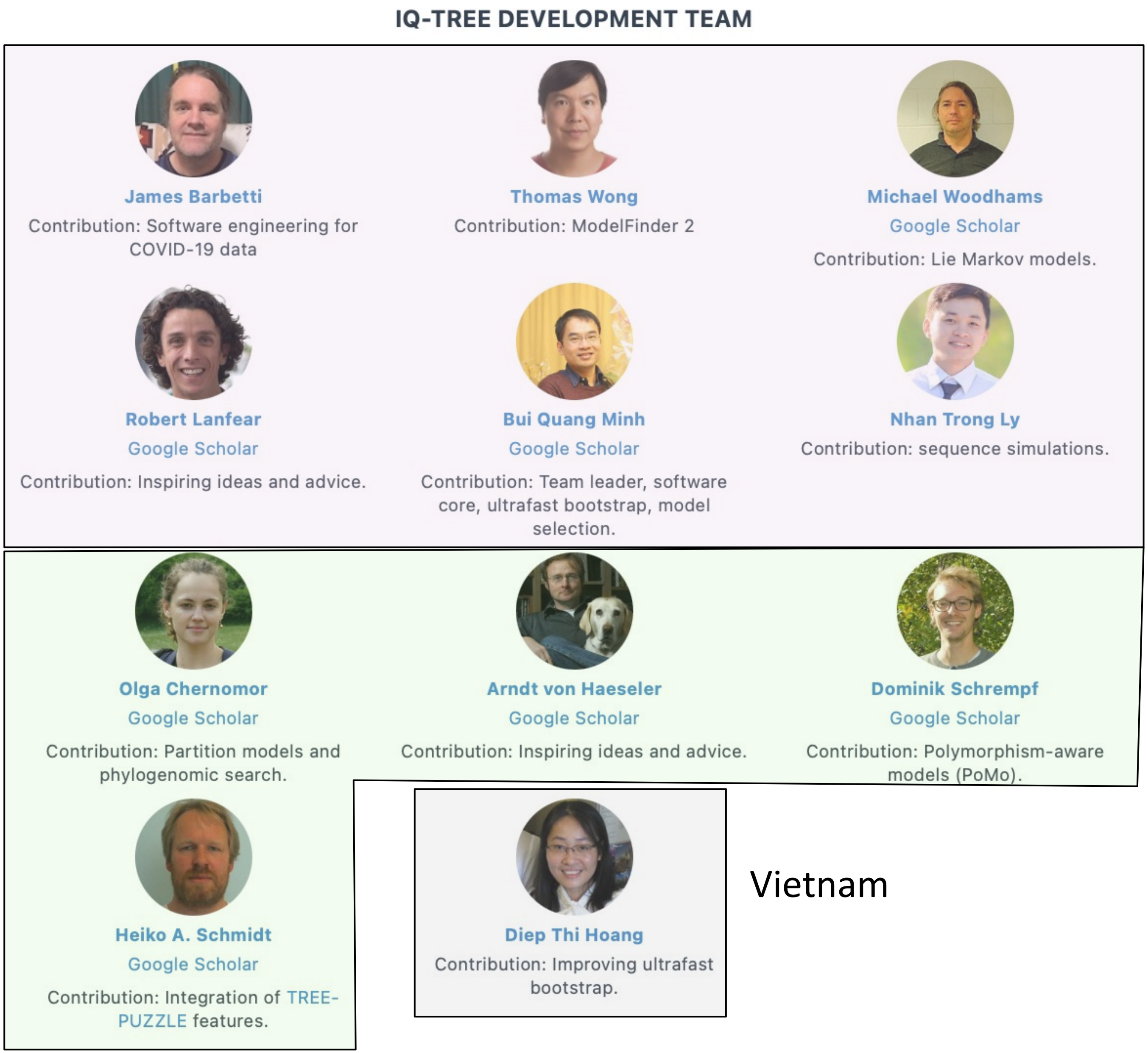
IQ-TREE

http://www.iqtree.org

Methods and Practice

Minh Bui
Australian National University

Sydney Phylogenetics Workshop
July 2022



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.

Nhan Trong Ly

Contribution: sequence simulations.

Austria

Olga Chernomor

Google Scholar

Contribution: Partition models and phylogenomic search.

Arndt von Haeseler

Google Scholar

Contribution: Inspiring ideas and advice.

Dominik Schrempf

Google Scholar

Contribution: Polymorphism-aware models (PoMo).

Heiko A. Schmidt

Google Scholar

Contribution: Integration of TREE-PUZZLE features.

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.
- Easy to use.

Typical phylogenetic analysis under maximum likelihood

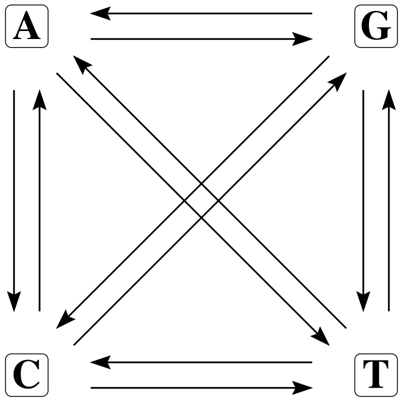
Multiple sequence alignment

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

Model selection

ModelFinder (2017)

Substitution model

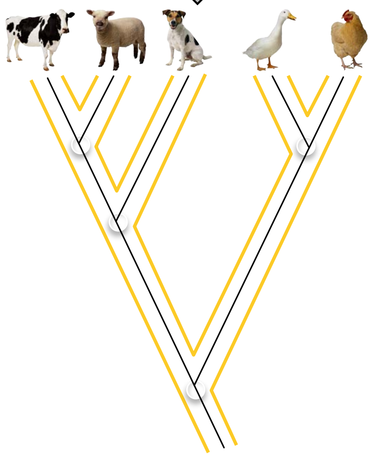


My work focused on improving all three steps for large datasets!

IQ-TREE (2015, 2020)

