Phylogenetic analysis with IQ-TREE

http://www.iqtree.org

Minh Bui

Australian National University, Canberra

Sydney Phylogenetics Workshop
29 July 2021

IQ-TREE DEVELOPMENT TEAM

Australia

James Barbetti
Contribution: Software engineering for COVID-19 data

Thomas Wong
Contribution: ModelFinder 2

Michael Woodhams
Google Scholar
Contribution: Lie Markov models.

Robert Lanfear
Google Scholar
Contribution: Inspiring ideas and advice.

Bui Quang Minh
Google Scholar
Contribution: Team leader, software core, ultrafast bootstrap, model selection.

Austria

Olga Chernomor
Google Scholar
Contribution: Partition models and phylogenomic search.

Heiko A. Schmidt
Google Scholar
Contribution: Integration of TREE-PUZZLE features.

Dominik Schrempf
Google Scholar
Contribution: Polymorphism-aware models (PoMo).

Arndt von Haeseler
Google Scholar
Contribution: Inspiring ideas and advice.

Vietnam

Diep Thi Hoang
Contribution: Improving ultrafast bootstrap.

Thanks to plenty of users for feedback and bug reports!

Why IQ-TREE?

Next generation sequencing data represent both a blessing and a curse:

- Blessing: (Phylo)genomic data help to elucidate many phylogenetic questions.
- Curse: Many model assumptions become increasingly distant from the truth due to growing data complexity.

"All models are wrong, but some are useful" (Box, 1976)

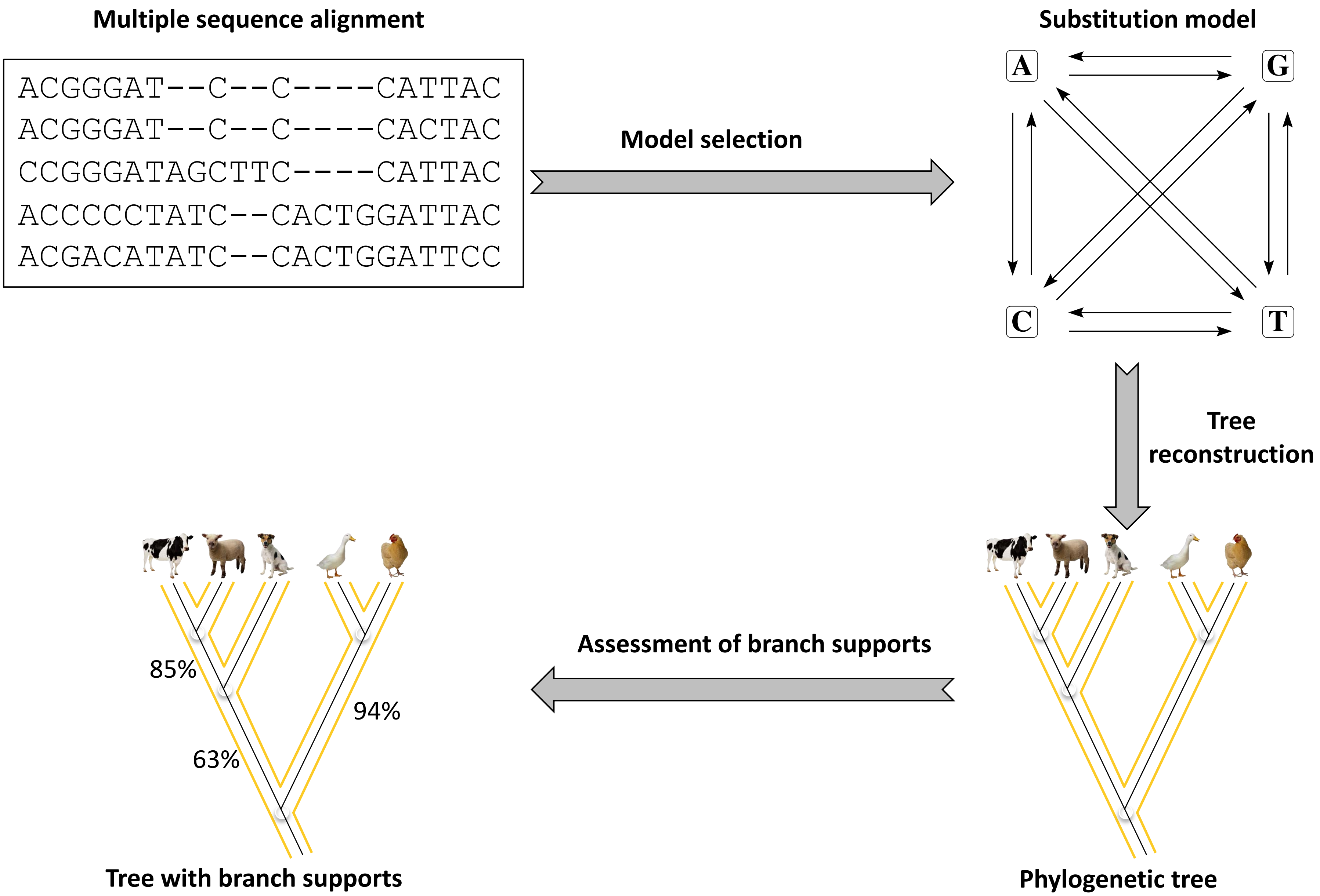
With IQ-TREE we aim to:

- Analyze ultra-large data sets.
- Provide many (if not most) "useful" models of sequence evolution.

But still, there are RAxML, PhyML out there, why do we need IQ-TREE?

- We better have at least 2 software independently developed for similar purpose. Only then, the pros and cons (sometimes **bugs**) can be identified. This creates a *friendly* competition, which helps to advance the field!
- Same as having MrBayes, RevBayes, BEAST for Bayesian inference.

Typical phylogenetic analysis under maximum likelihood



Multiple sequence alignment

```
ACGGGAT--C--C----CATTAC
ACGGGAT--C--C----CACTAC
CCGGGATAGCTTC----CATTAC
ACCCCCTATC--CACTGGATTAC
ACGACATATC--CACTGGATTCC
```

Model selection

Substitution model

Tree reconstruction

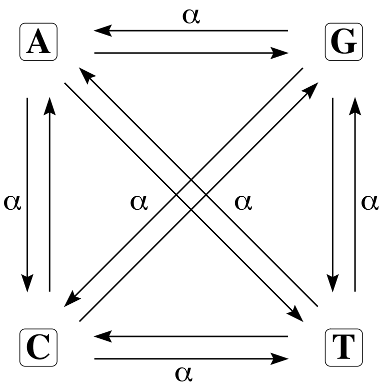
Phylogenetic tree

Assessment of branch supports

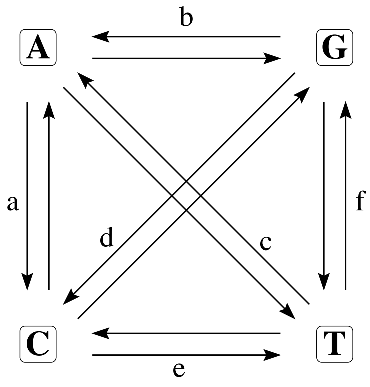
Tree with branch supports

Step 1: Model selection

Question: Which model fits best to the data?



