

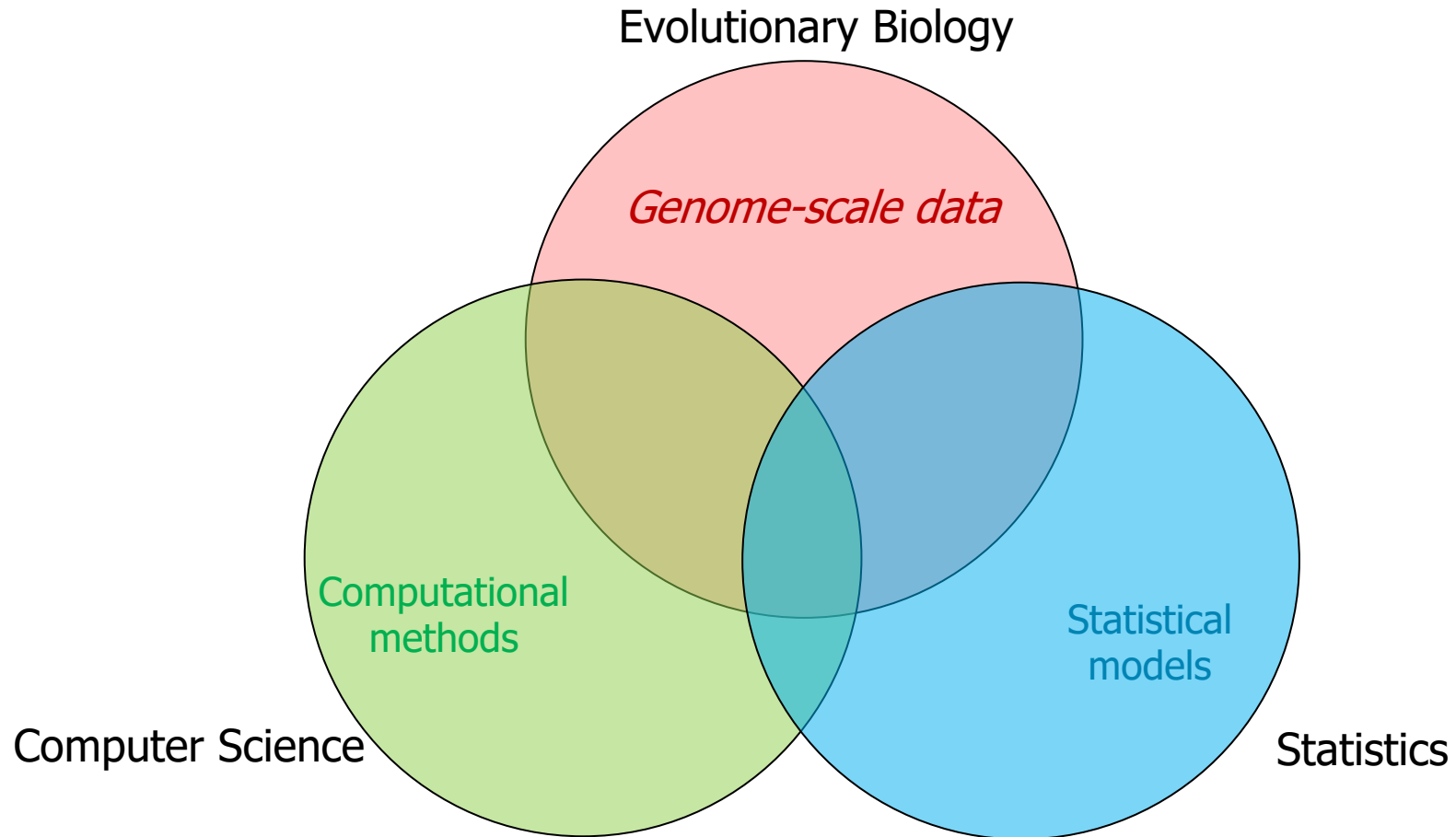
Addressing some computational and modeling challenges in phylogenomic inference

Minh Q. Bui

Australian National University

Computational Phylogenomics Lab @ ANU

"To enable evolutionary research in the genomic era"



<https://anu-phylogenomics.github.io>

Methods for phylogenetic reconstruction

Multiple Sequence Alignment (MSA)

Rhesus	TCC-CATTGTTCCACACAATGCC
Chimp	CGCTCATTCC--CCACACAACGCC
Human	CTCTC--TTTTCCACCCTCCGCT
Gorilla	CTCTCATTA-TTTTCA--ACGCT
Orang	CCCT-ATTGTTCC-CACCACACT

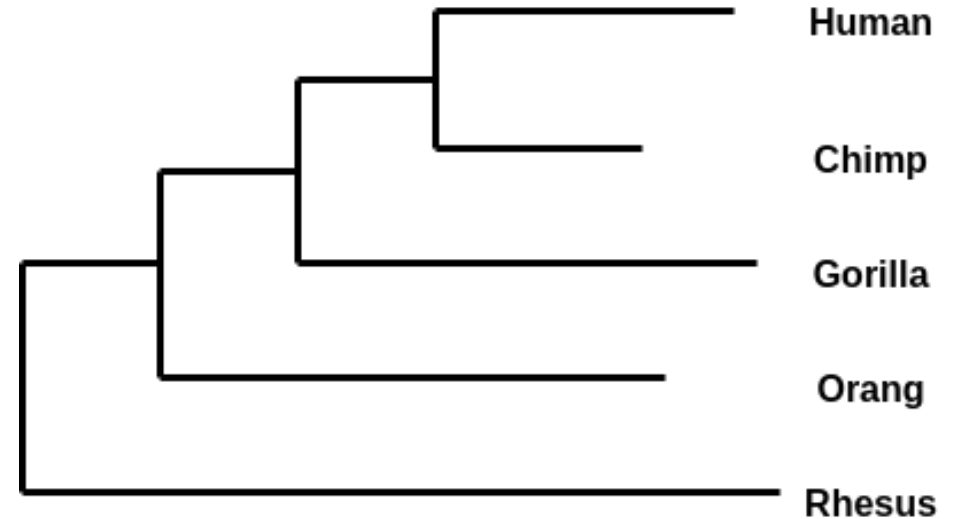
Maximum parsimony

Distance based

Maximum likelihood

Bayesian

Phylogenetic Tree



Challenges in phylogenetic inference from genomic data

Multiple sequence alignments

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

20,000 publications using genome-scale data until 2015!

