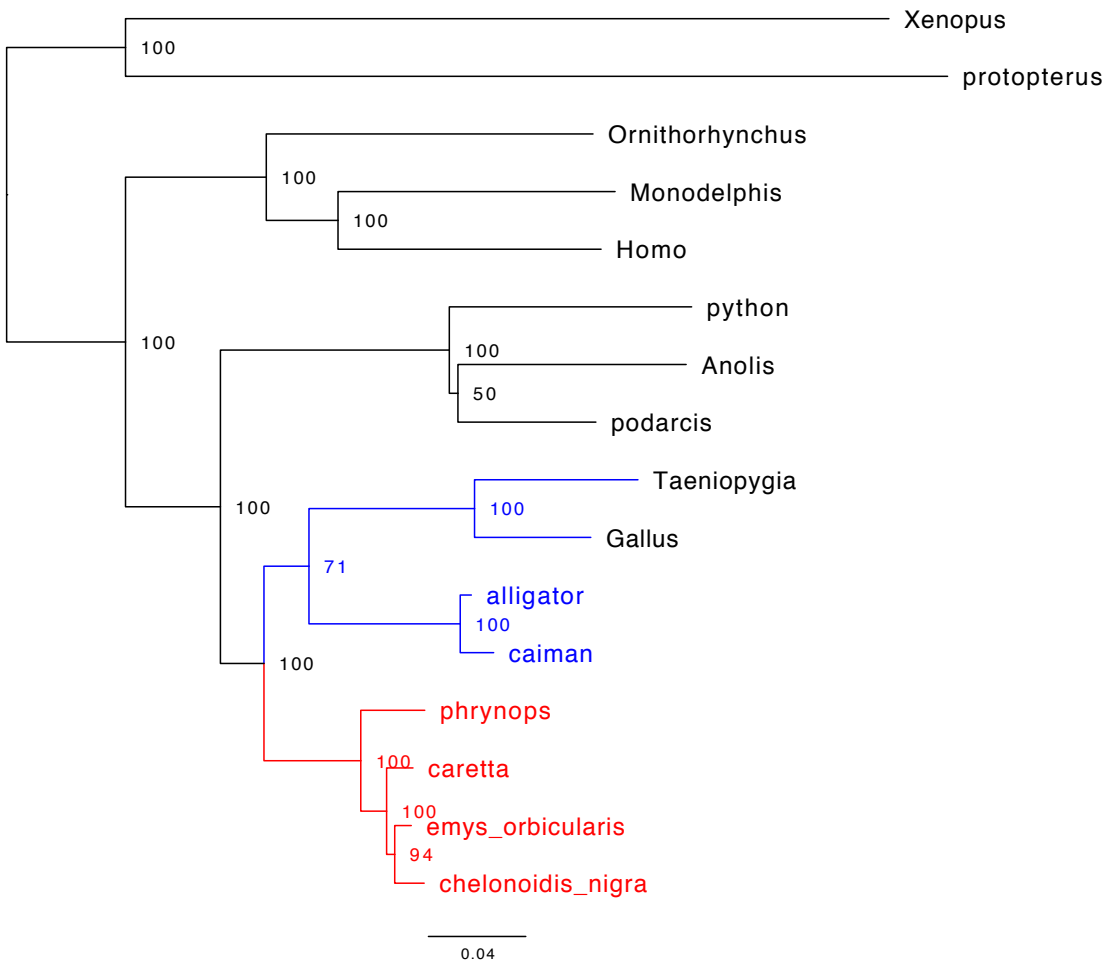
Solutions - Applying partition model

- The slowest evolving gene is the 10th gene, with a rate of 0.4683.
- The fastest evolving gene is the 18th gene, with a rate of 1.8421.

- The BIC of the non-partitioned model is 232837.7889.
- The BIC of the partitioned model is 233126.4205.

Even though the partitioned model has a higher likelihood than the non-partitioned model, the non-partitioned model has a smaller (better) BIC, and on that basis should be preferred. This is because the partitioned model has 221 free parameters, compared to just 41 for the non-partitioned model.