Jukes-Cantor 1969 (JC)



General time reversible (GTR)

22 DNA models, 36 protein models, 12 codon models, 4 binary/morphological models

Combined with rate heterogeneity across sites:

- +I: a proportion of invariable sites (e.g., JC+I)
- +Γ: Gamma distribution (e.g., GTR+G)
- +I+Γ: mixture of +I and +Γ (e.g., GTR+I+G)
- +R: distribution-free rate model (e.g., GTR+R5)

Complex models:

- Non-reversible models
- Partition models
- Mixture models

https://www.nature.com/articles/nmeth.4285 (Nature Methods 2017)

Step 1: ModelFinder

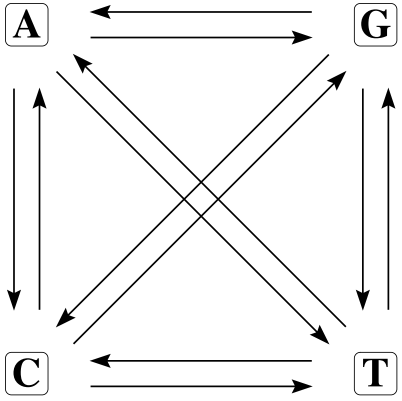
Multiple sequence alignment

```
ACGGGAT--C--C----CATTAC
ACGGGAT--C--C----CACTAC
CCGGGATAGCTTC----CATTAC
ACCCCCTATC--CACTGGATTAC
ACGACATATC--CACTGGATTCC
```

Model selection

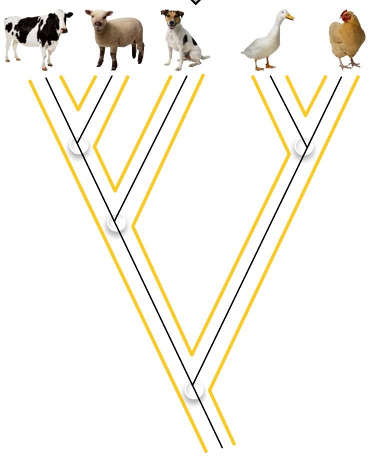
ModelFinder

Substitution model



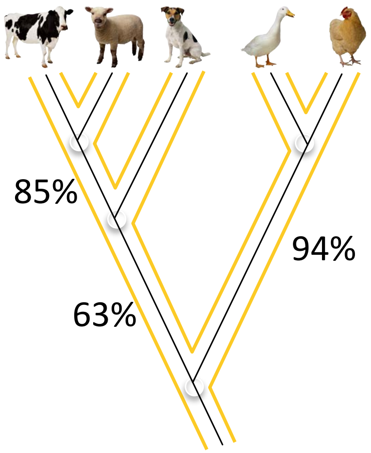
- More biologically plausible models.
- Faster alternative to jModelTest, ProtTest, and PartitionFinder

Tree reconstruction



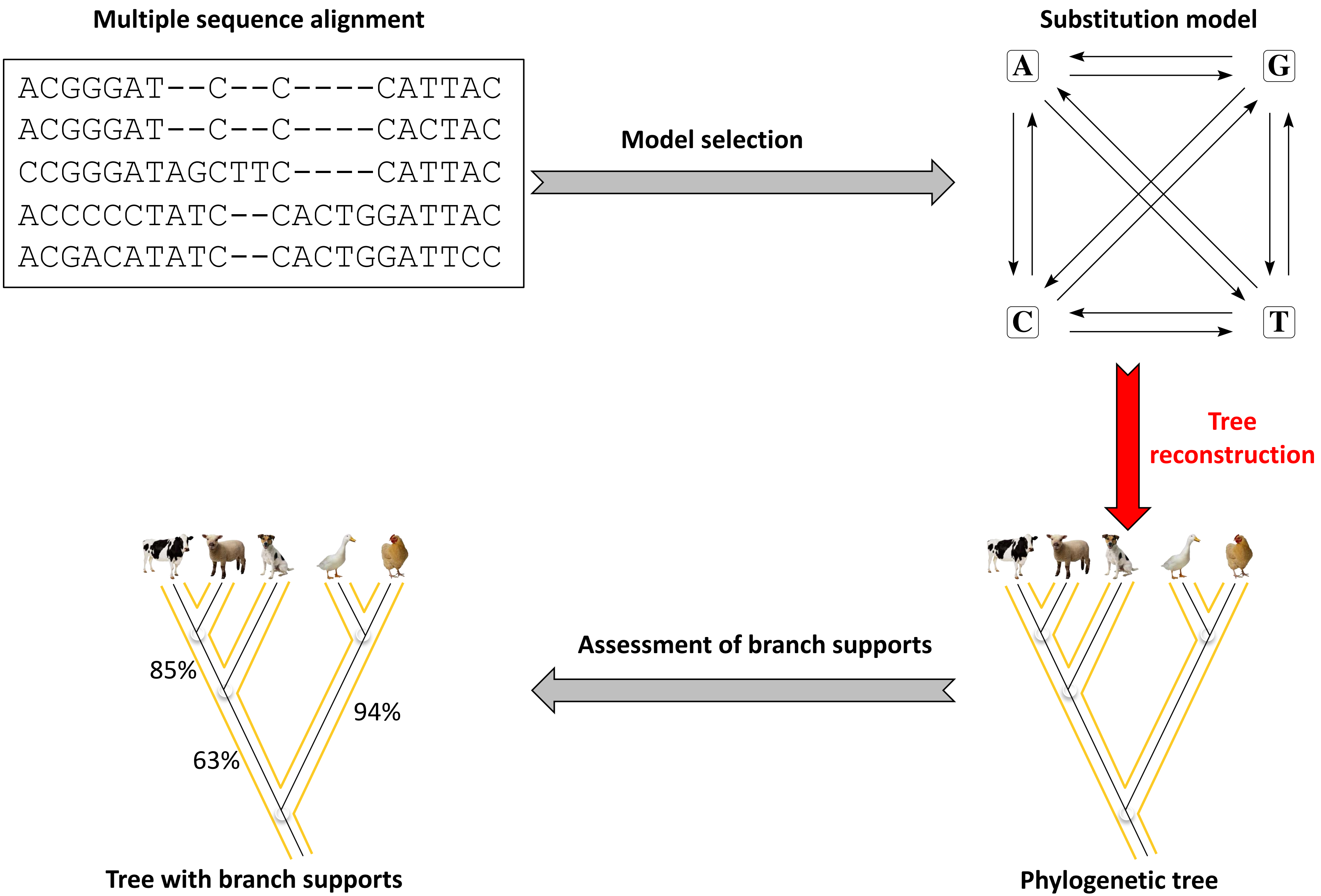
Phylogenetic tree

Assessment of branch supports



Tree with branch supports

Step 2: Tree reconstruction with IQ-TREE



Multiple sequence alignment

```
ACGGGAT--C--C----CATTAC
ACGGGAT--C--C----CACTAC
CCGGGATAGCTTC----CATTAC
ACCCCCTATC--CACTGGATTAC
ACGACATATC--CACTGGATTCC
```