COVID-19 data: n > 5 million!

Computational Challenges

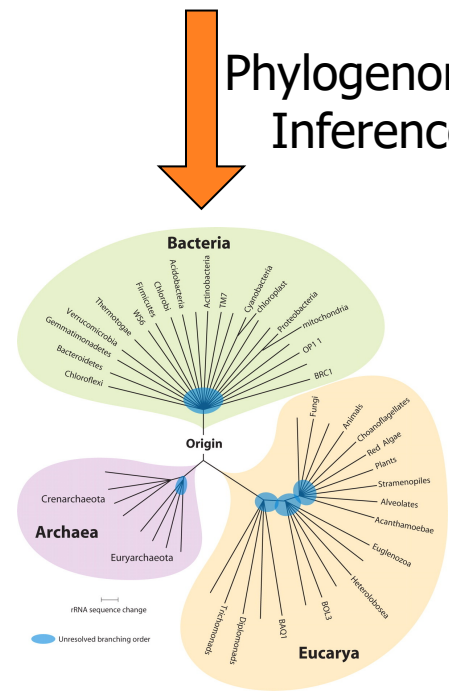
Phylogenomic Inference

Modeling Challenges

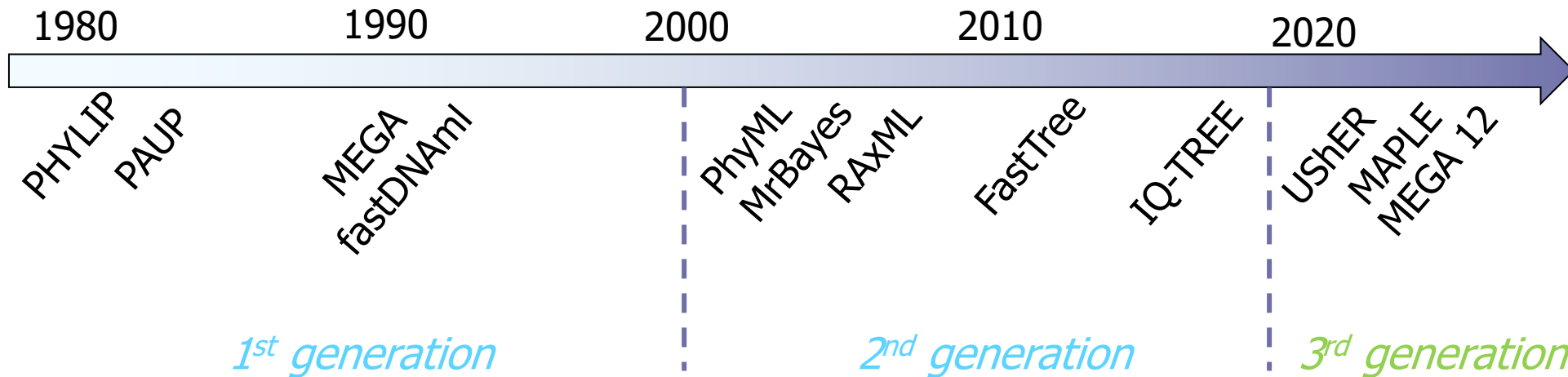
Part 1: Methods for pathogen data

Part 2: Models for protein phylogeny

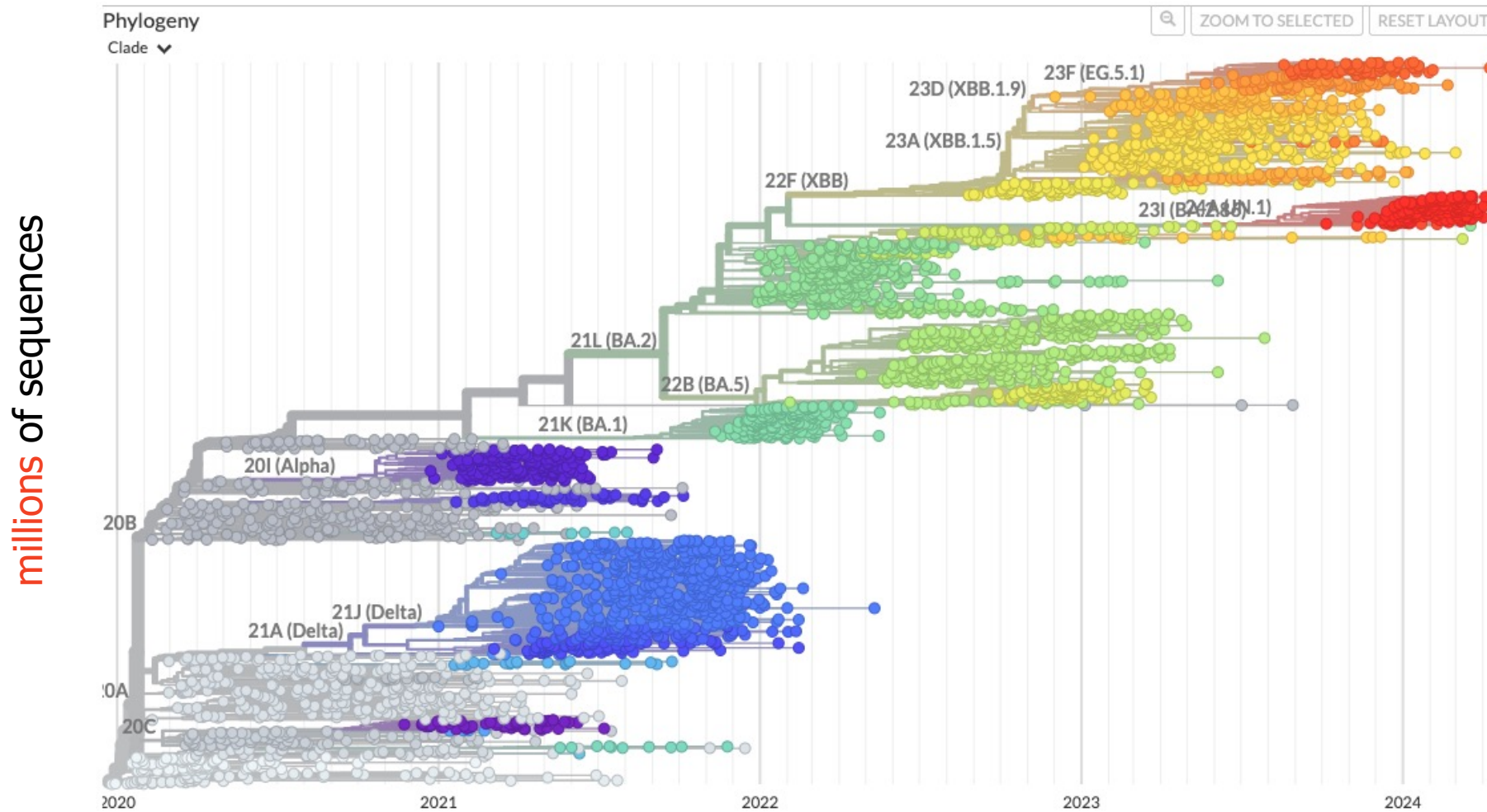
Part 3: Machine learning for protein model selection