Tree reconstruction

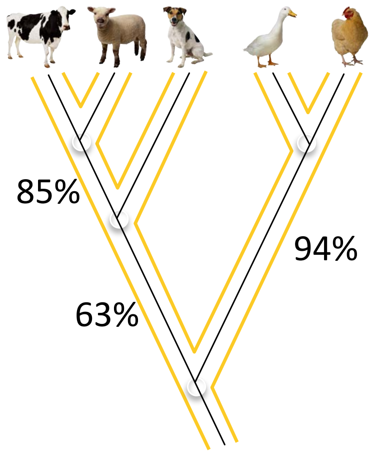
```
iqtree2 -s ALN_FILE -B 1000
```



Phylogenetic tree

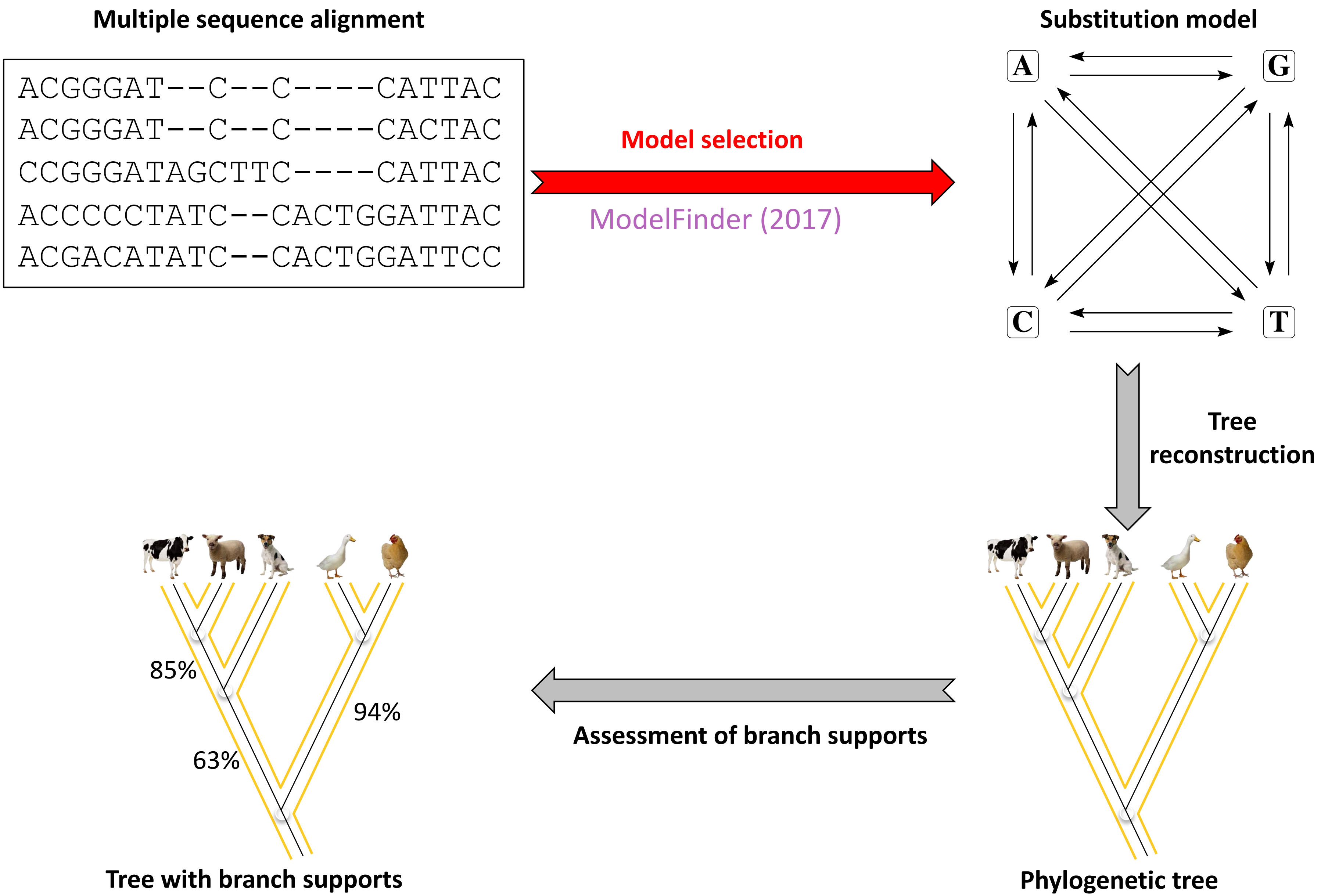
Ultrafast bootstrap (2013, 2018)

Assessment of branch supports



Tree with branch supports

Step 1: Model selection



Multiple sequence alignment

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

Model selection

ModelFinder (2017)

Substitution model

Tree reconstruction

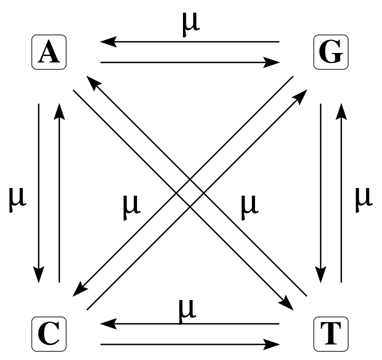
Phylogenetic tree

Assessment of branch supports

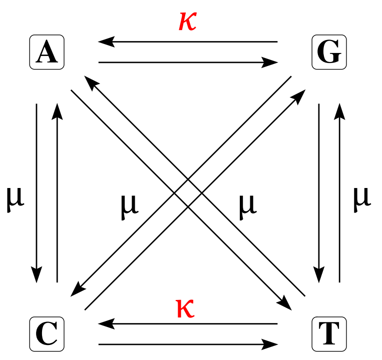
Tree with branch supports

Step 1: Which model fits best to the data?

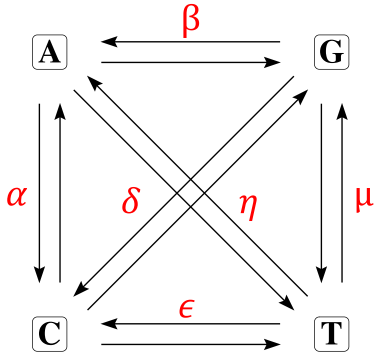
A model = substitution model + rate heterogeneity, e.g. “GTR+G”



JC
(Jukes & Cantor 1969)



HKY
(Hasegawa, Kishino, Yano 1985)

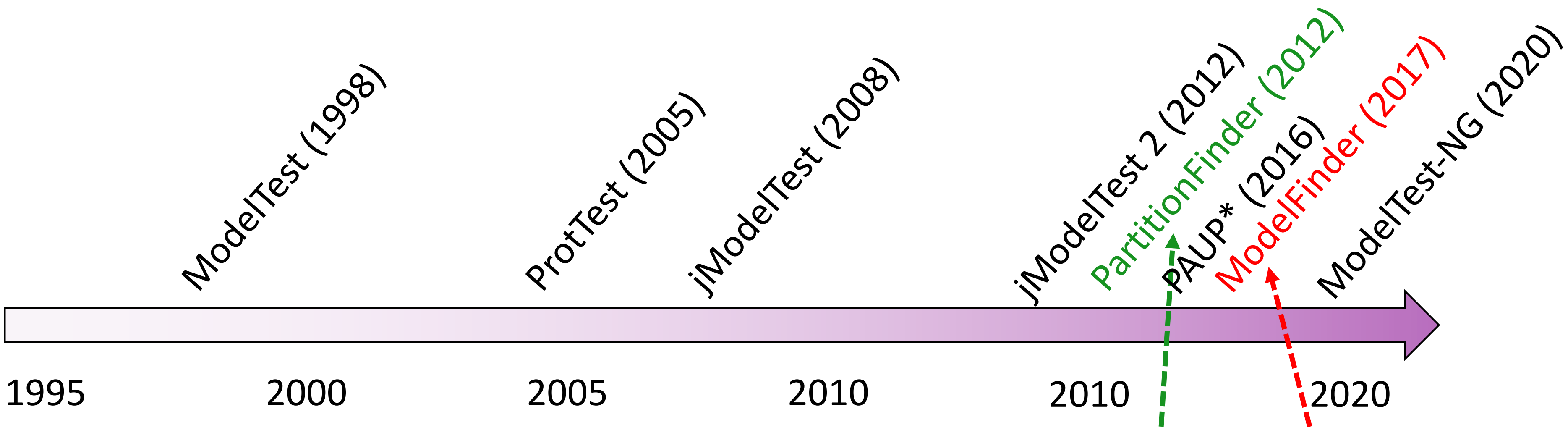


GTR
(General Time Reversible, 1986)

Rate heterogeneity: alignment sites evolved 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).

Model selection approaches



ModelTest (1998)

ProtTest (2005)

jModelTest (2008)

jModelTest 2 (2012)

PartitionFinder (2012)

PAUP* (2016)

ModelFinder (2017)

ModelTest-NG (2020)

1995

2000

2005

2010

2010

2020

Robert Lanfear
(ANU)

Lars Jermiin
(ANU & CSIRO)

Thomas Wong
(ANU)

- (j)Modeltest / ProtTest: slow and limited on models.
- PartitionFinder: better models for genomic data but still slow.
- ModelFinder: >10x faster and more realistic models.
- Current work: ModelFinder 2 = ModelFinder + PartitionFinder + ModelRevelator

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

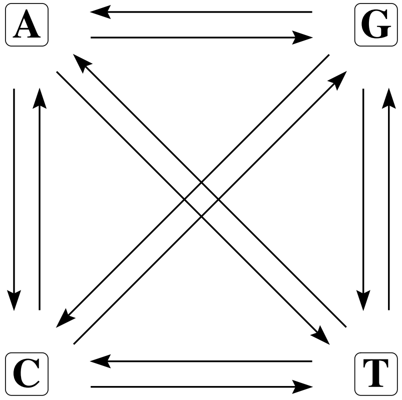
Step 2: Tree reconstruction

Multiple sequence alignment

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

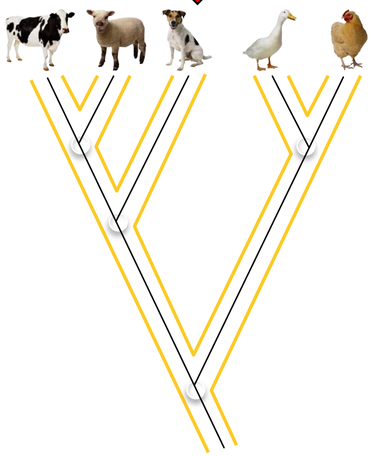
Model selection

Substitution model



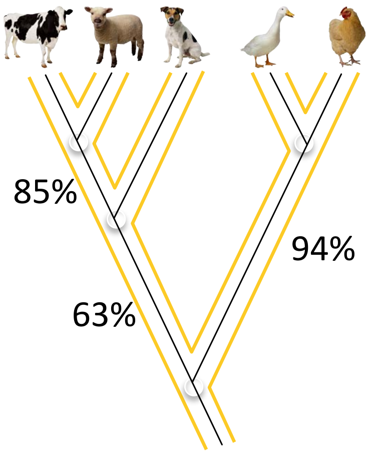
IQ-TREE (2015, 2020)