Model selection

Substitution model

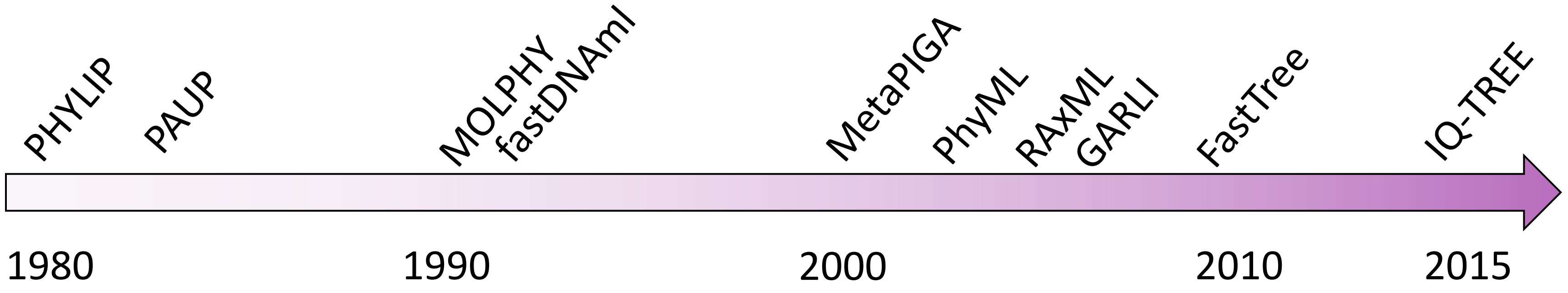
Tree reconstruction

Phylogenetic tree

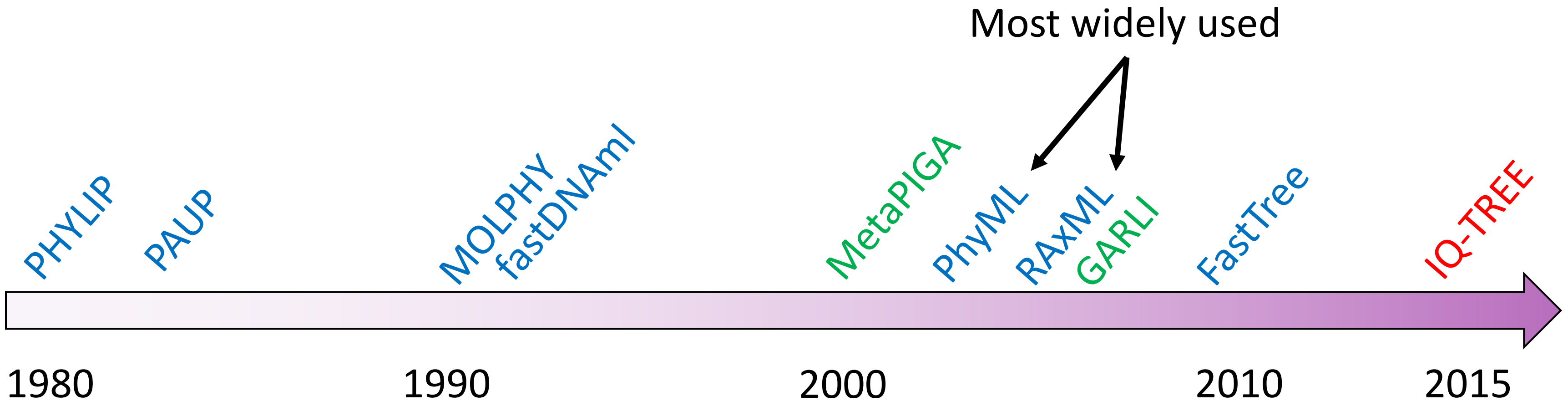
Assessment of branch supports

Tree with branch supports

Search heuristics for finding maximum likelihood trees



Search heuristics for finding maximum likelihood trees



Most widely used

PHYLIP

PAUP

MOLPHY

fastDNAml

MetaPIGA

PhyML

RAxML

GARLI

FastTree

IQ-TREE

1980

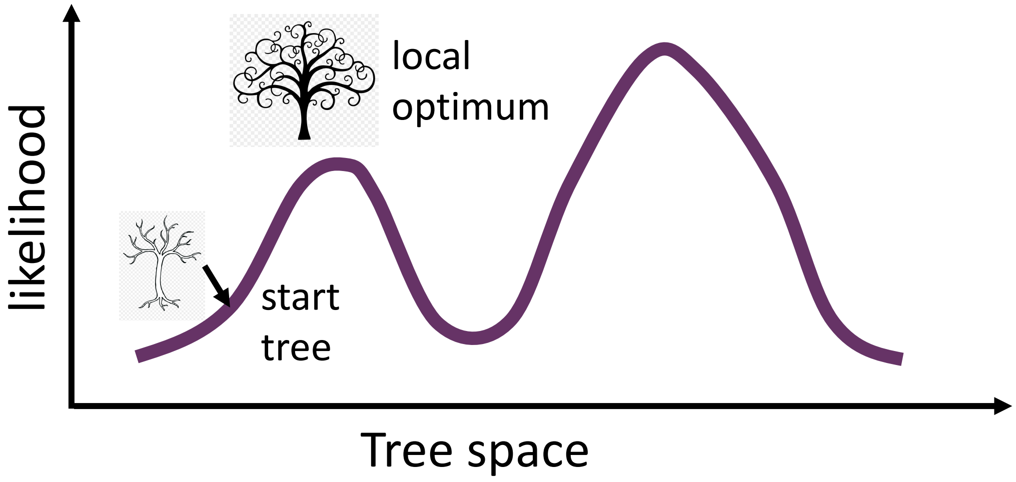
1990

2000

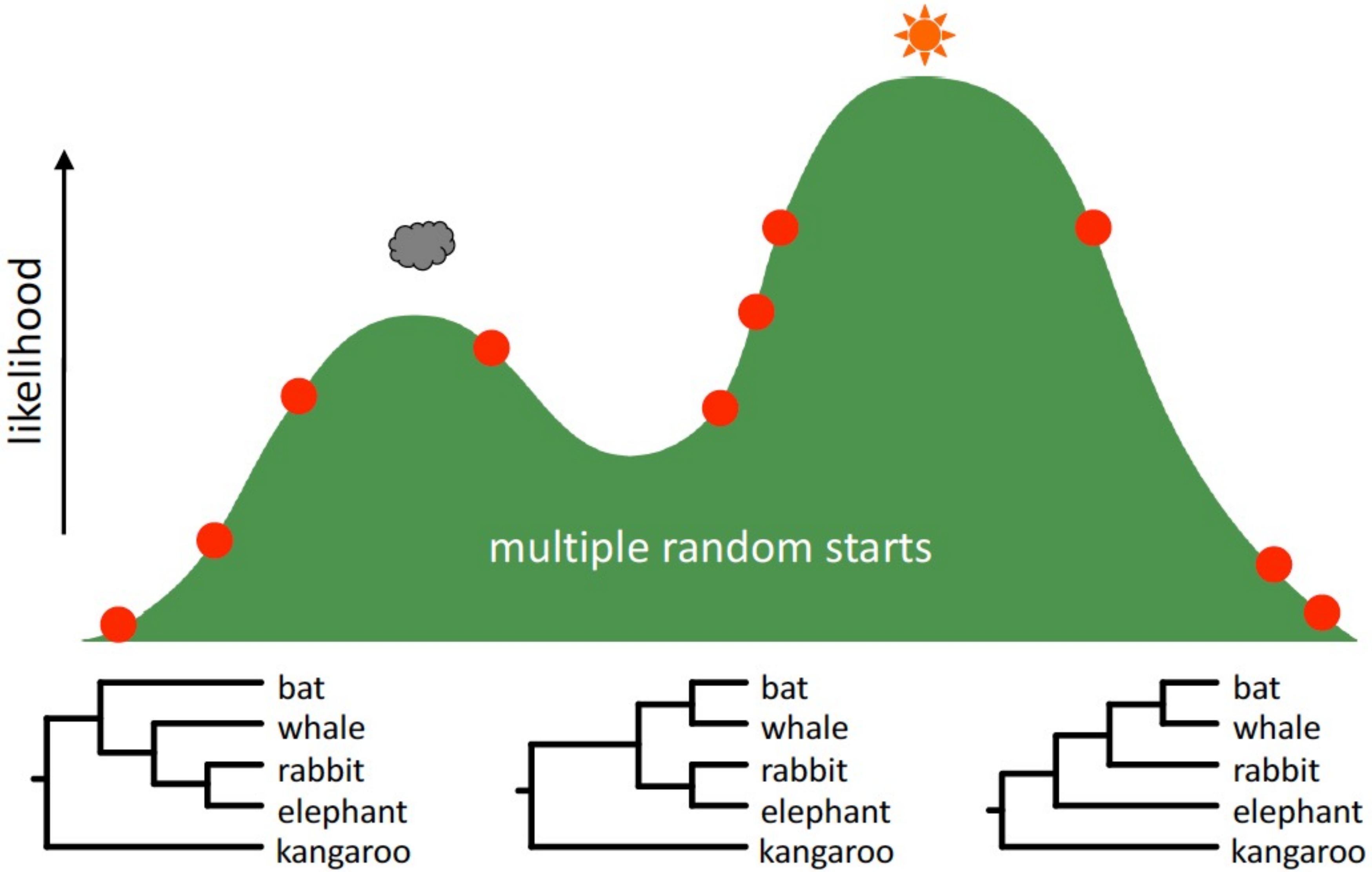
2010

2015

1. Hill-climbing / greedy algorithms: Fast but local optimum
2. Genetic algorithm: Slow but escaping local optima
3. IQ-TREE: Fast and escaping local optima