Generations of phylogenetic tools



SARS-CoV-2 era



- UShER
- matOptimize

- MAPLE
- CMAPLE



Nicola De Maio

nextstrain.org (May 9th 2024)

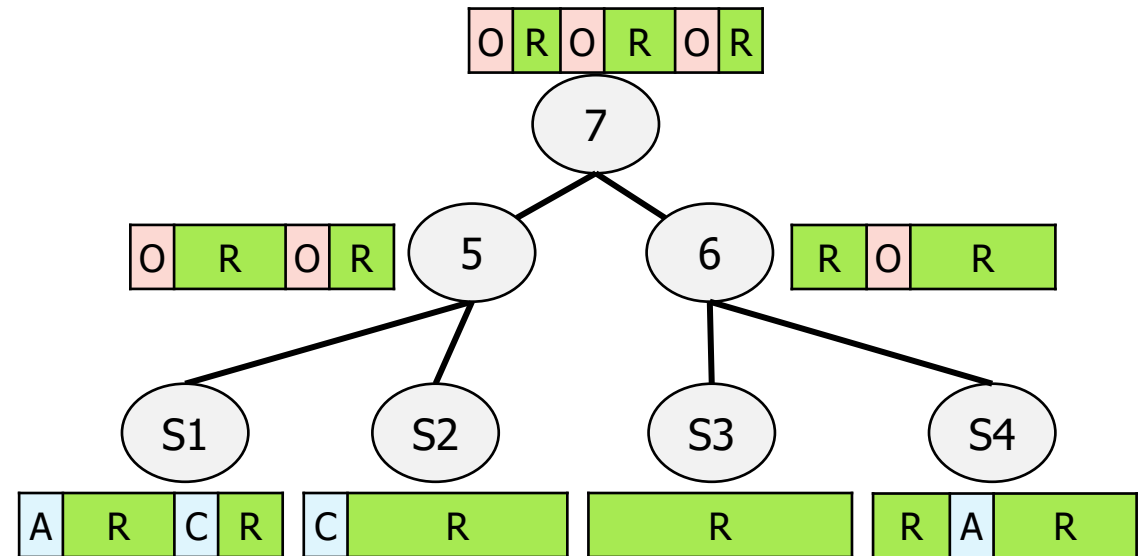
MAPLE: M^AXimum P^Arsimonious L^Ikelihood E^Stimation

- **High sequence similarity:** MAPLE format stores sequence differences to a reference.

```
>Ref
GTCCCACAGCCAGGA
>S1
A 1
C 13
>S2
C 1
>S3
>S4
A 4
```

- 27.84 GB FASTA file -> 224.6 MB MAPLE file (124-fold reduction)

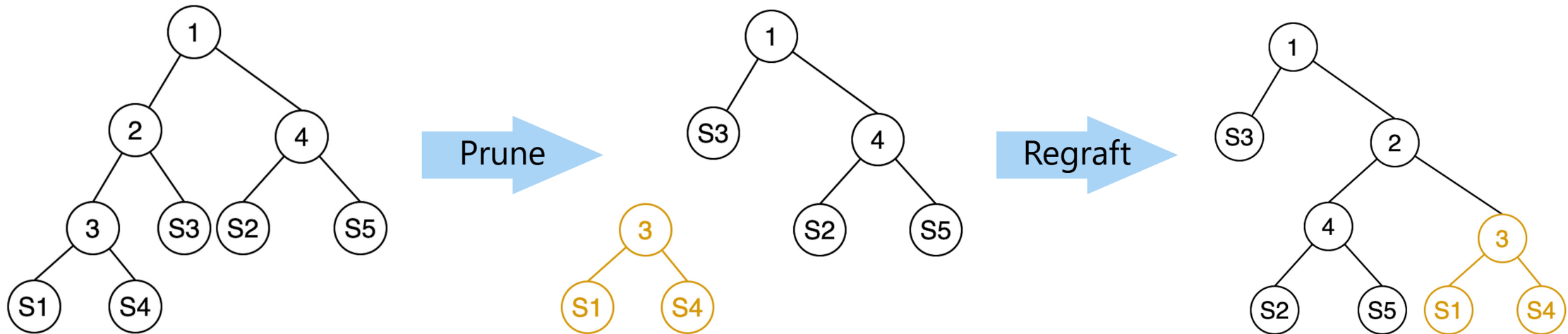
- **Tailored Felsenstein's pruning algorithm** to compute *approximate* likelihoods.



- Compute likelihoods at each ancestral node for ~2.7 O-positions instead of all 30,000 positions