Solutions - Applying partition model



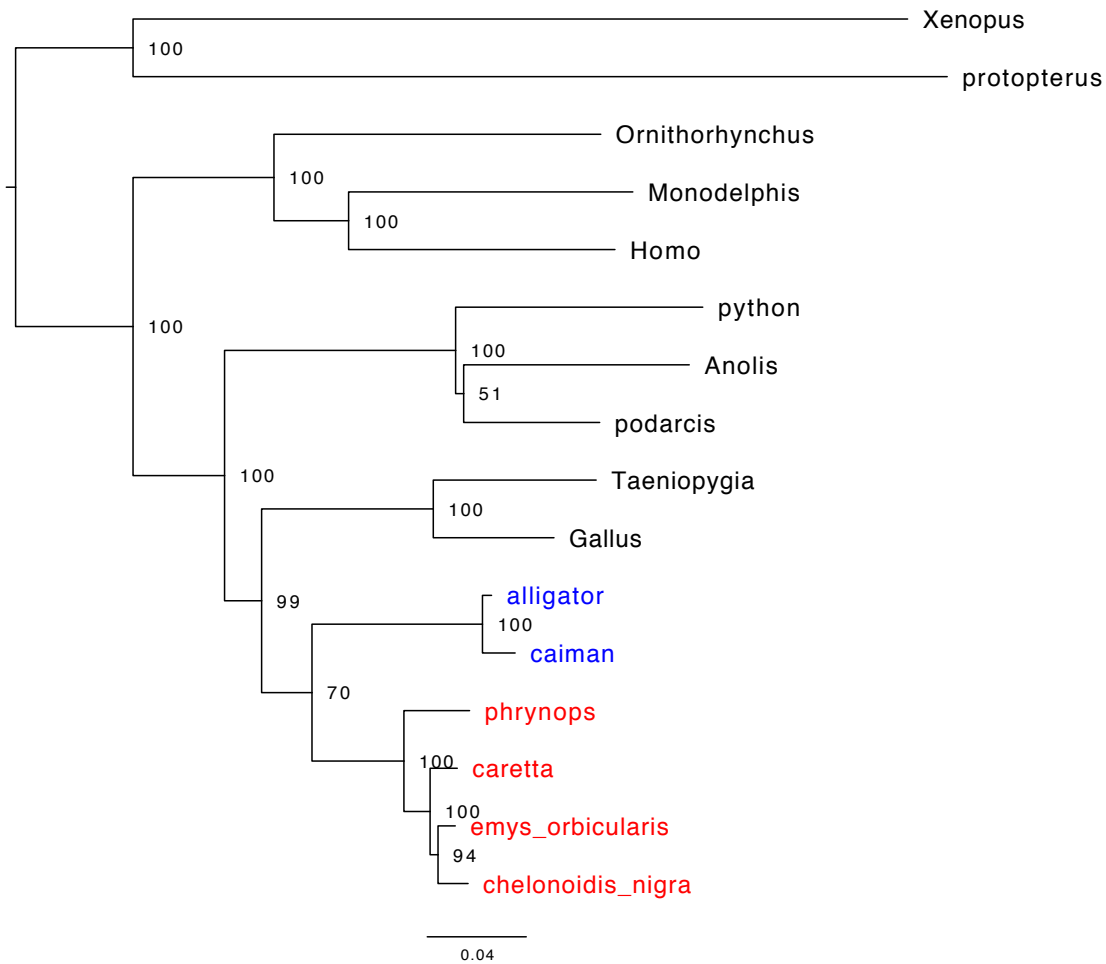
Tree inferred under the partition model. Turtles (red) are sister to archosaurs (blue), concurring with the published tree.