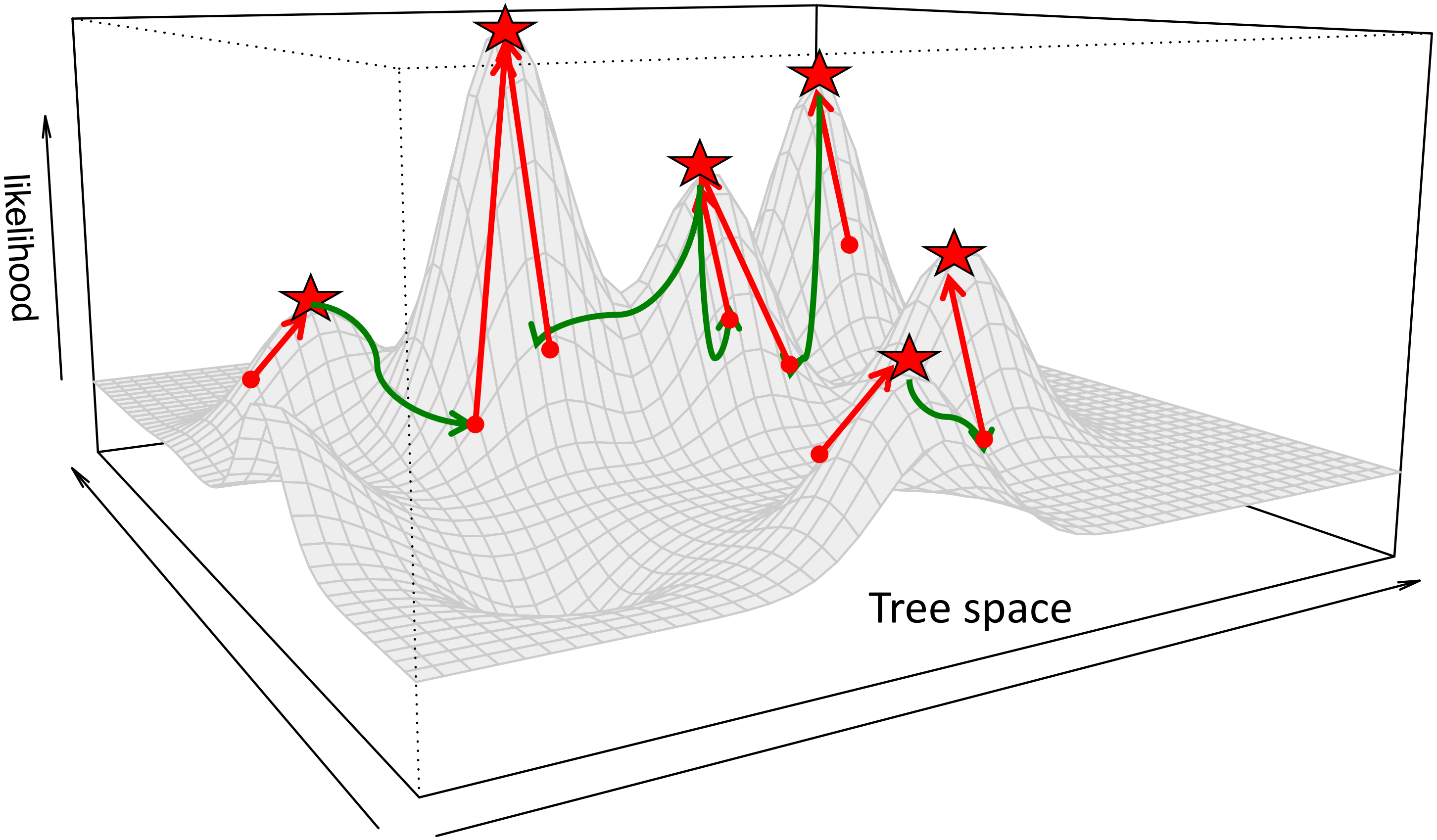


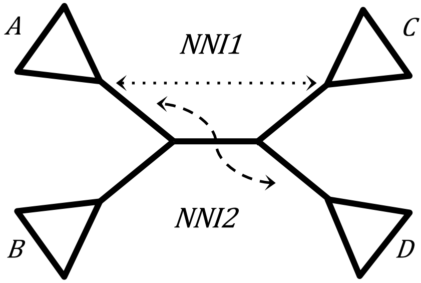
Heuristic search



multiple random starts

IQ-TREE: A new stochastic algorithm





Nearest neighbor interchange

Metaheuristics:
Iterated local search, Evolution strategy

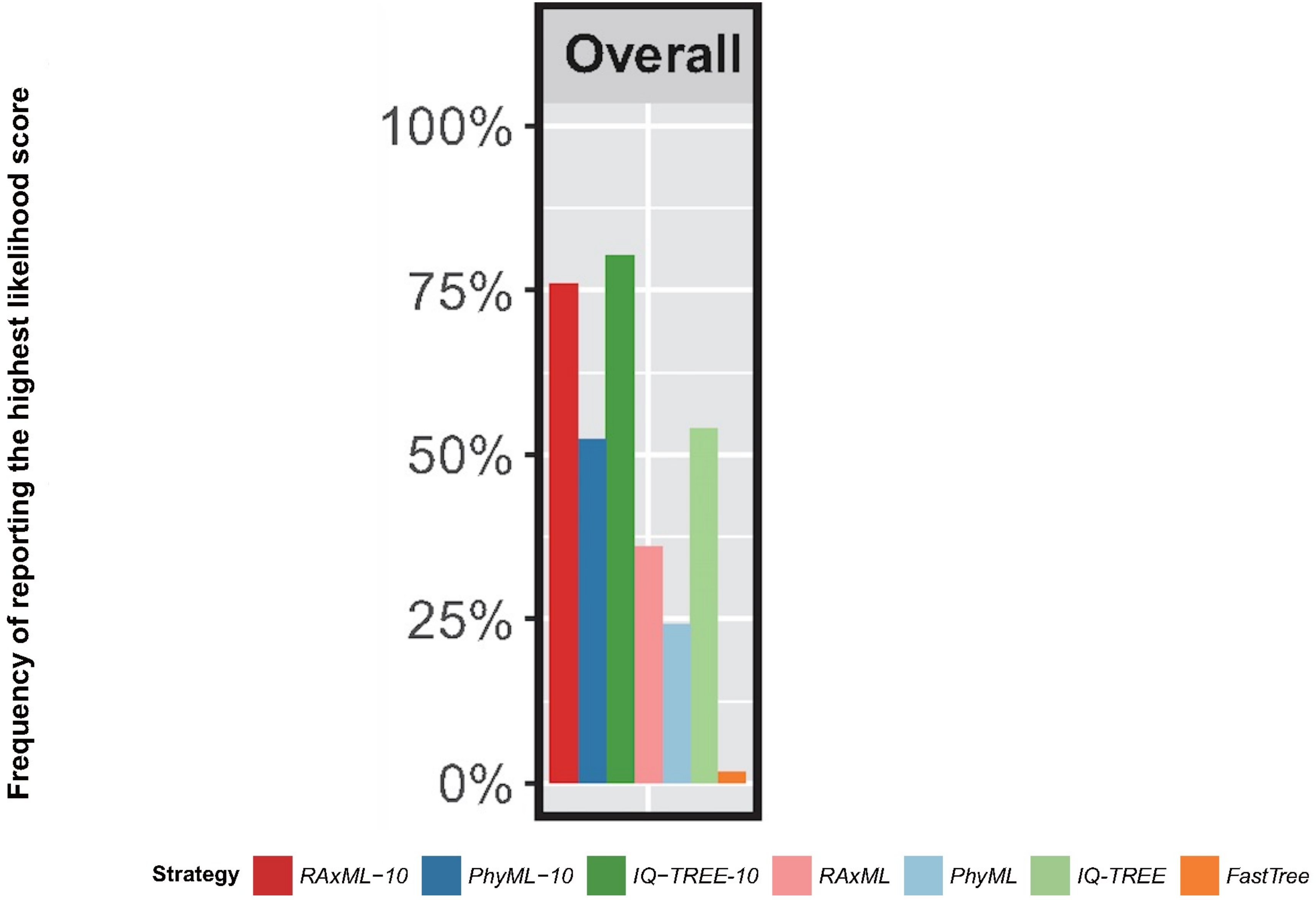
https://doi.org/10.1093/molbev/msu300
(*Mol. Biol. Evol.* 2015)