MAPLE: Fast tree search algorithm

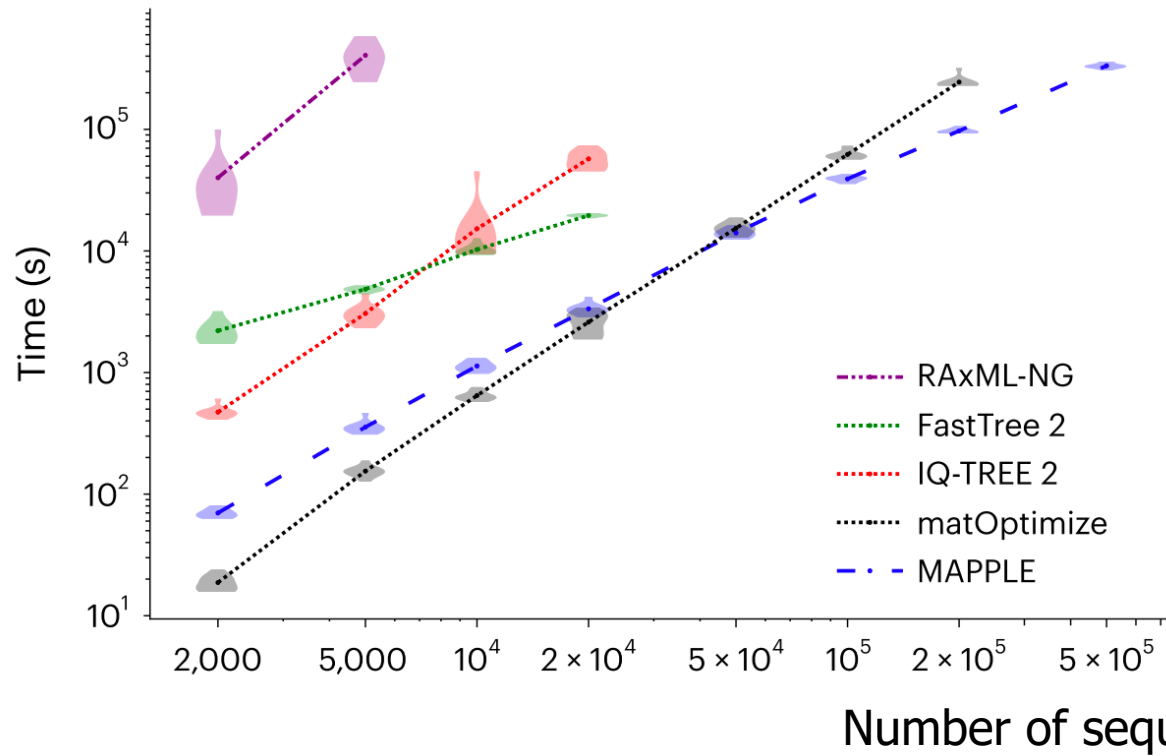
Subtree pruning and regrafting (SPR)



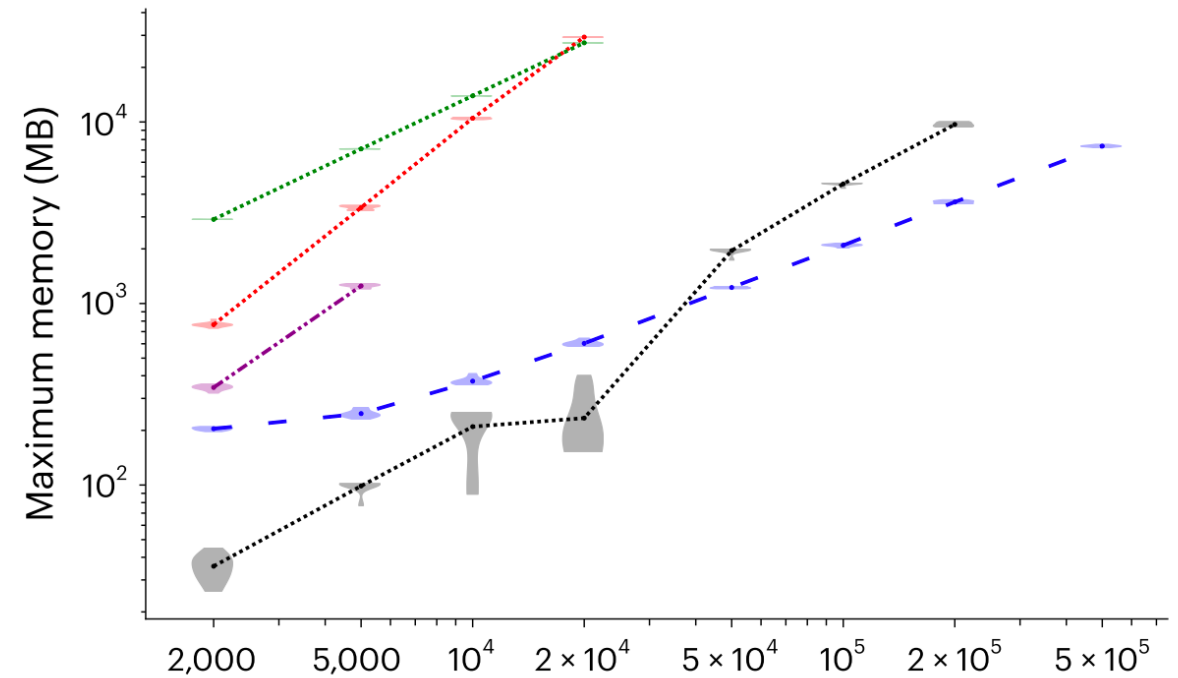
MAPLE stops moving a subtree into further branches if likelihood decreases twice consecutively