Tree reconstruction



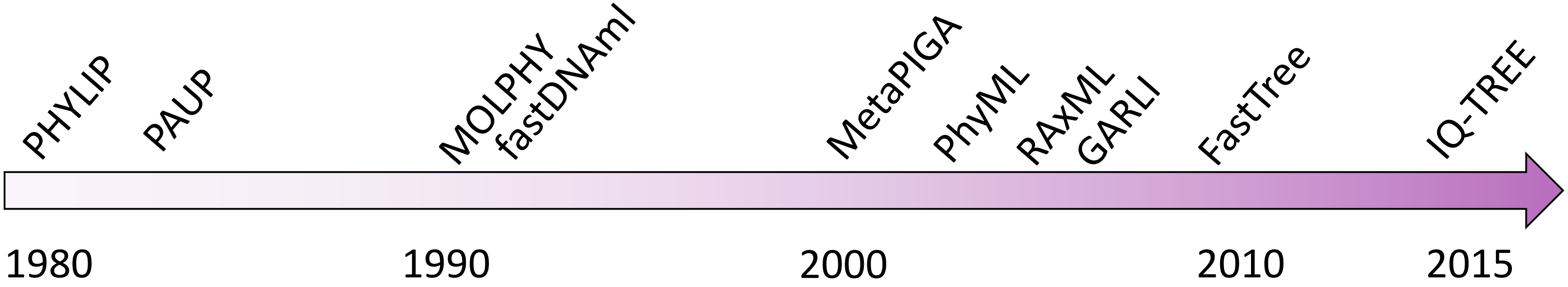
Phylogenetic tree

Assessment of branch supports



Tree with branch supports

Search heuristics for finding maximum likelihood trees



PHYLIP

PAUP

MOLPHY

fastDNAml

MetaPIGA

PhyML

RAxML

GARLI

FastTree

IQ-TREE

1980

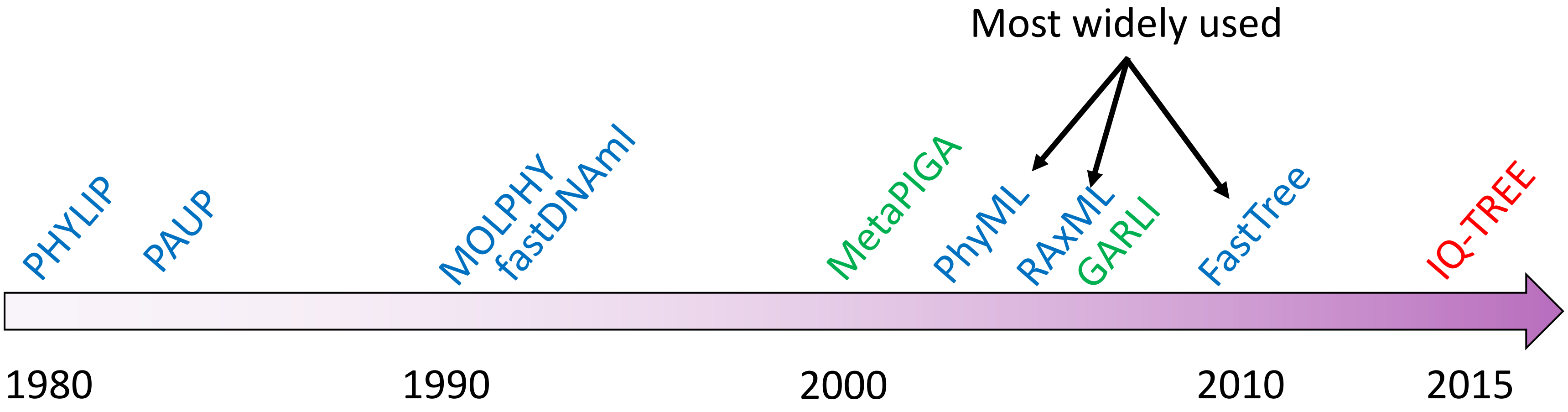
1990

2000

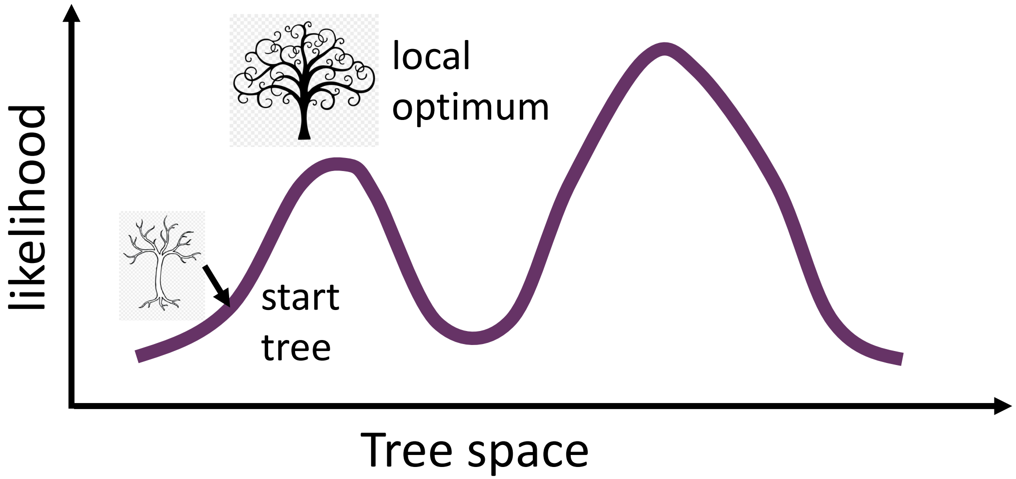
2010

2015

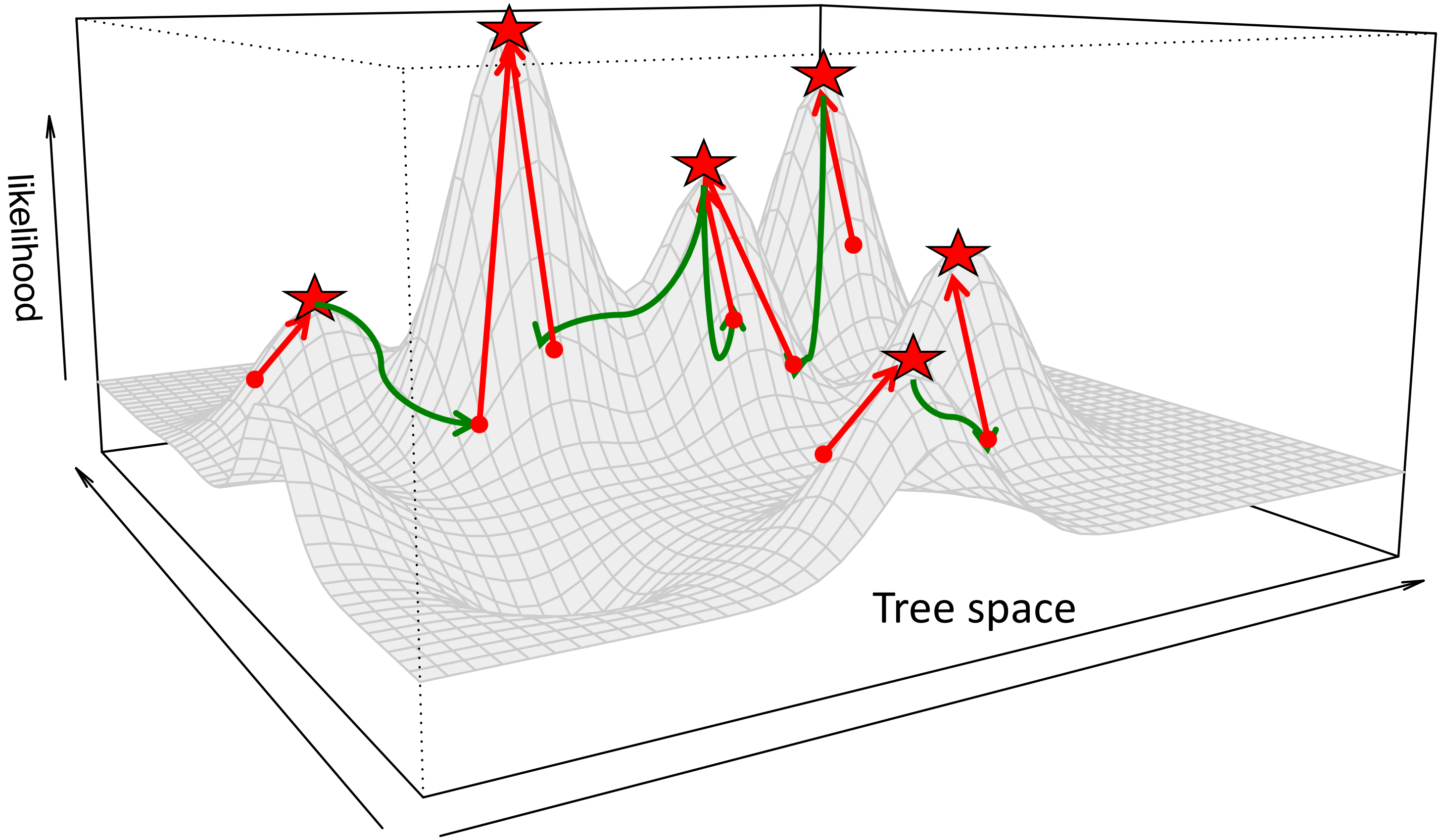
Search heuristics for finding maximum likelihood trees