Lam-Tung Nguyen

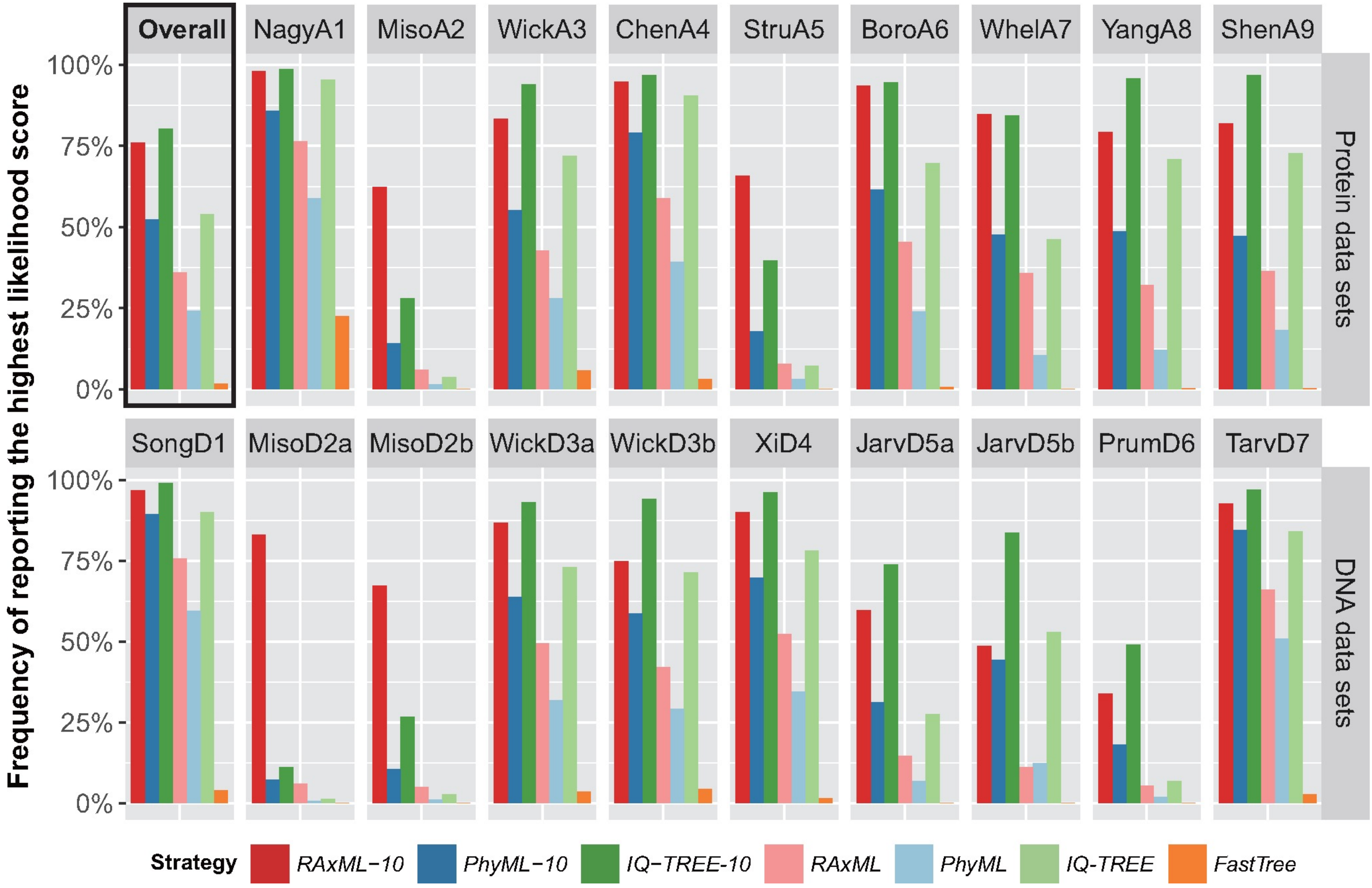
Heiko Schmidt

Arndt von Haeseler

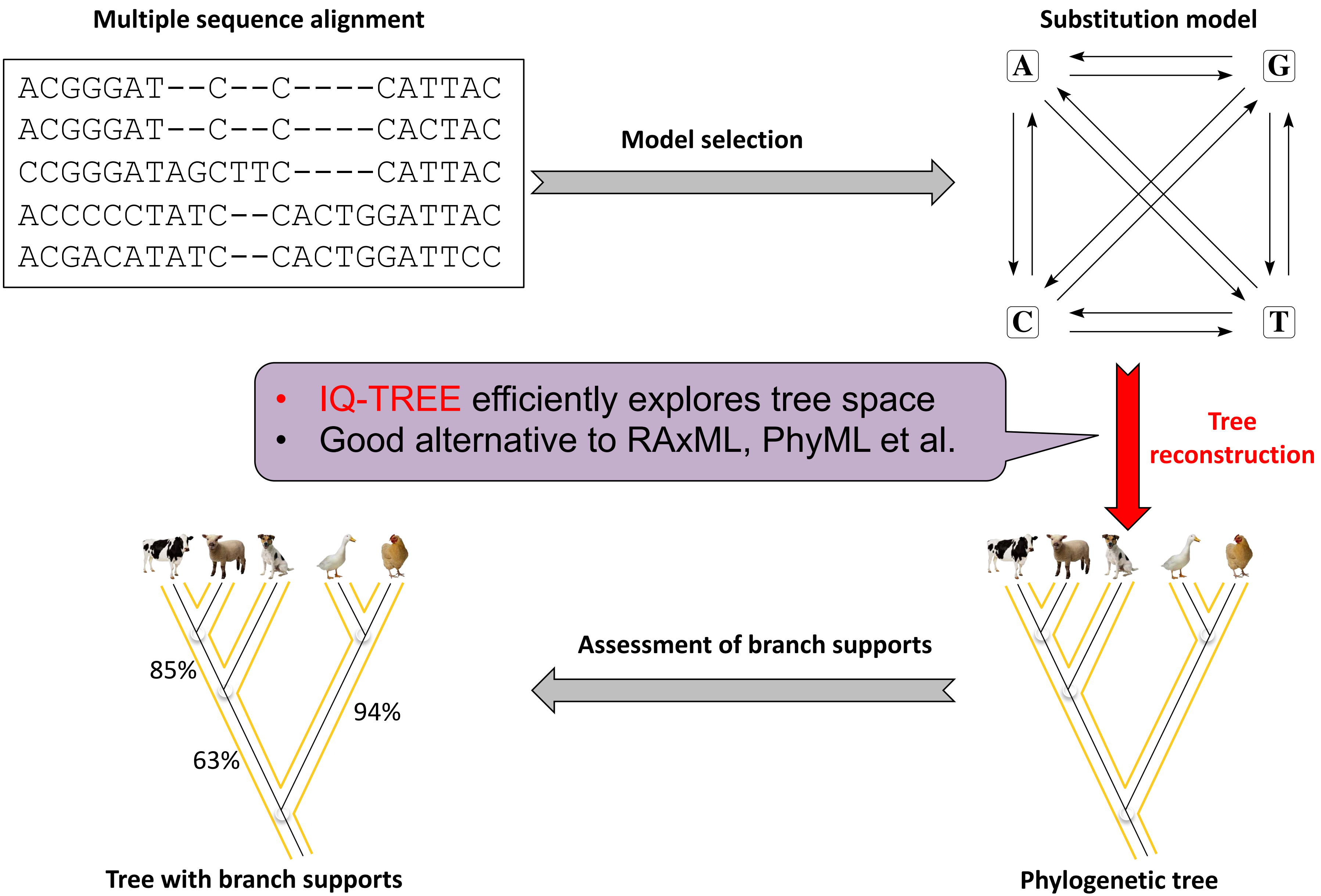
An independent benchmark by Zhou et al. (2018)



An independent benchmark by Zhou et al. (2018)



Step 2: IQ-TREE summary



Multiple sequence alignment

```
ACGGGAT--C--C----CATTAC
ACGGGAT--C--C----CACTAC
CCGGGATAGCTTC----CATTAC
ACCCCCTATC--CACTGGATTAC
ACGACATATC--CACTGGATTCC
```

Model selection

Substitution model

- IQ-TREE efficiently explores tree space
- Good alternative to RAxML, PhyML et al.

Tree reconstruction

Phylogenetic tree

Assessment of branch supports

Tree with branch supports

Step 3: Ultrafast bootstrap

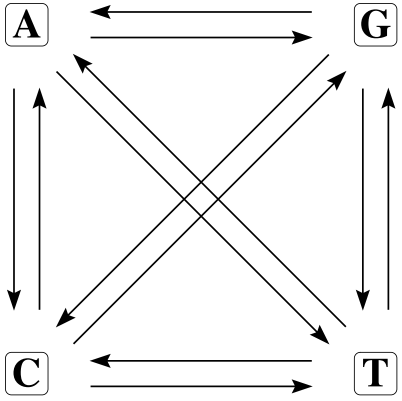
Multiple sequence alignment

```
ACGGGAT--C--C----CATTAC
ACGGGAT--C--C----CACTAC
CCGGGATAGCTTC----CATTAC
ACCCCCTATC--CACTGGATTAC
ACGACATATC--CACTGGATTCC
```