Solutions - Choosing the best partitioning scheme

- The BIC of the non-partitioned model is 232837.7889.
- The BIC of the partitioned model is 233126.4205.
- The BIC of the best partition model is 232401.3940

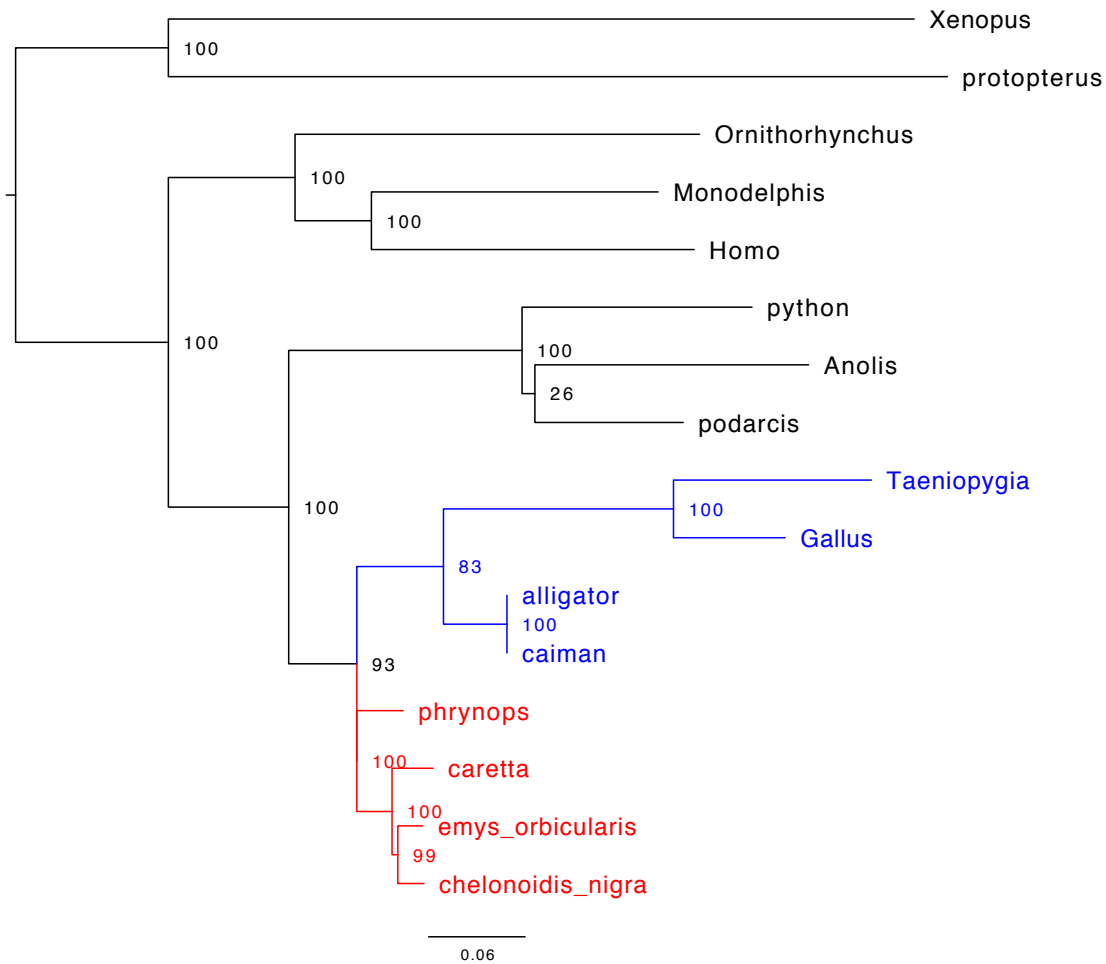
By merging similar genes, we have reduced the number of partitions from 29 to 10. This has reduced the number of parameters in the model from 221 to 106 and consequently, the best partition scheme now has the lowest BIC score of the three models considered so far.

Solutions - Applying partition model



Tree inferred under best partition scheme. This topology agrees with that inferred by the non-partitioned model, and conflicts with the published tree. The bootstrap support for the conflicting branch has fallen from 83 to 70.