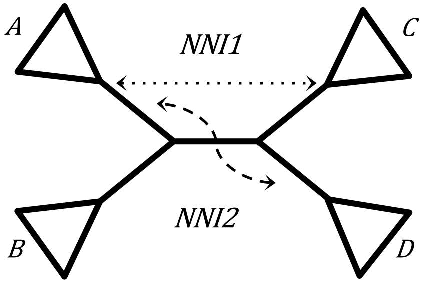


1. Hill-climbing / greedy algorithms: Fast but local optimum
2. Genetic algorithm: Slow but escaping local optima
3. IQ-TREE: Fast and escaping local optima



IQ-TREE: A new stochastic algorithm





Nearest neighbor interchange

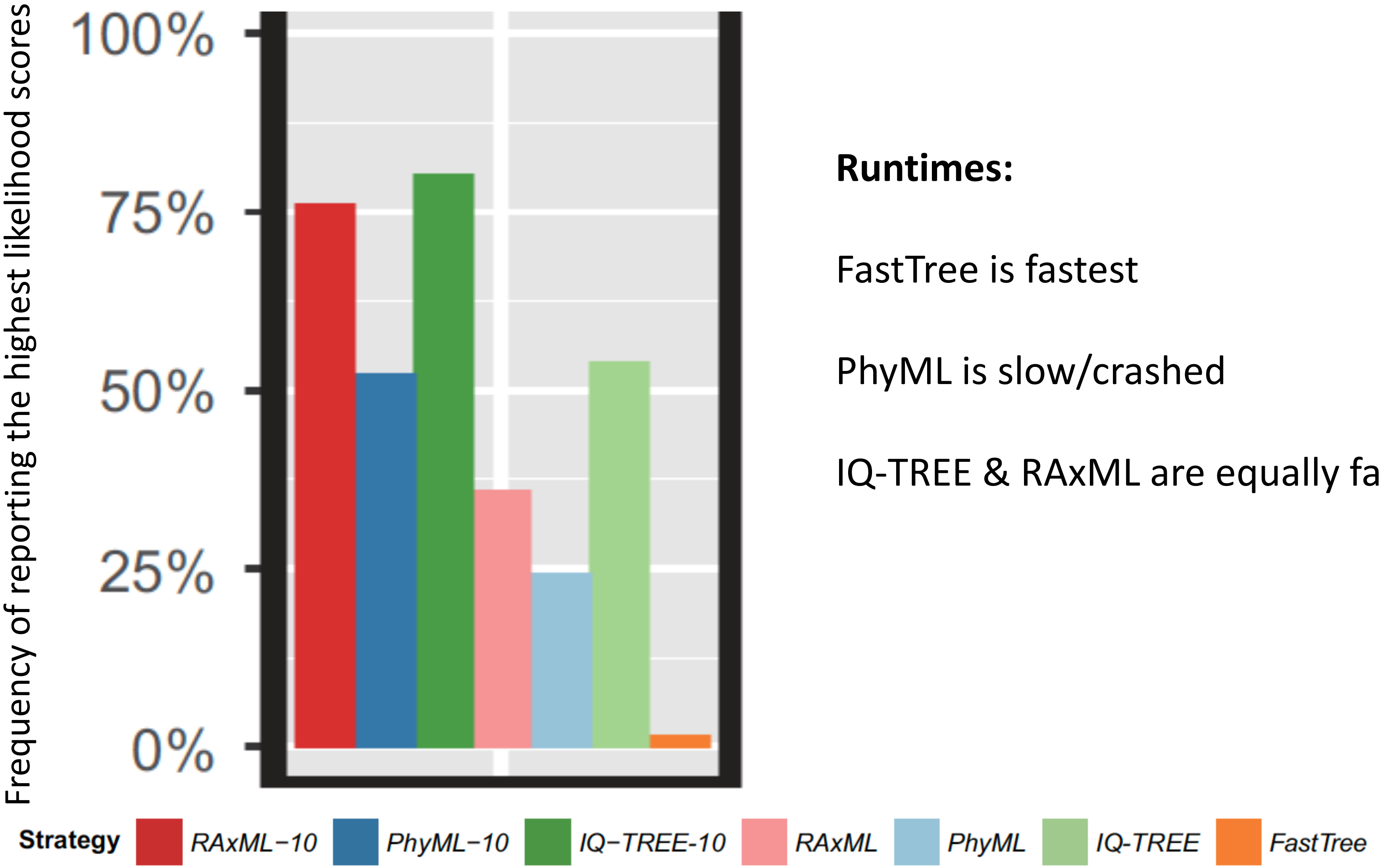
Metaheuristics:
Random restart, Iterated local search,
Evolution strategy

Lam-Tung Nguyen

Heiko Schmidt

Arndt von Haeseler

An independent benchmark by Zhou et al. (2018)



Runtimes:

FastTree is fastest

PhyML is slow/crashed

IQ-TREE & RAxML are equally fast

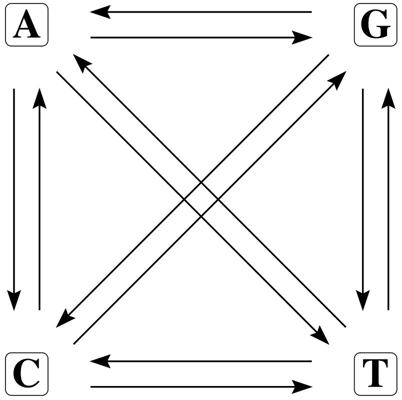
Step 2: Summary

Multiple sequence alignment

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

Model selection

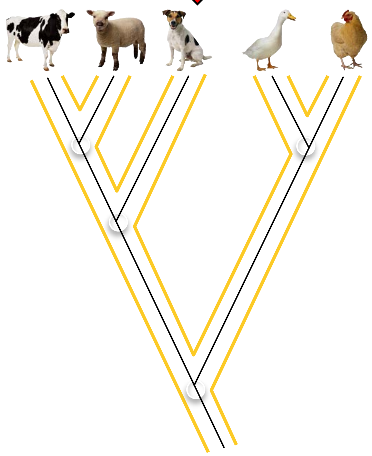
Substitution model



- IQ-TREE algorithm efficiently explores tree space

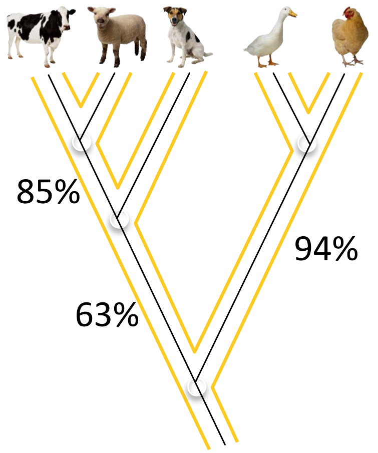
IQ-TREE (2015, 2020)