MAPLE's performance vs. existing tools

a

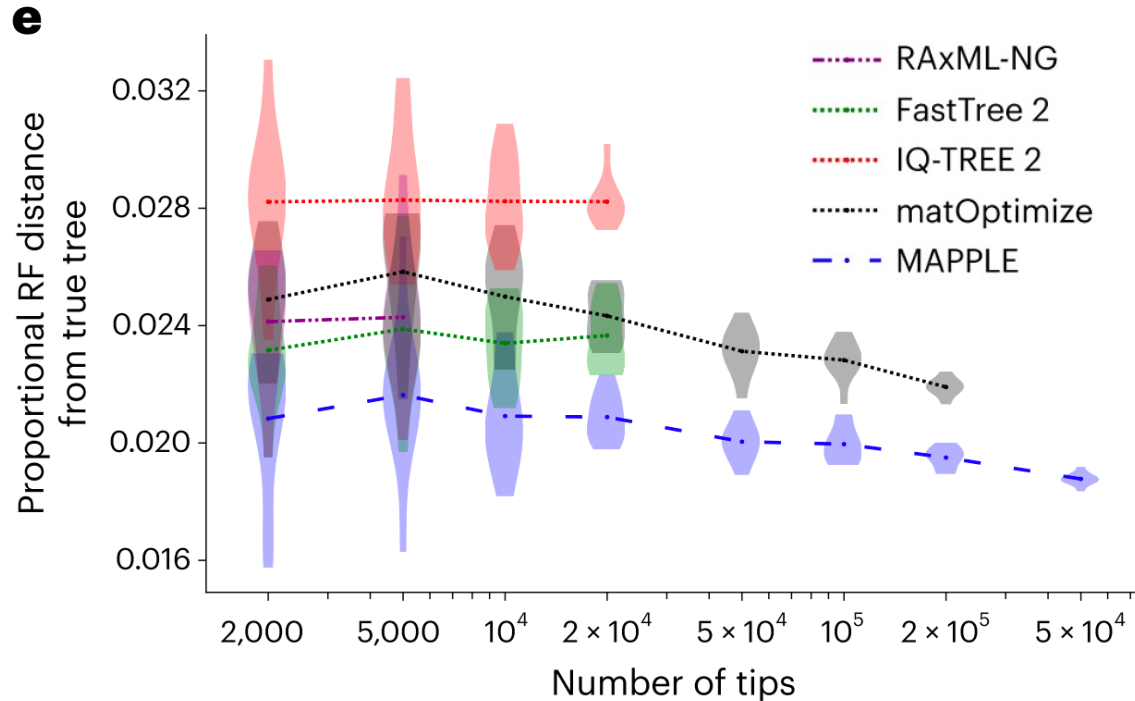


b

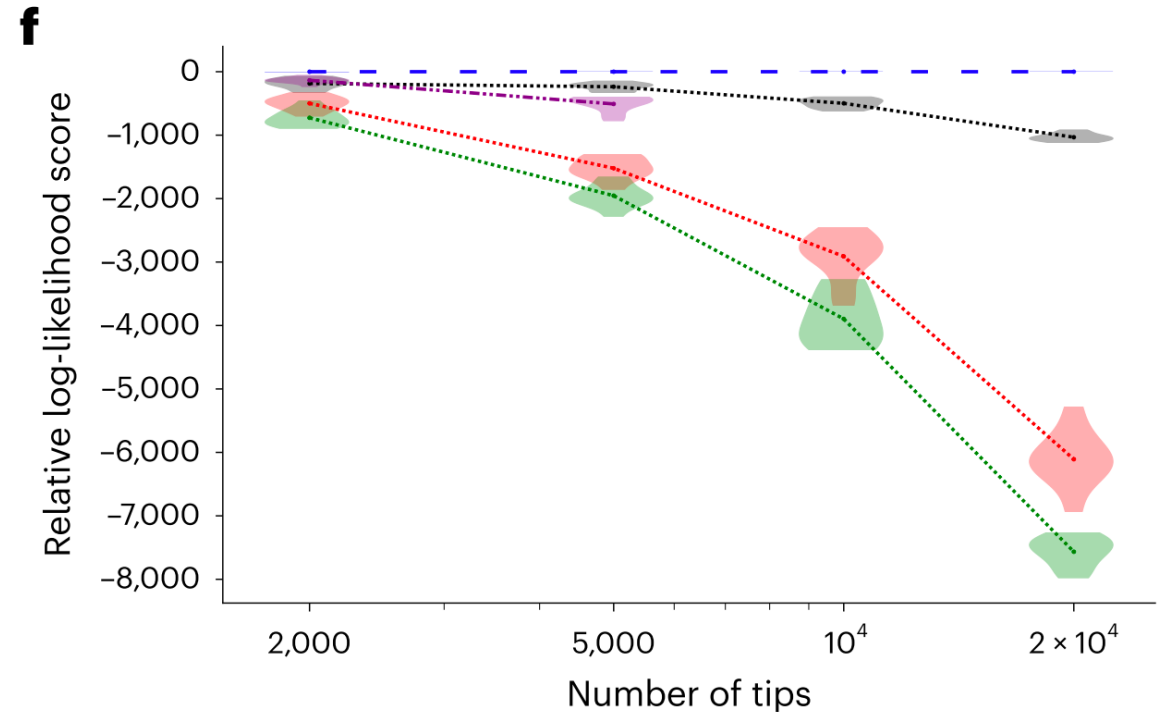


MAPLE is more accurate than existing tools

Simulated data



Real SARS-CoV-2 data



[Maximum likelihood pandemic-scale phylogenetics](#)

N. De Maio, P. Kalaghatgi, Y. Turakhia, R. Corbett-Detig, [B.Q. Minh](#), N. Goldman

Nature Genetics 55:746–752 (2023)



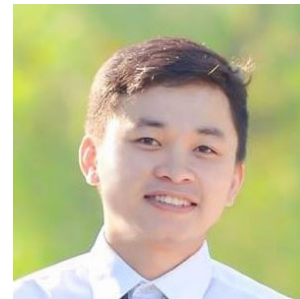
From MAPLE to CMAPLE

- MAPLE “only” works up to 500,000 sequences.
- CMAPLE is a complete rewrite of MAPLE (Python) into C++.
 - A lot of low-level code optimizations, e.g., minimize CPU-cache misses, specialized memory allocator.
- New substitution models for protein.
- Branch supports with Shimodaira-Hasegawa approximate likelihood ratio test.