Solutions - Applying a mixture model - GHOST

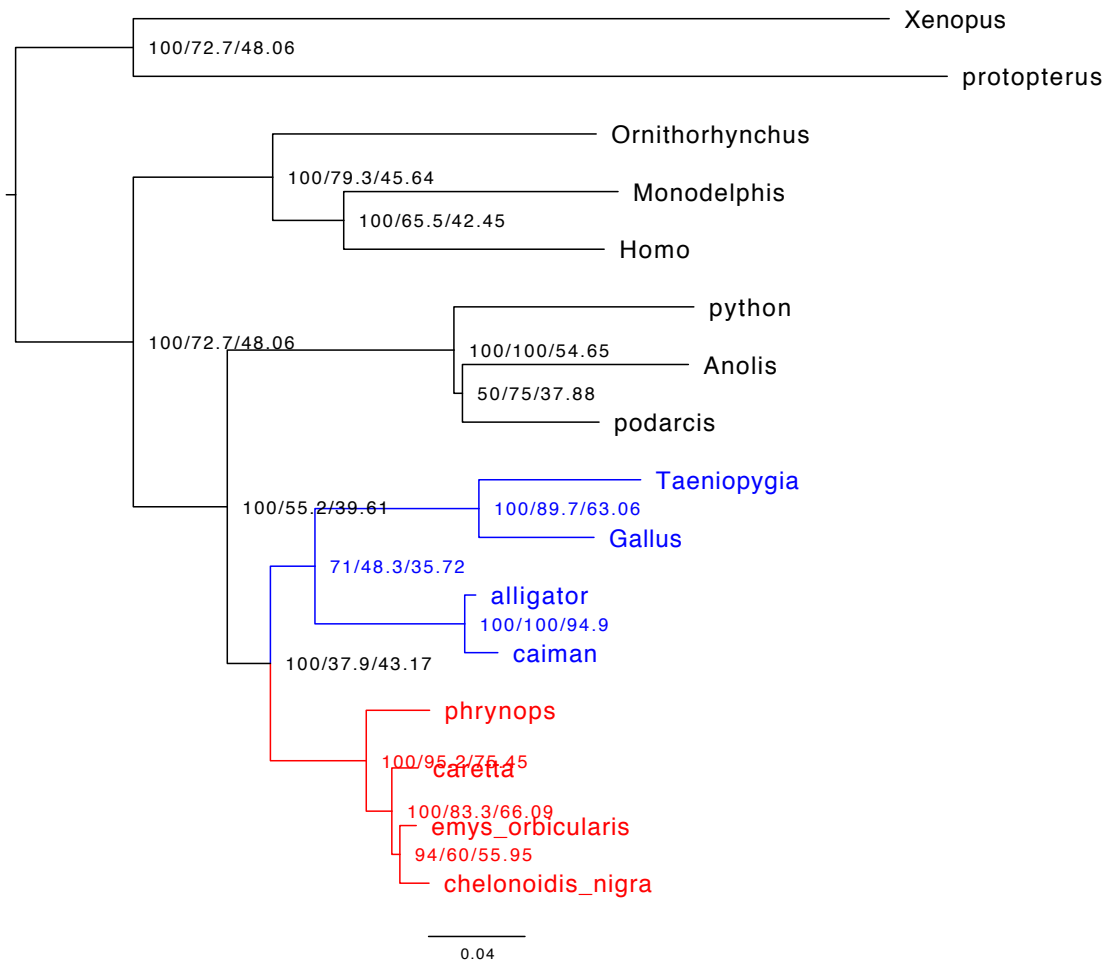


Tree inferred under GHOST model. This topology agrees with the inferred tree under the full partition model, and the published tree. Compared to the full partition model, the bootstrap support for the contentious branch has increased from 71 to 93.