Model selection

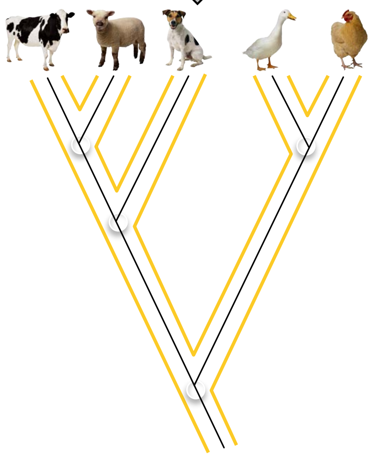
ModelFinder

Substitution model



IQ-TREE

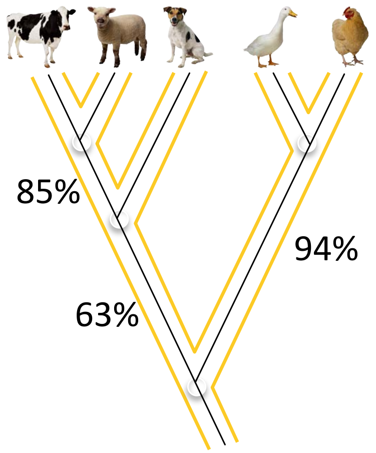
Tree reconstruction



Phylogenetic tree

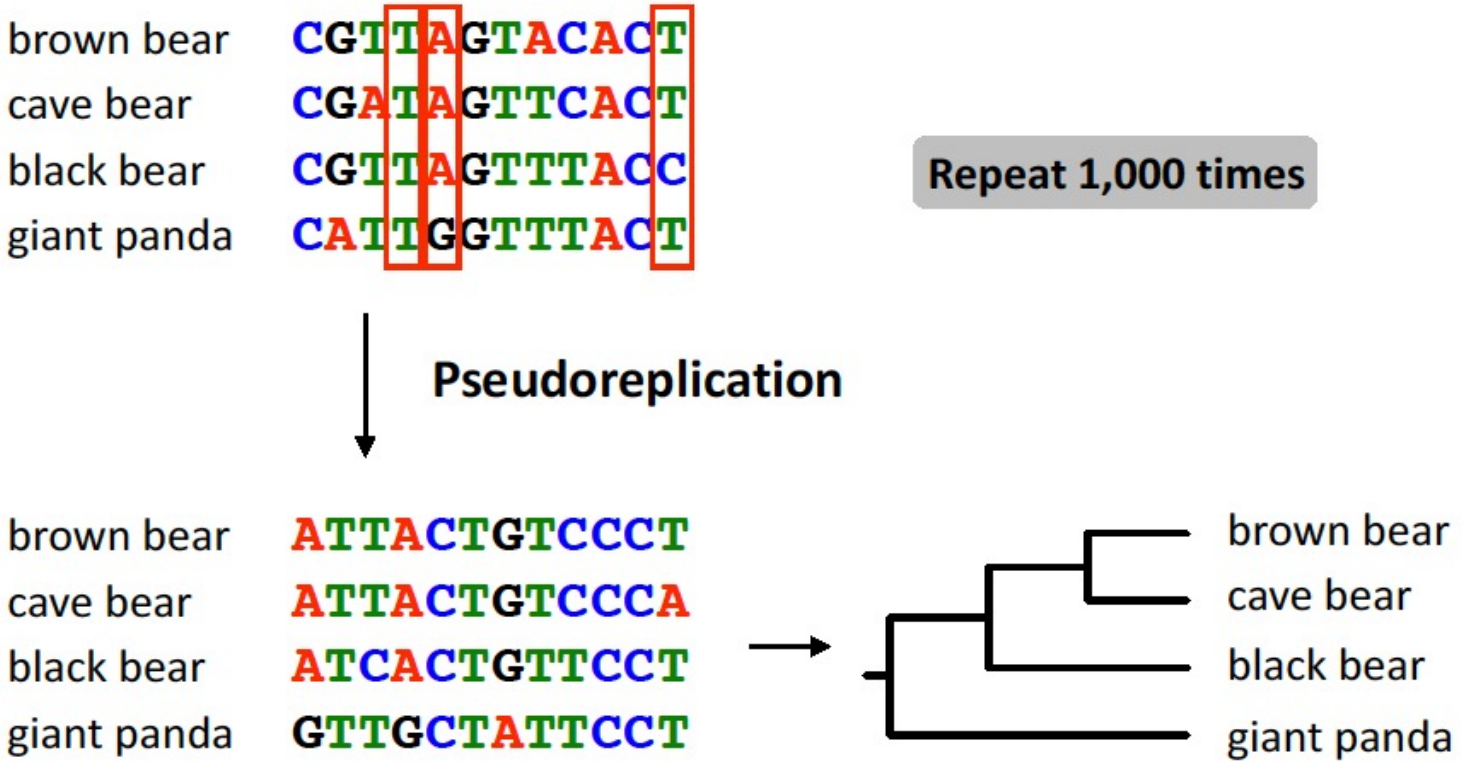
Ultrafast bootstrap

Branch supports



Tree with branch supports

Bootstrapping



brown bear	CGTTAGTACACT
cave bear	CGATAGTTCACT
black bear	CGTTAGTTTACC
giant panda	CATTGGTTTACT

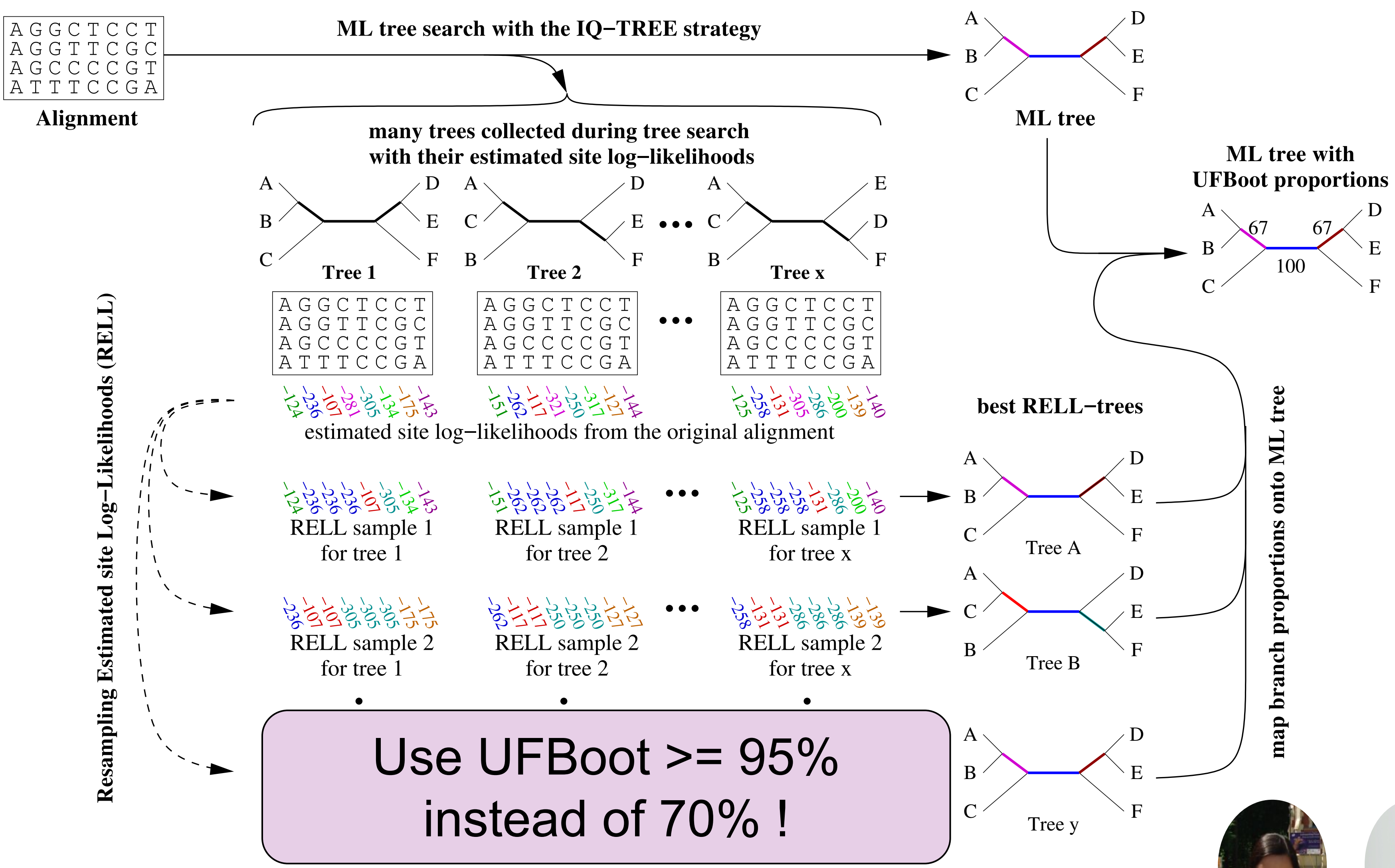
Repeat 1,000 times

Pseudoreplication

brown bear	ATTACTGTCCCT
cave bear	ATTACTGTCCCA
black bear	ATCACTGTTCCT
giant panda	GTTGCTATTCCT

Bootstrap analysis is extremely time-consuming!