CMAPLE: Efficient phylogenetic inference in the pandemic era 

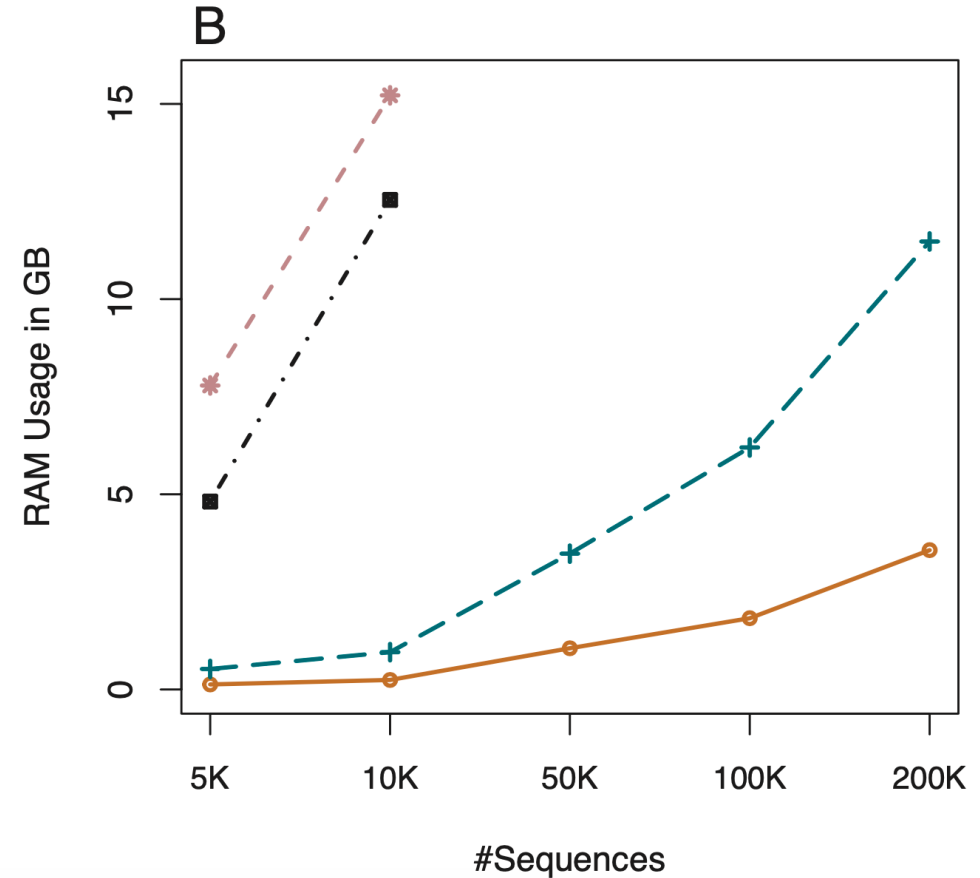
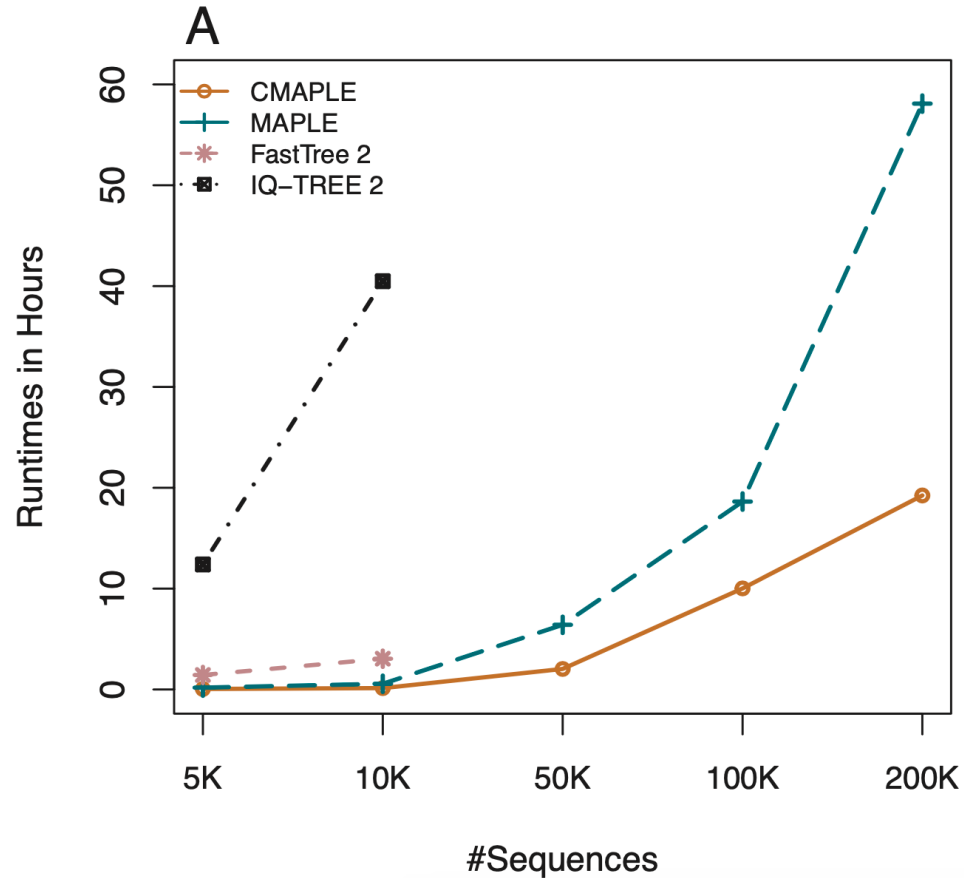
N. Ly-Trong, C. Bielow, N. De Maio, B.Q. Minh

Molecular Biology and Evolution msae134 (2024)