Solutions - Tree topology tests

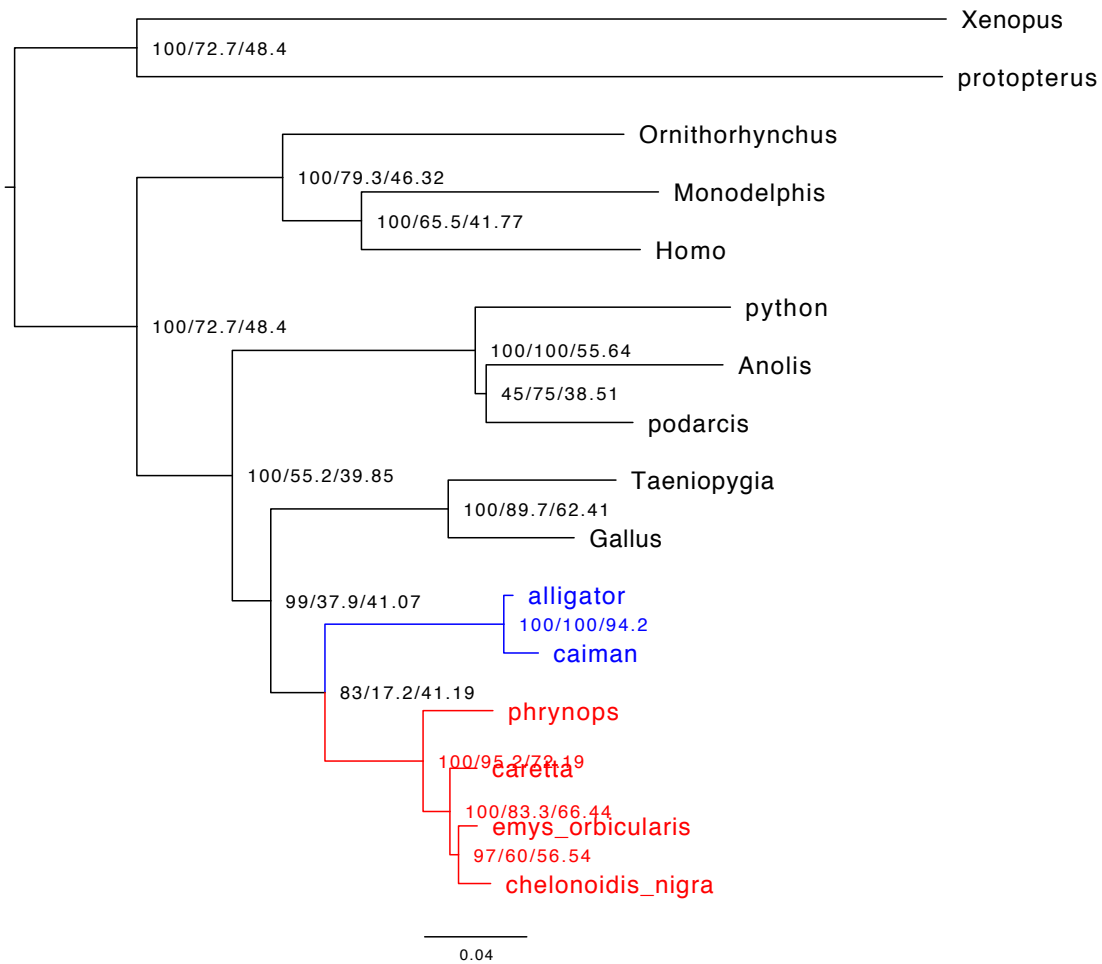
```
USER TREES
----------

See turtle.test.trees for trees with branch lengths.

Tree      logL    deltaL  bp-RELL    p-KH     p-SH       c-ELW
--------------------------------------------------------------
  1 -115476.8396  6.7446   0.399 +  0.394 +  0.394 +     0.401 +
  2  -115470.095       0   0.601 +  0.606 +      1 +     0.599 +

deltaL  : logL difference from the maximal logl in the set.
bp-RELL : bootstrap proportion using RELL method (Kishino et al. 1990).
p-KH    : p-value of one sided Kishino-Hasegawa test (1989).
p-SH    : p-value of Shimodaira-Hasegawa test (2000).
c-ELW   : Expected Likelihood Weight (Strimmer & Rambaut 2002).

Plus signs denote the 95% confidence sets.
Minus signs denote significant exclusion.
All tests performed 1000 resamplings using the RELL method.
```