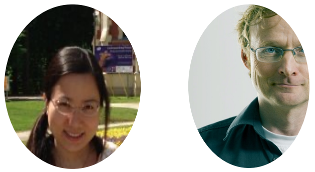
UFBoot: Ultrafast bootstrap approximation



```
A G G C T C C T
A G G T T C G C
A G C C C C G T
A T T T C C G A
```

Use UFBoot >= 95%
instead of 70% !

https://doi.org/10.1093/molbev/mst024 (2013)



M.A.T. Nguyen, A. von Haeseler

Step 3: Ultrafast bootstrap

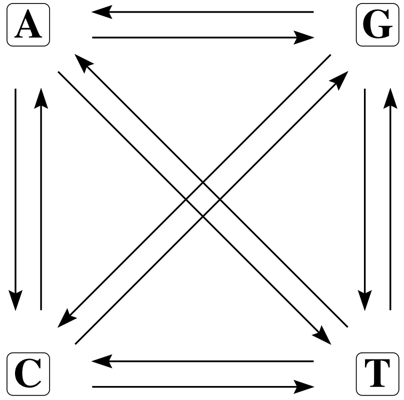
Multiple sequence alignment

```
ACGGGAT--C--C----CATTAC
ACGGGAT--C--C----CACTAC
CCGGGATAGCTTC----CATTAC
ACCCCCTATC--CACTGGATTAC
ACGACATATC--CACTGGATTCC
```

Model selection

ModelFinder

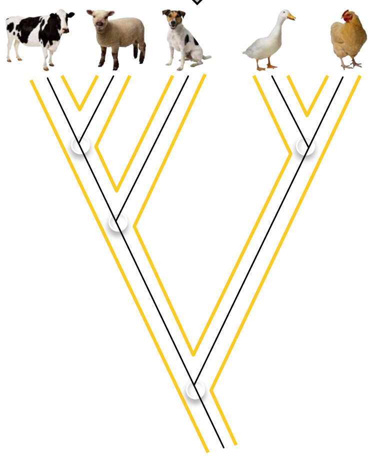
Substitution model



IQ-TREE

Tree reconstruction

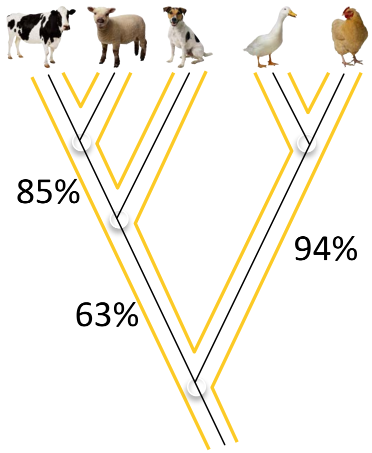
- A very fast alternative to standard bootstrap.
- More direct interpretation of bootstrap supports.



Phylogenetic tree

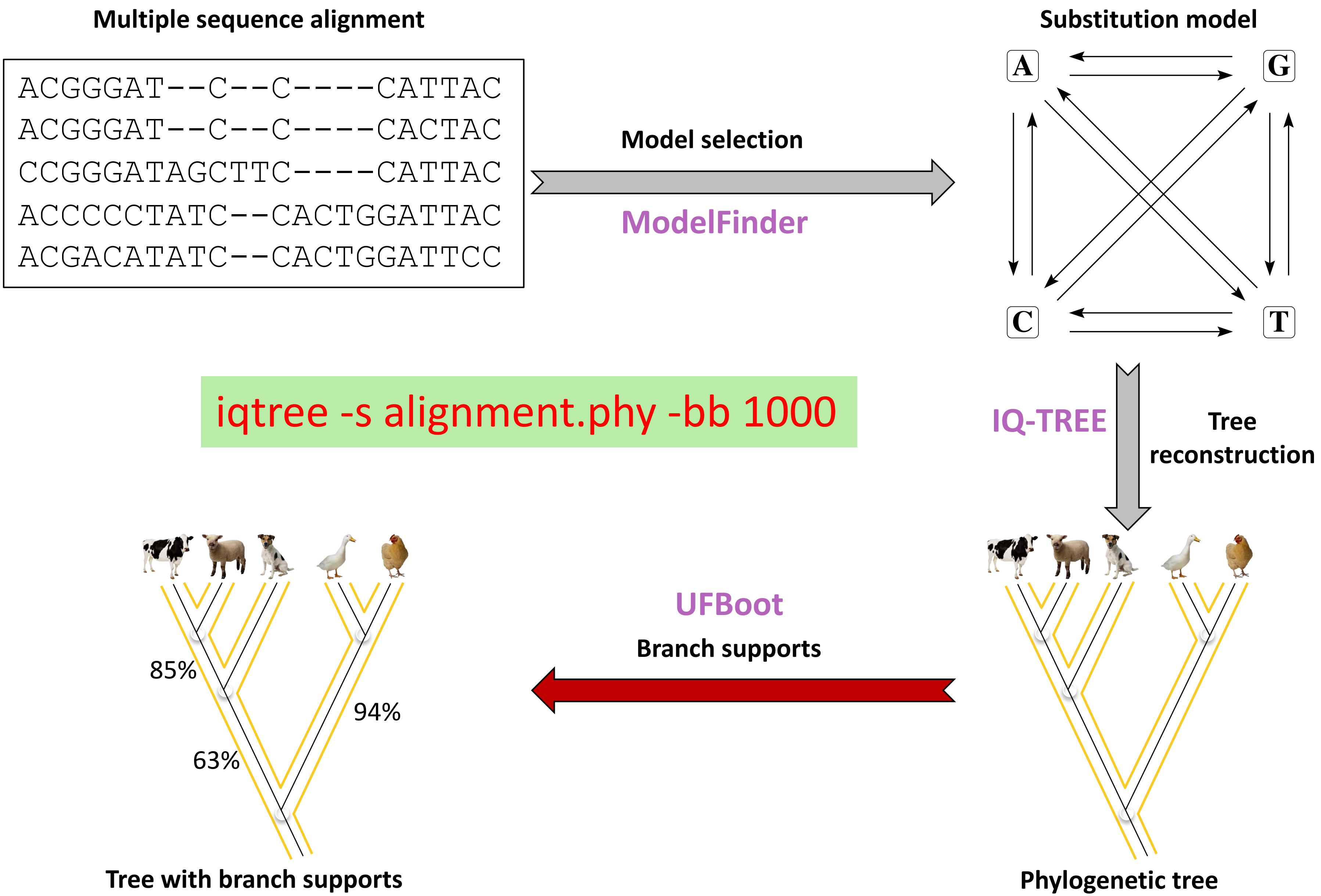
UFBoot

Branch supports



Tree with branch supports

Typical analysis in one IQ-TREE run



Multiple sequence alignment

```
ACGGGAT--C--C----CATTAC
ACGGGAT--C--C----CACTAC
CCGGGATAGCTTC----CATTAC
ACCCCCTATC--CACTGGATTAC
ACGACATATC--CACTGGATTCC
```

Model selection

ModelFinder

Substitution model

IQ-TREE

Tree reconstruction

```
iqtree -s alignment.phy -bb 1000
```

Phylogenetic tree

UFBoot

Branch supports

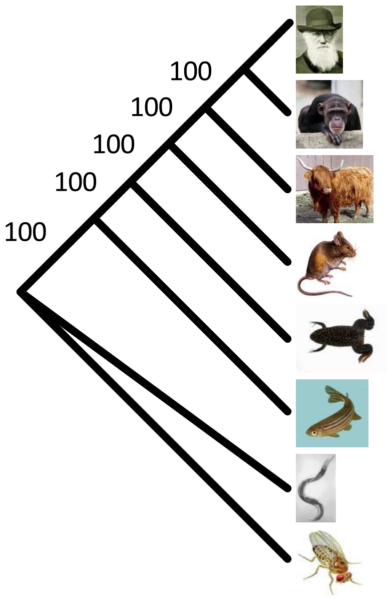
Tree with branch supports

Concatenation methods: Limitation

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



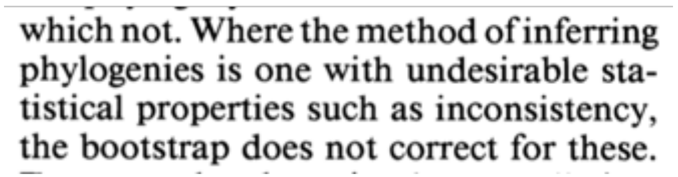
Species tree of life

Bootstrap supports and Bayesian posteriors tend to 100% as #genes increases!