Tree reconstruction



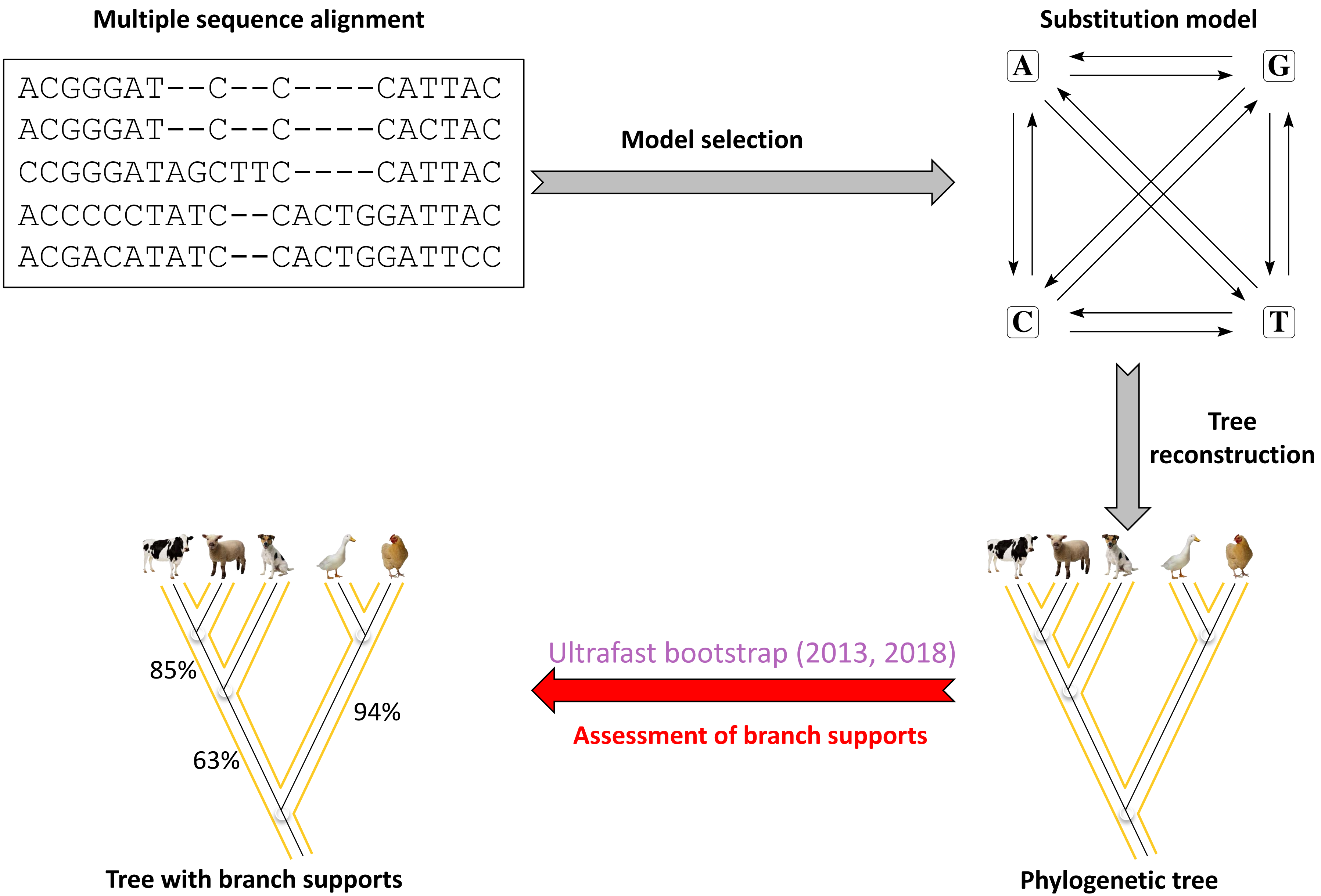
Phylogenetic tree

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

Substitution model

Tree reconstruction

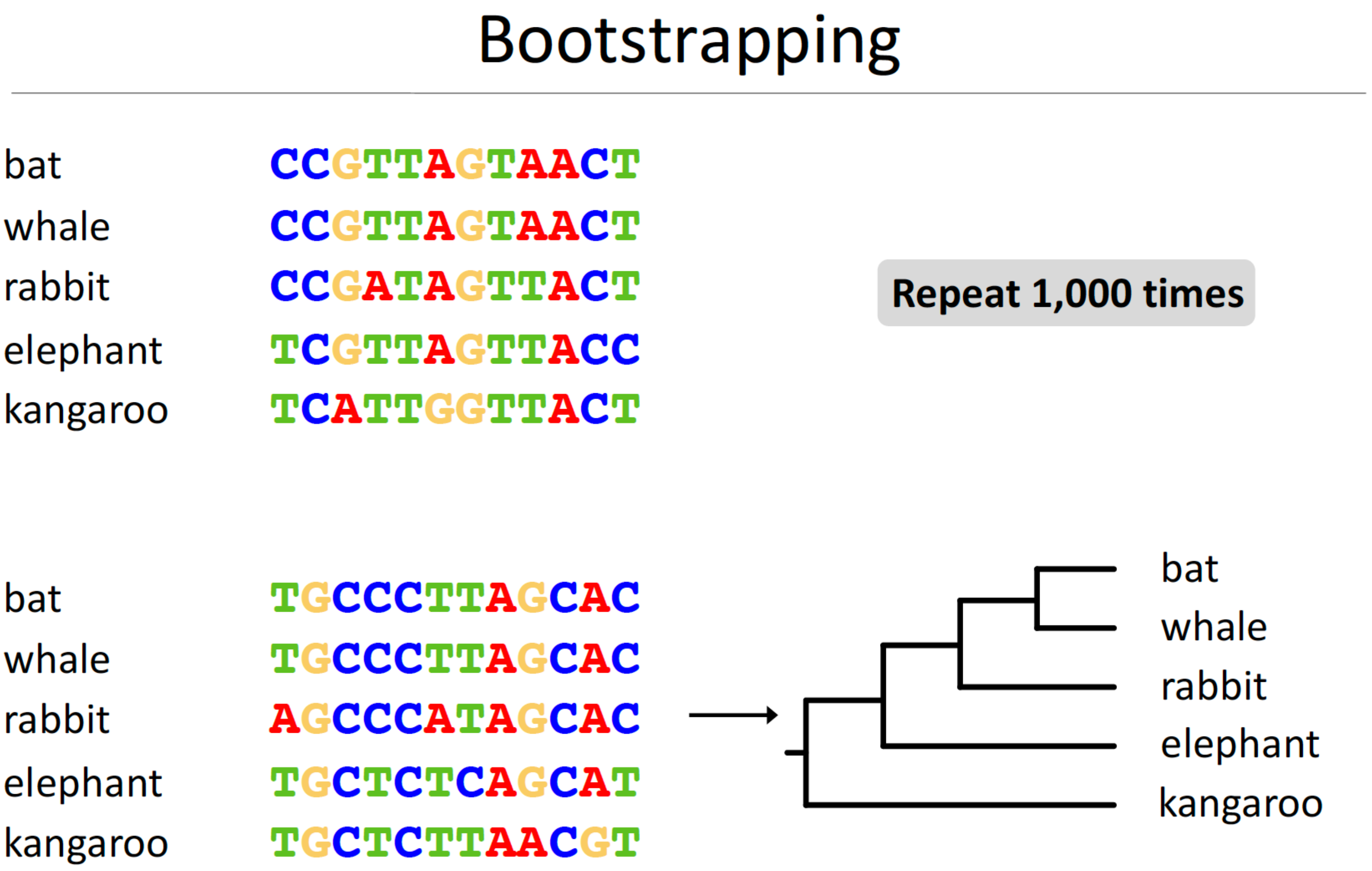
Phylogenetic tree

Ultrafast bootstrap (2013, 2018)

Assessment of branch supports

Tree with branch supports

Bootstrap: How reliable are branches of the tree?