Solutions - Concordance factors



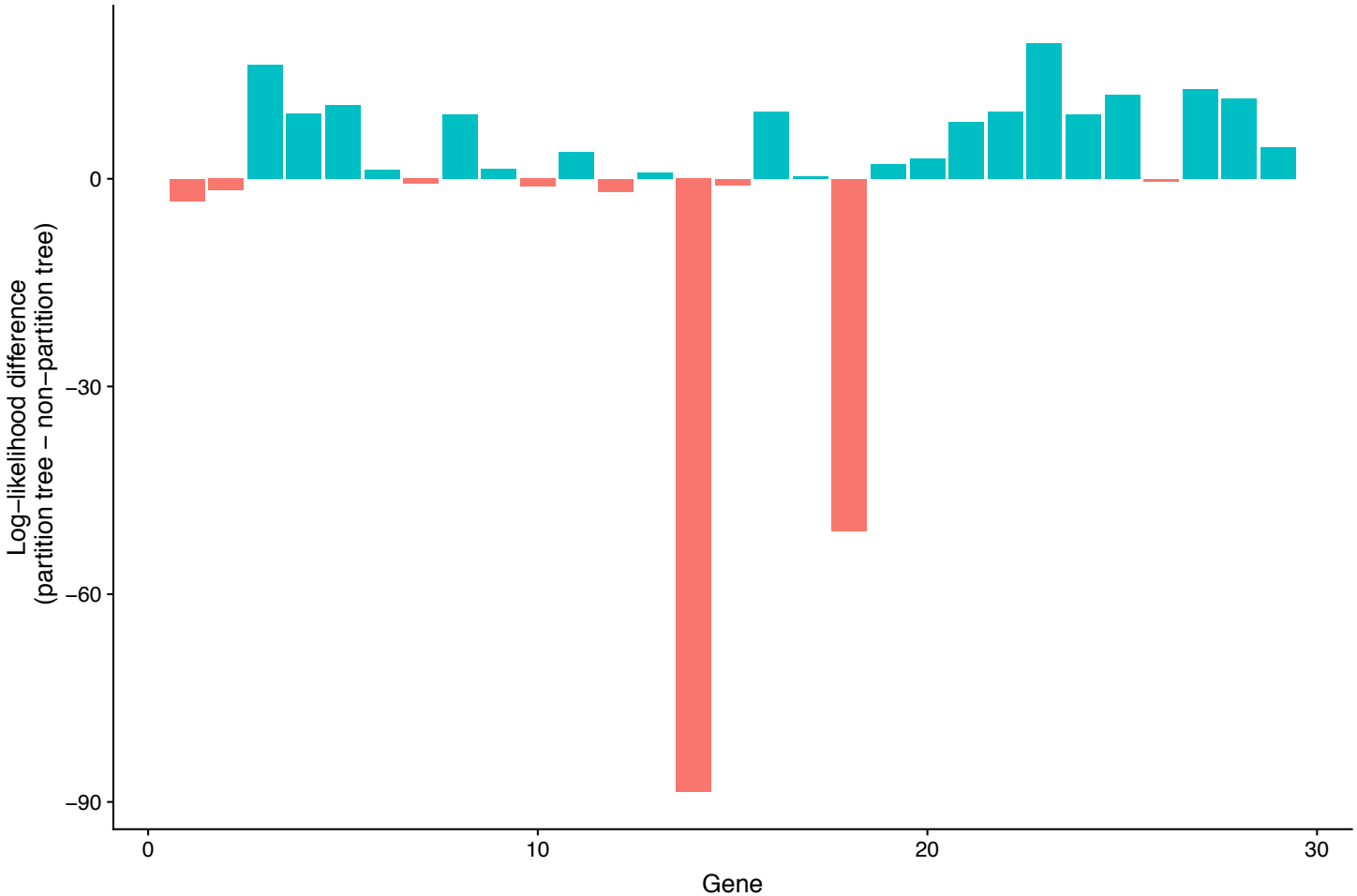
Inferred tree for the full partition model. The node annotation indicates BS/GCF/SCF scores. The contentious branch splitting turtles and archosaurs has GCF of 37.9%, which equates to 11 of the 29 genes supporting this topology.

Solutions - Concordance factors



Inferred tree for the non-partition model. The node annotation indicates BS/GCF/SCF scores. The contentious branch splitting turtles and archosaurs has GCF of 17.2%, which equates to 5 of the 29 genes supporting this topology.

Solutions - Identifying most influential genes



When we examine the difference in log-likelihoods between the two trees on a per-gene basis, we notice that two particular genes strongly support turtles as sister to crocodiles, whereas most other genes are either neutral, or support turtles as sister to archosaurs.