Concatenation assumes a single tree across all loci

Potential *systematic bias*

Felsenstein (1985):

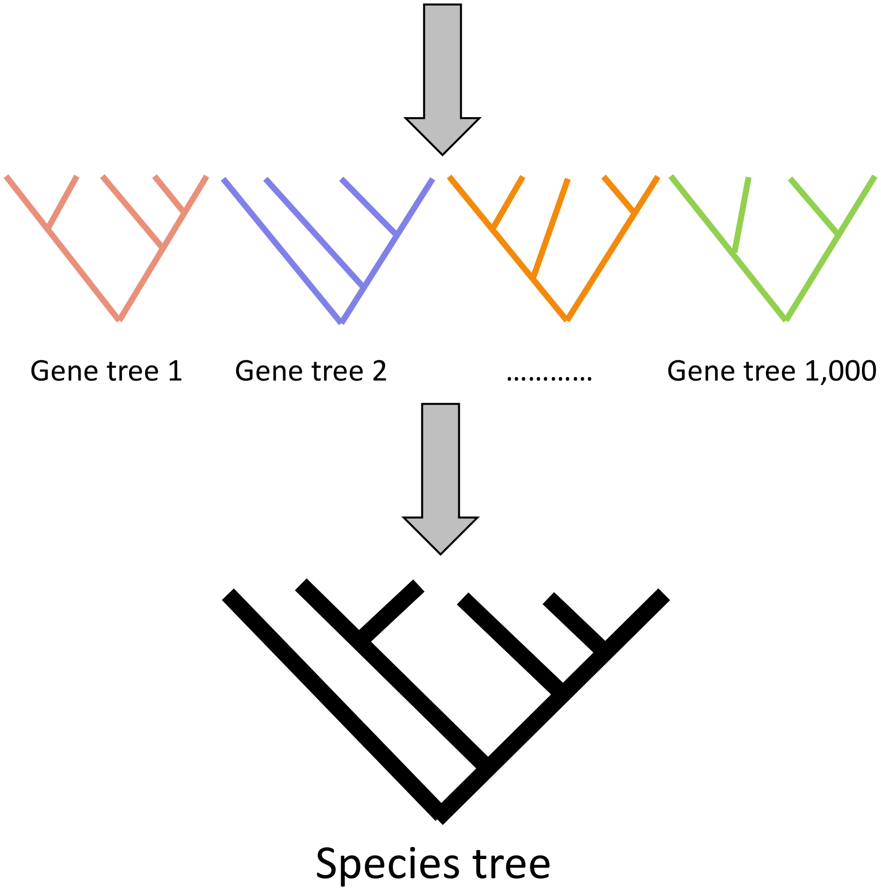


which not. Where the method of inferring phylogenies is one with undesirable statistical properties such as inconsistency, the bootstrap does not correct for these.

Coalescent/reconciliation methods

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	----------



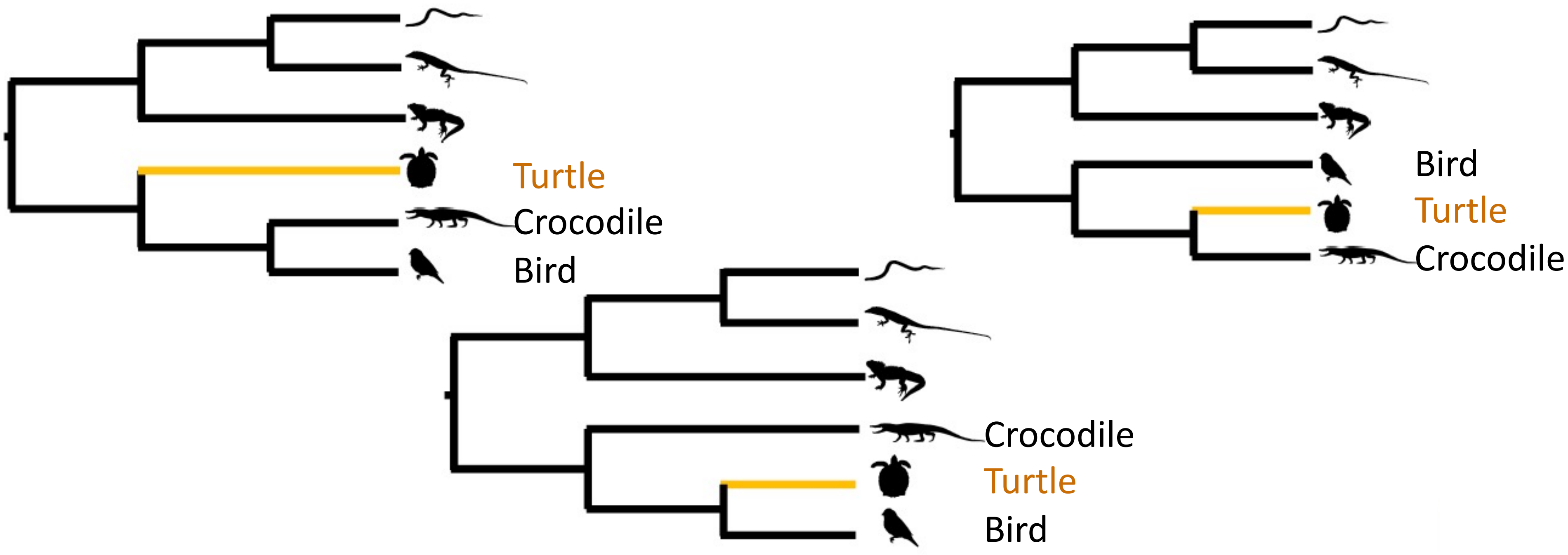
Gene Concordance Factor (gCF): How often a branch in species tree is found among gene trees?

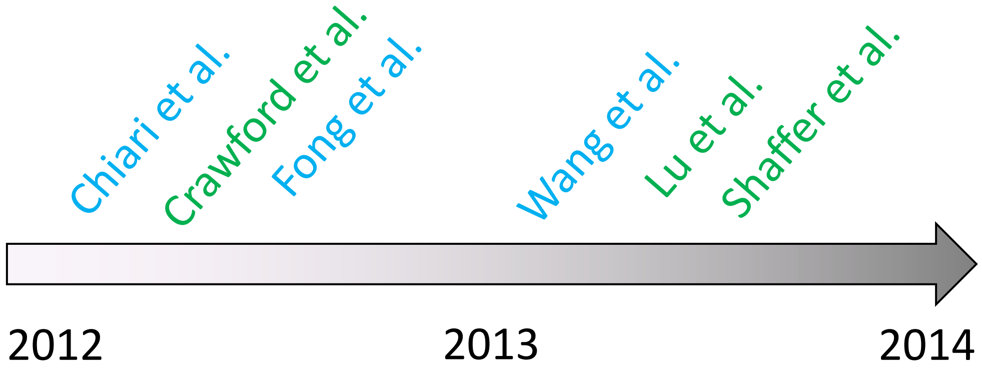
0% ≤ gCF ≤ 100%

Implementation in IQ-TREE fully accounts for missing data

Problem: Uncertainties in gene trees!

Suggested Tutorial: Turtle dataset





Chiari et al.

Crawford et al.

Fong et al.

Wang et al.

Lu et al.

Shaffer et al.

2012

2013

2014

Different studies led to different trees!

Molecular Evolution 2019 Woods Hole, MA

Suggested Tutorial: Exercises

1. Input data
2. Inferring the first phylogeny
3. Applying partition model
4. Choosing the best partitioning scheme
5. Tree topology tests
6. Concordance factors
7. Resampling partitions and sites
8. Identifying most influential genes
9. Wrapping up

http://www.iqtree.org/workshop/molevol2019

Molecular Evolution
2019
Woods Hole, MA