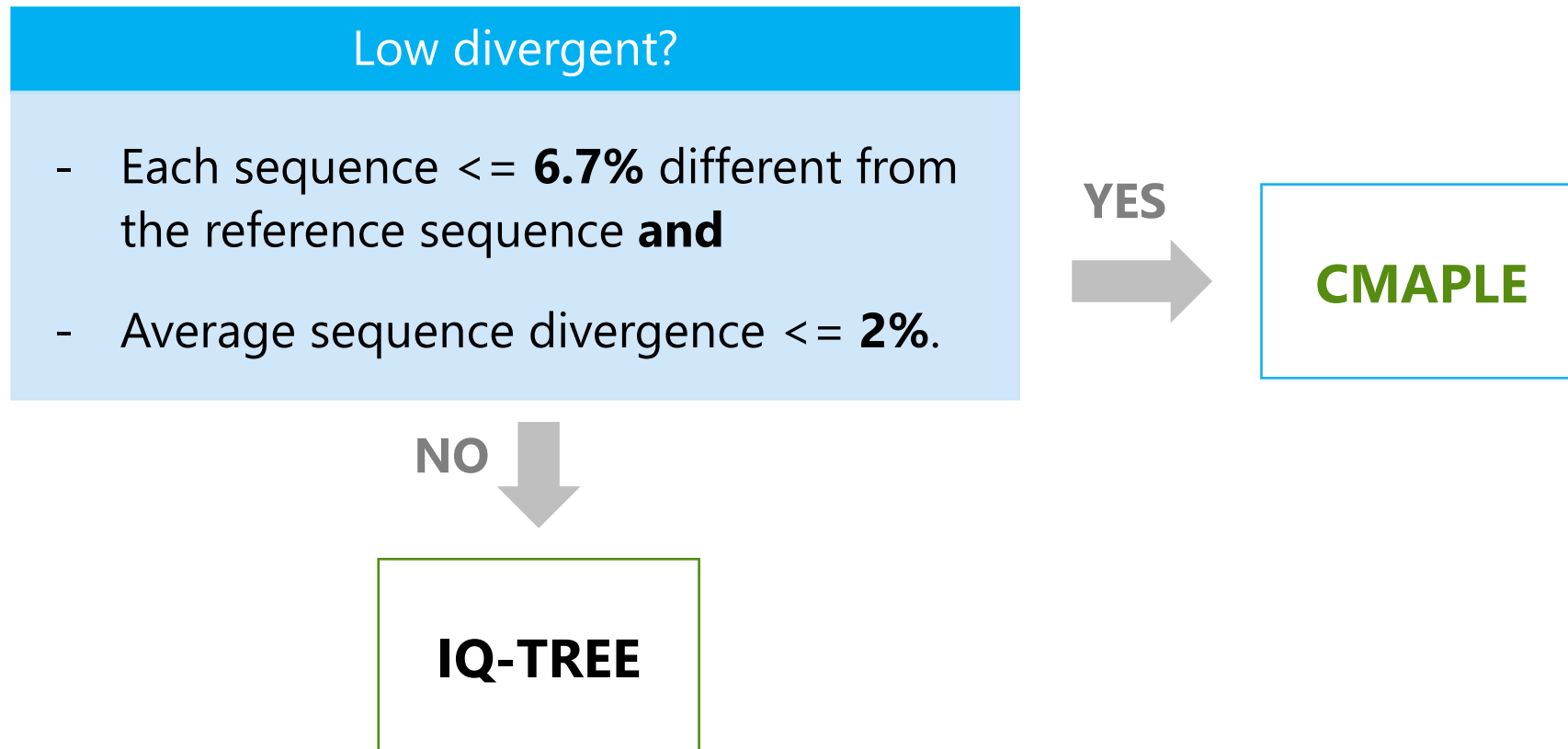
Nhan Ly-Trong

Benchmark CMAPLE




- CMAPLE takes 11 days and 15.4 GB RAM to infer a tree of 1M SARS-CoV-2 “from scratch”
- CMAPLE takes 14 min to place 10K “new” sequences into a 500,000-tip tree (online phylogenetics)


CMAPLE API is integrated in IQ-TREE (--pathogen option)



Recent developments

Rate variation and recurrent sequence errors in pandemic-scale phylogenetics 

N. De Maio, M. Willemsen, Z. Guo, A. Saha, M. Hunt, [N. Ly-Trong](#), [B.Q. Minh](#), Z. Iqbal, N. Goldman
bioRxiv (2024)

This is SPRTA: assessing phylogenetic confidence at pandemic scales 

N. De Maio, [N. Ly-Trong](#), [B.Q. Minh](#), N. Goldman
bioRxiv (2024)

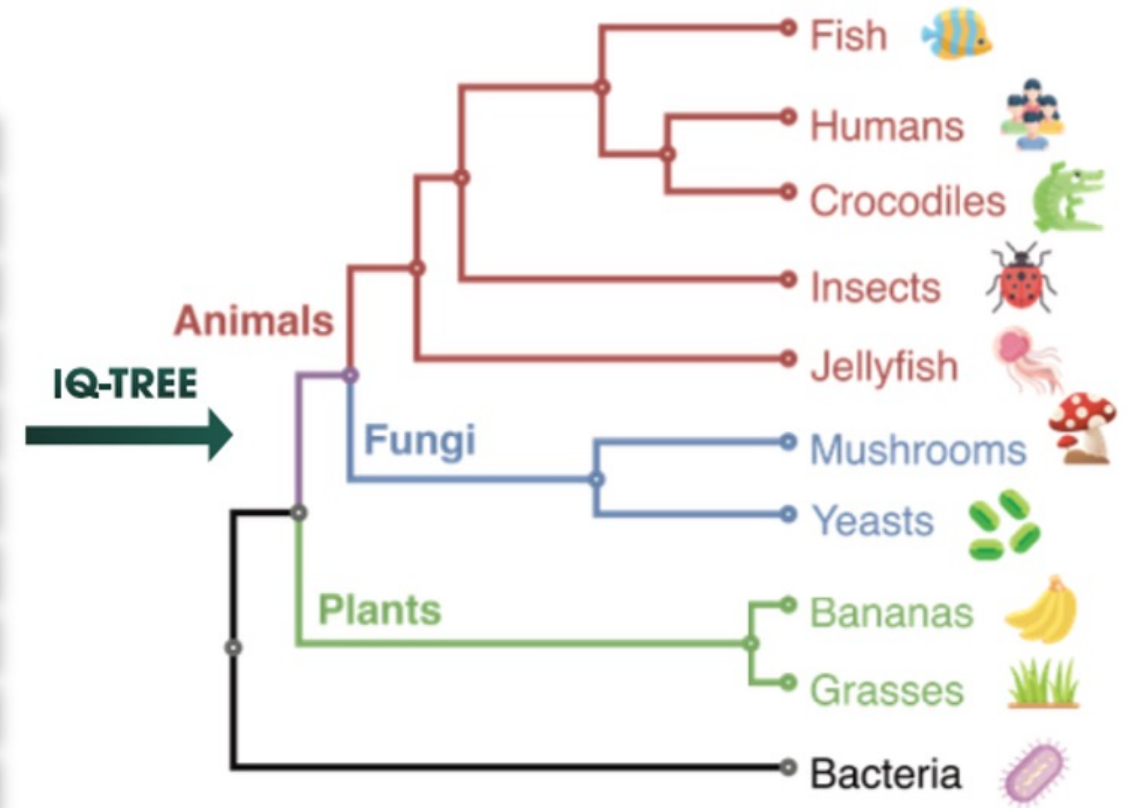
Others

- Integrated (C)MAPLE into nextstrain.org
- Integrate (C)MAPLE into BEAST 2
- Parallelize tree search algorithm

Part 2: Models for protein phylogeny

Protein Sequence Alignment

Humans	RMHYD...LIHTS...LARVV...FTLEYGYS...EYRS
Crocodiles	RMHYD...LIHTS...LARRV...FTLEYGYS...EYRS
Fish	RMHYD...LIHTS...LARRV...FTQ EYGTS...EYRS
Insects	RMHYD...LIHTS...LARRV...FTLEYGTS...EYYS
Jellyfish	RMHYD...LIHTS...LARRV...FTLEYGTS...EYRS
Mushrooms	RMHYD...L LNRS...LARRV...FTLEYGTS...EYRS
Yeasts	RMHYD...L INRS...LARRV...FTLEYGTS...EYRS
Bananas	RMHYD...LIHTS...LASRV...FTLRYGTS...EYRS
Grasses	RMHYD...LIHTS...LASRV...FTLEYGTS...EYRS
Bacteria	RMHHD...LIHTS...LARRV...FTLEYGTS...EYRS