Bootstrapping

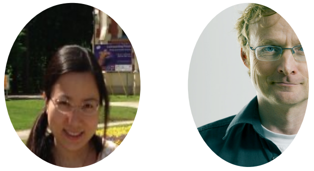
bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT

Repeat 1,000 times

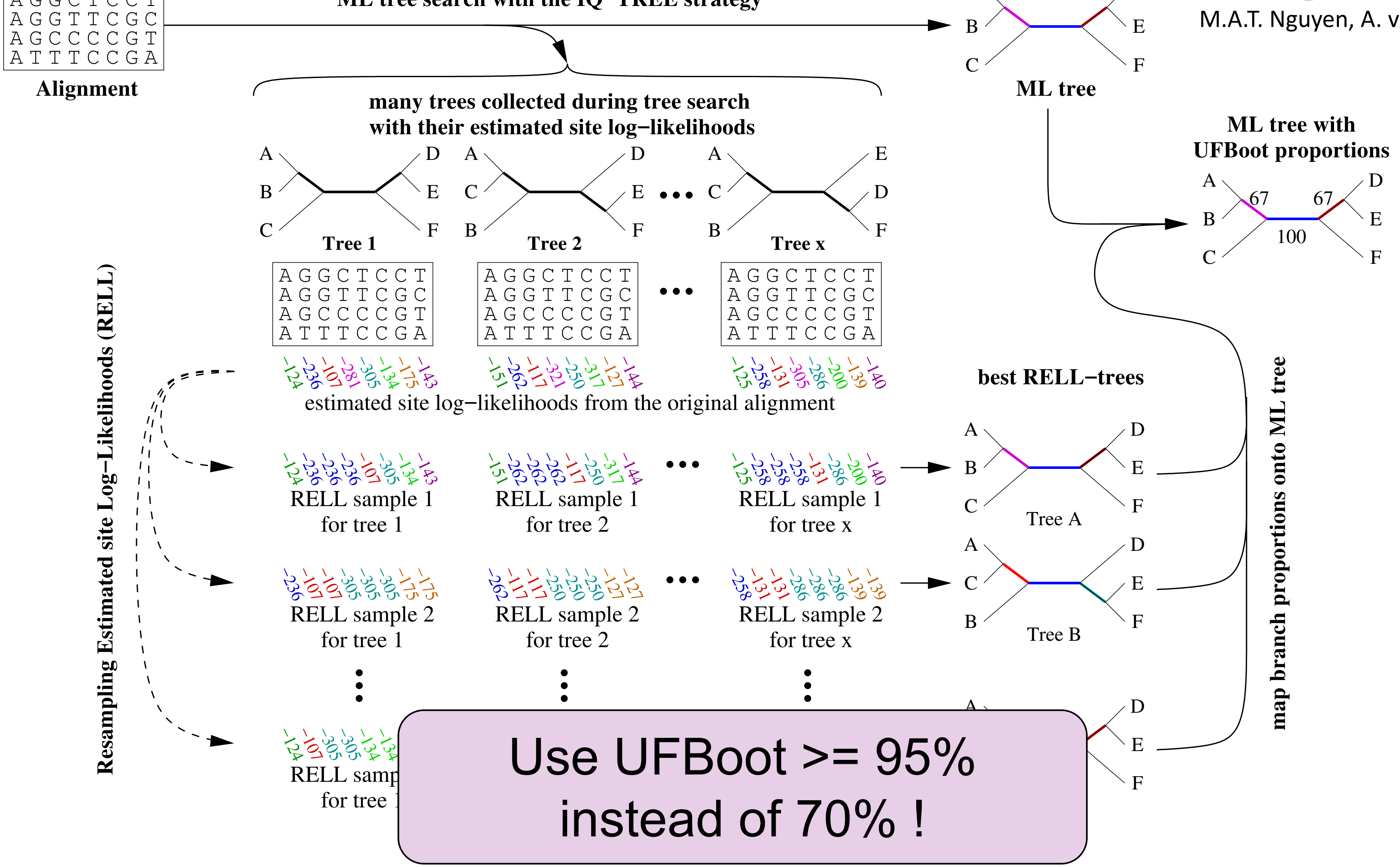
bat	TGCCCTTAGCAC
whale	TGCCCTTAGCAC
rabbit	AGCCCATAGCAC
elephant	TGCTCTCAGCAT
kangaroo	TGCTCTTAACGT

Bootstrap analysis is extremely time-consuming!

UFBoot: Ultrafast bootstrap approximation



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


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
Alignment
ML tree search with the IQ−TREE strategy
ML tree
many trees collected during tree search
with their estimated site log−likelihoods
Tree 1
Tree 2
Tree x
estimated site log−likelihoods from the original alignment
Resampling Estimated site Log−Likelihoods (RELL)
RELL sample 1 for tree 1
RELL sample 1 for tree 2
RELL sample 1 for tree x
RELL sample 2 for tree 1
RELL sample 2 for tree 2
RELL sample 2 for tree x
best RELL−trees
Tree A
Tree B
ML tree with UFBoot proportions
67
67
100
map branch proportions onto ML tree

Use UFBoot >= 95%
instead of 70% !

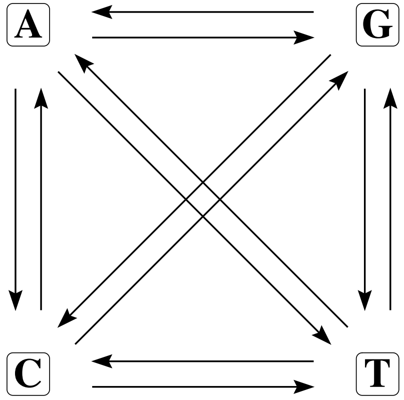
Step 3: Ultrafast bootstrap

Multiple sequence alignment

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

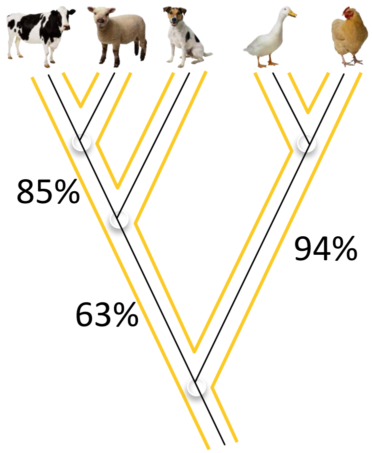
Model selection

Substitution model



Tree reconstruction

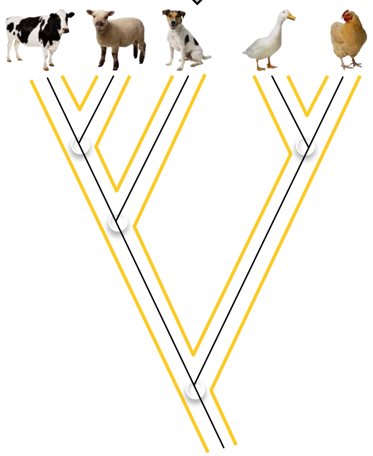
- Very fast alternative to standard bootstrap
- More direct interpretation of support values



Tree with branch supports

Ultrafast bootstrap (2013, 2018)

Assessment of branch supports



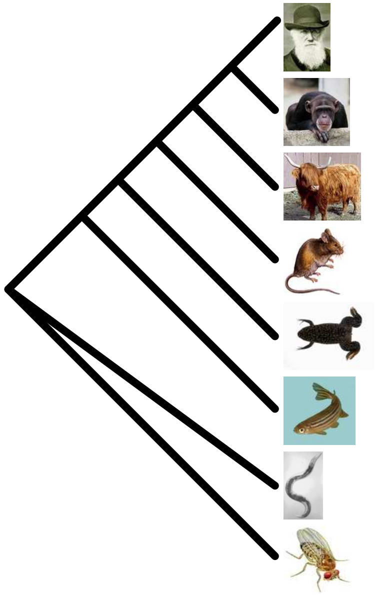
Phylogenetic tree

Genome-scale data: Concatenation 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	----------

Phylogenomic Inference



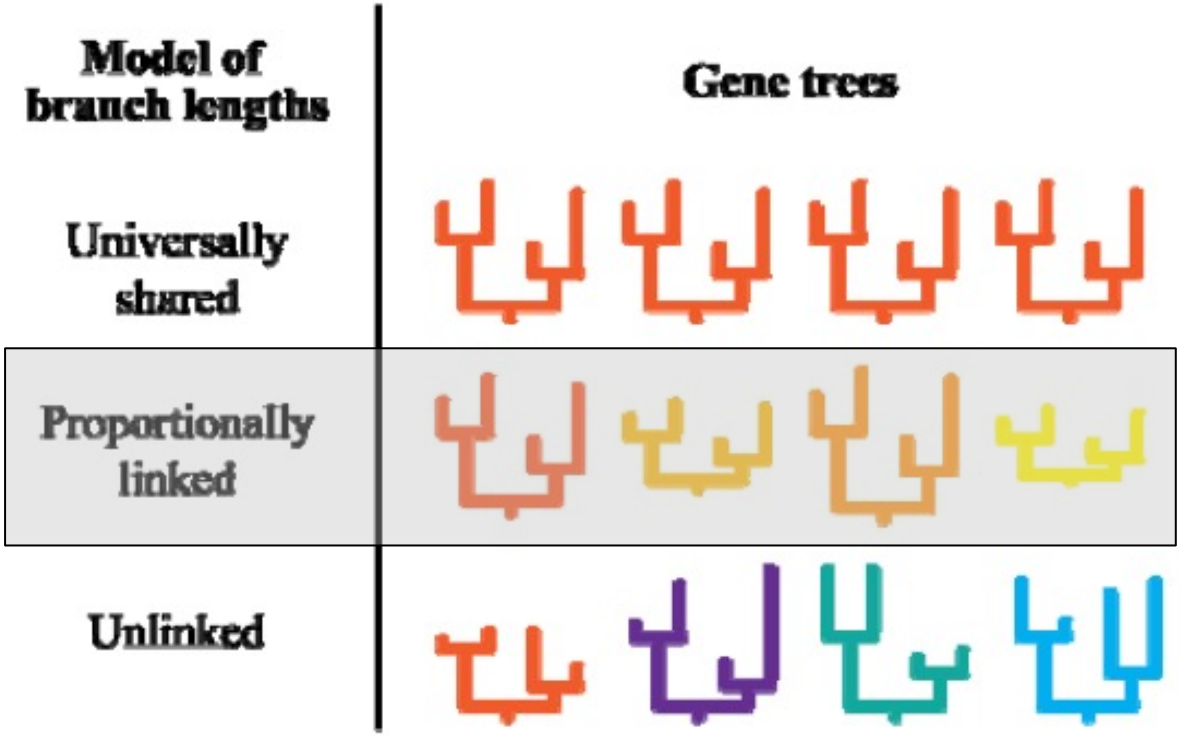
Species tree of life

Partition model

Supermatrix

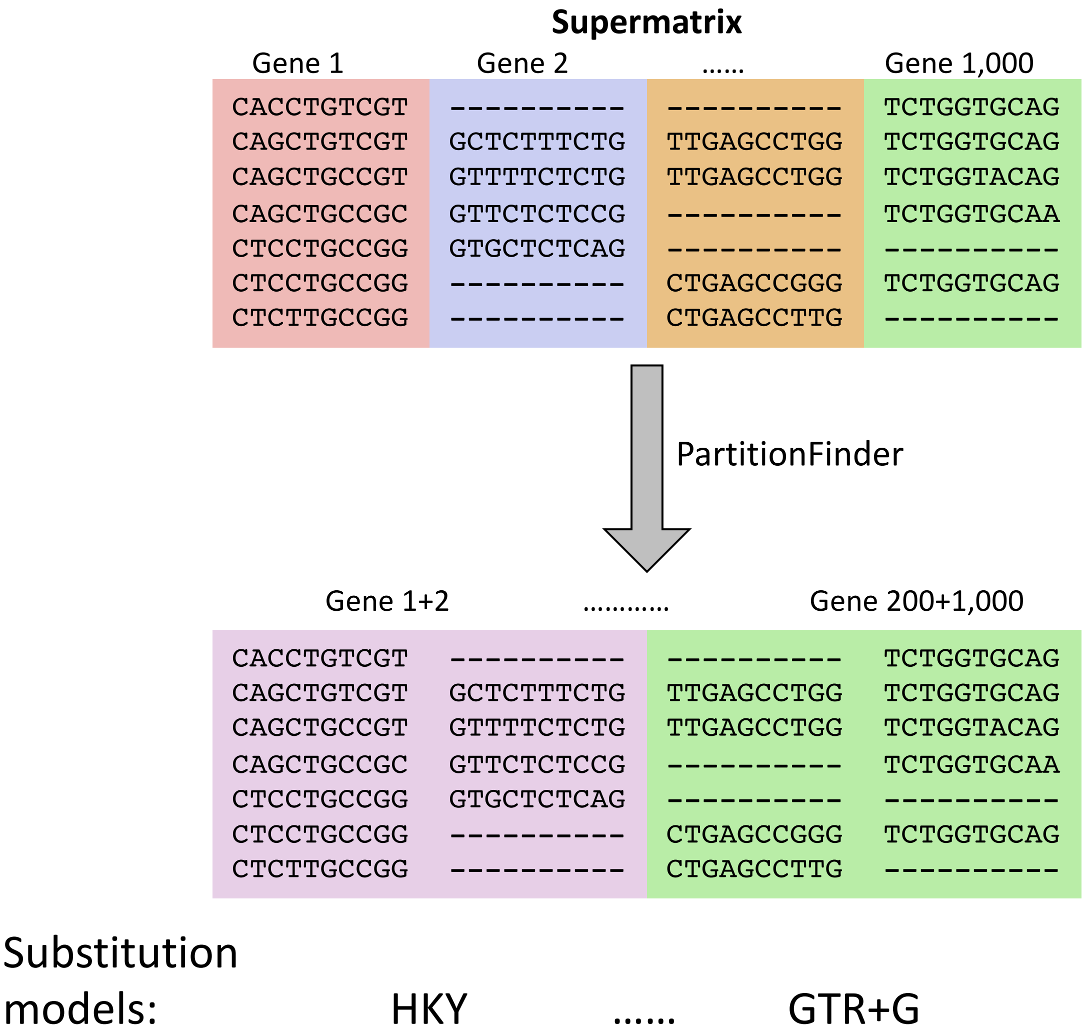
Gene 1	Gene 2	......	Gene 1,000
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG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

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



Recommended for typical analysis (Duchene et al. 2020)

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	----------

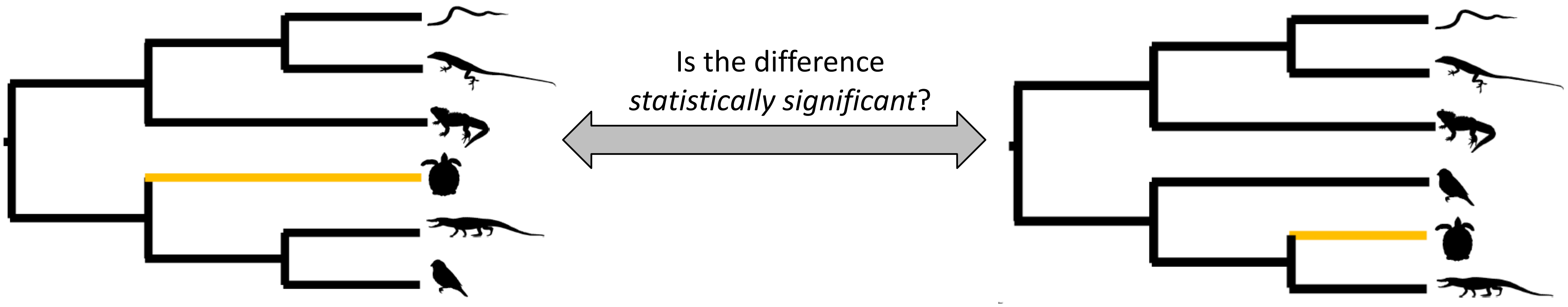
Gene 1+2	....	Gene 200+1,000	
CACCTGTCGT	----------	----------	TCTGGTGCAG
CAGCTGTCGT	GCTCTTTCTG	TTGAGCCTGG	TCTGGTGCAG
CAGCTGCCGT	GTTTTCTCTG	TTGAGCCTGG	TCTGGTACAG
CAGCTGCCGC	GTTCTCTCCG	----------	TCTGGTGCAA
CTCCTGCCGG	GTGCTCTCAG	----------	----------
CTCCTGCCGG	----------	CTGAGCCGGG	TCTGGTGCAG
CTCTTGCCGG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

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

PartitionFinder algorithm (Lanfear et al. 2012): Greedy algorithm: repeatedly merge the ‘best’ pairs of partitions until AIC/BIC is not improved.

Relaxed clustering algorithm (Lanfear et al. 2014): Only examine the top k% of most “promising” pairs when merging them.

Tree topology tests

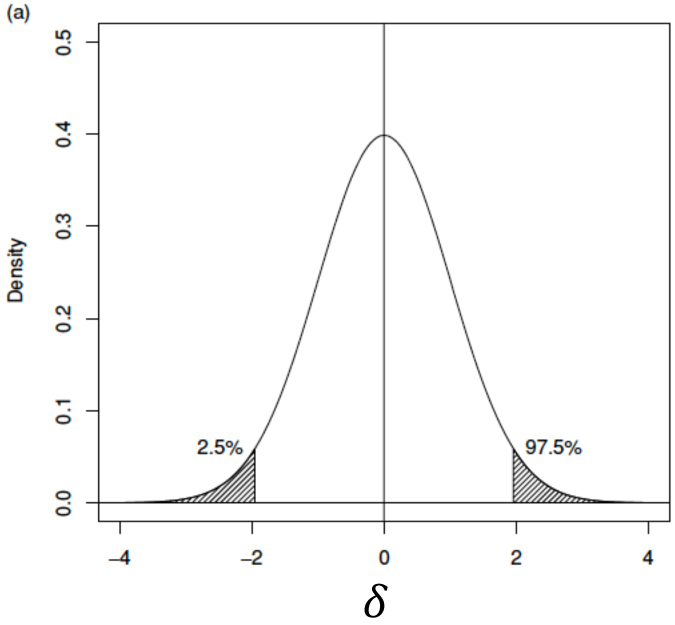


Is the difference *statistically significant*?

Testing two trees (Kishino & Hasegawa, 1989):

Is $\delta = \log\big(likelihood(T_1)\big) - \log\big(likelihood(T_0)\big)$ significantly different from zero?

1. Generate distribution of δ from many “random” data (e.g. by 1000 bootstrap resampling).
2. Compare the statistic between original and random data to obtain *p-value*.
3. If $p\text{-value} < 0.05$: YES! two trees are significantly different.
4. If p-value >= 0.05: NO! they are not.

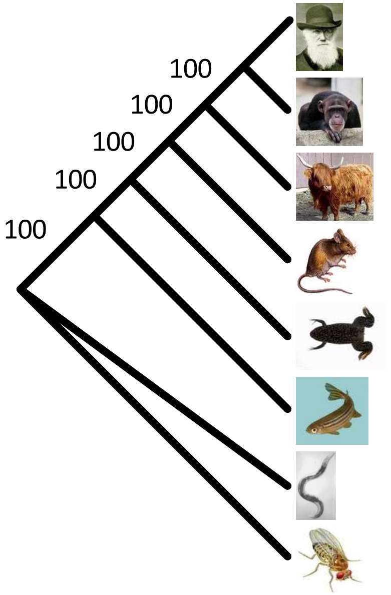


Concatenation methods: Limitation

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

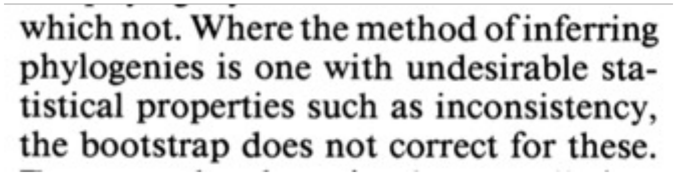
Bootstrap supports and Bayesian posteriors tend to 100% as #genes increases!

Concatenation assumes a single tree across all loci

⇩

Potential *systematic bias*

Felsenstein (1985):

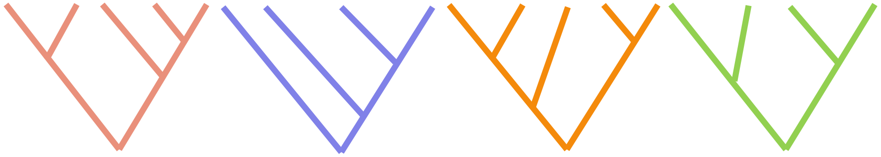


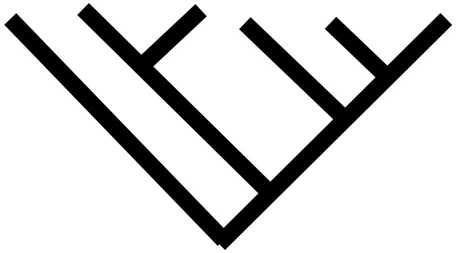
which not. Where the method of inferring phylogenies is one with undesirable statistical properties such as inconsistency, the bootstrap does not correct for these.

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	----------





Species tree

Gene Concordance Factor (gCF):
How often a branch in species tree is found among gene trees?
0% ≤ gCF ≤ 100%

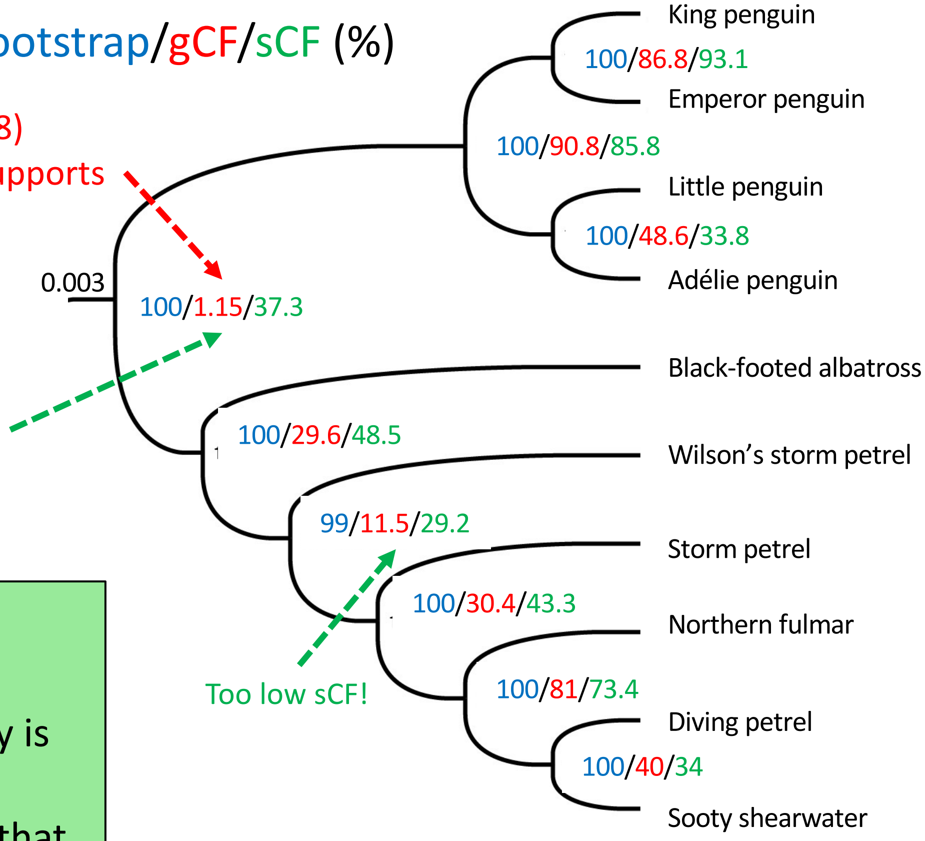
Site Concordance Factor (sCF):
How often a branch is "supported" by alignment sites?
33.3% ≨ sCF ≤ 100%

An example birds data set (Reddy et al., 2017)

Bootstrap/gCF/sCF (%)

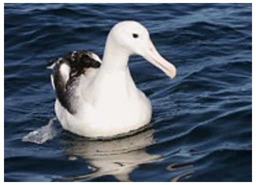
Only 1 (of 88) gene tree supports this branch!

- 131 sites support this branch
- 105 sites support NNI branch 1
- 114 sites support NNI branch 2



Too low sCF!

Penguins



Tubenoses

Felsenstein (1985): a difference of 20 sites favouring one topology is enough to give 100% bootstrap support for that one topology!

- gCF and sCF are useful when bootstrap supports reach 100%.
- CAUTION when gCF ~ 0% or sCF ~ 33%, even if BS ~ 100%.
- GREAT when gCF and sCF > 50%.