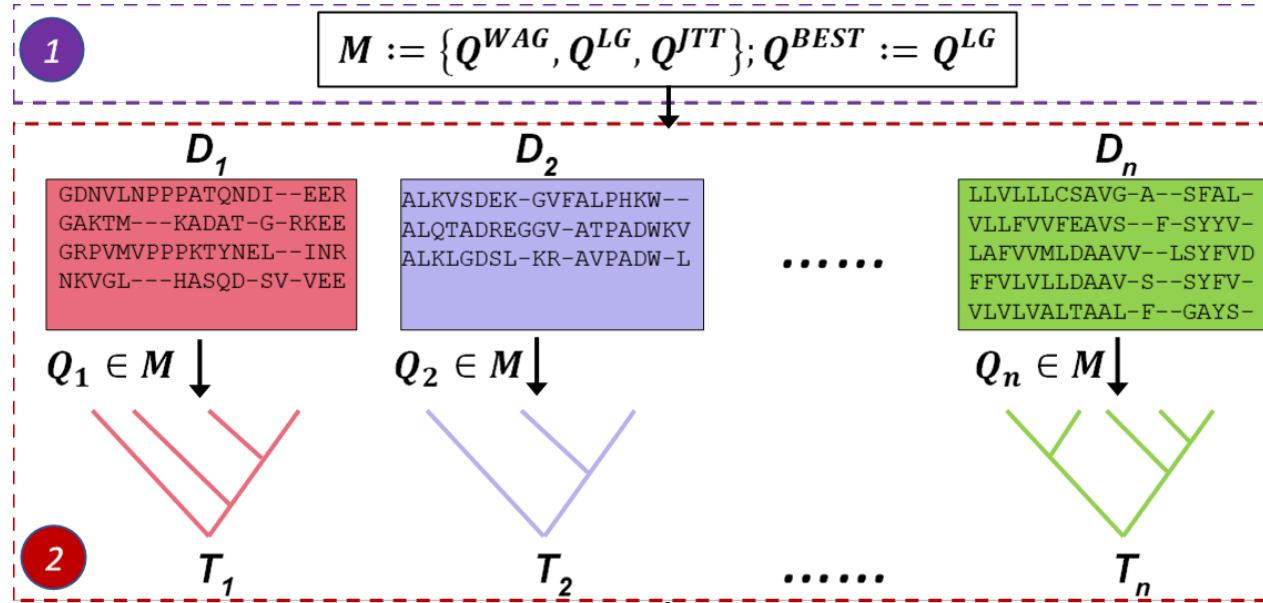
Datasets

Data set	References	Seqs	Sites	Loci	Training	Testing
Pfam	El-Gebali et al. (2019)	1,150,099	3,433,343	13,308	6654	6654
Plant	Ran et al. (2018)	38	432,014	1308	1000	308
Bird	Jarvis et al. (2015)	52	4,519,041	8295	1000 x 2	6295
Mammal	Wu et al. (2018)	90	3,050,199	5162	1000 x 2	3162
Insect	Misof et al. (2014)	144	595,033	2868	1000	1868
Yeast	Shen et al. (2018)	343	1,162,805	2408	1000 x 100 seqs	1408



Robert Lanfear

QMaker: Model training



QMaker: Fast and accurate method to estimate empirical models of protein evolution 

B.Q. Minh, C. Cao Dang, L.S. Vinh, R. Lanfear

Systematic Biology 70:1046–1060 (2021)

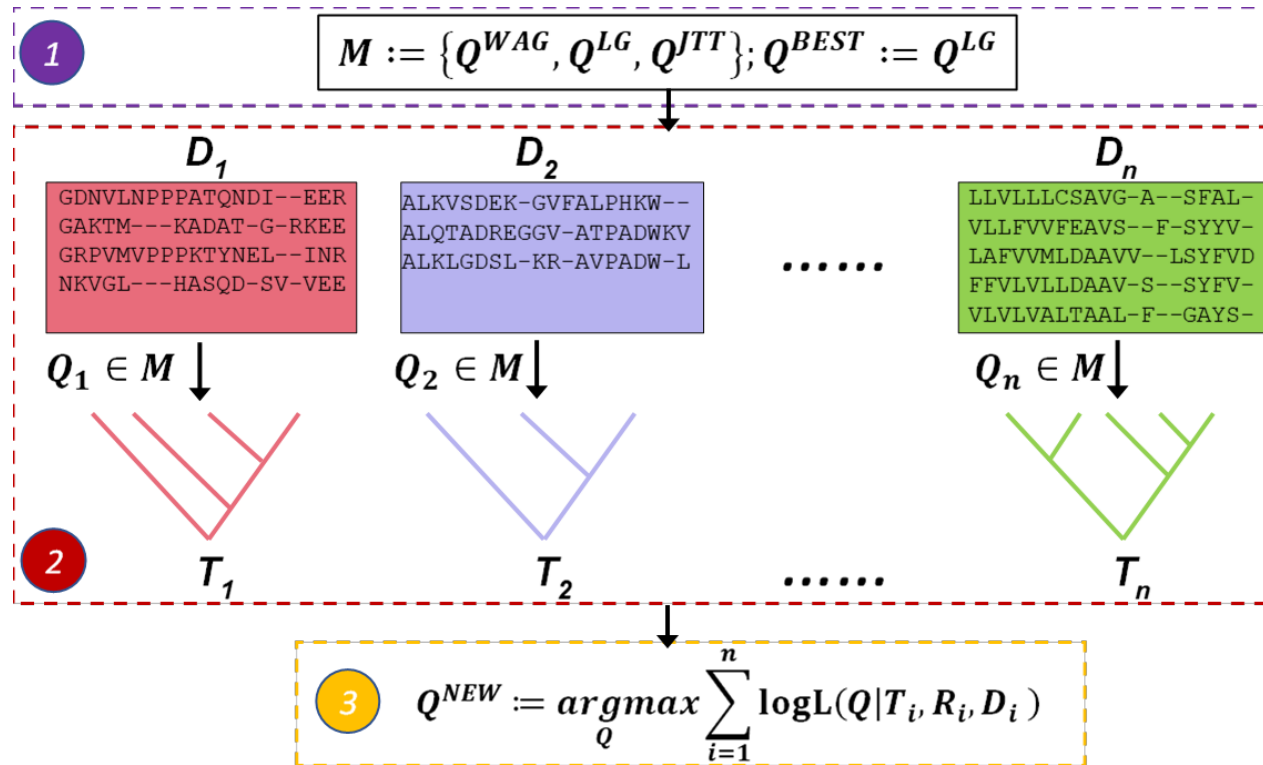
Altmetric

6

Citations

67

QMaker: Model training



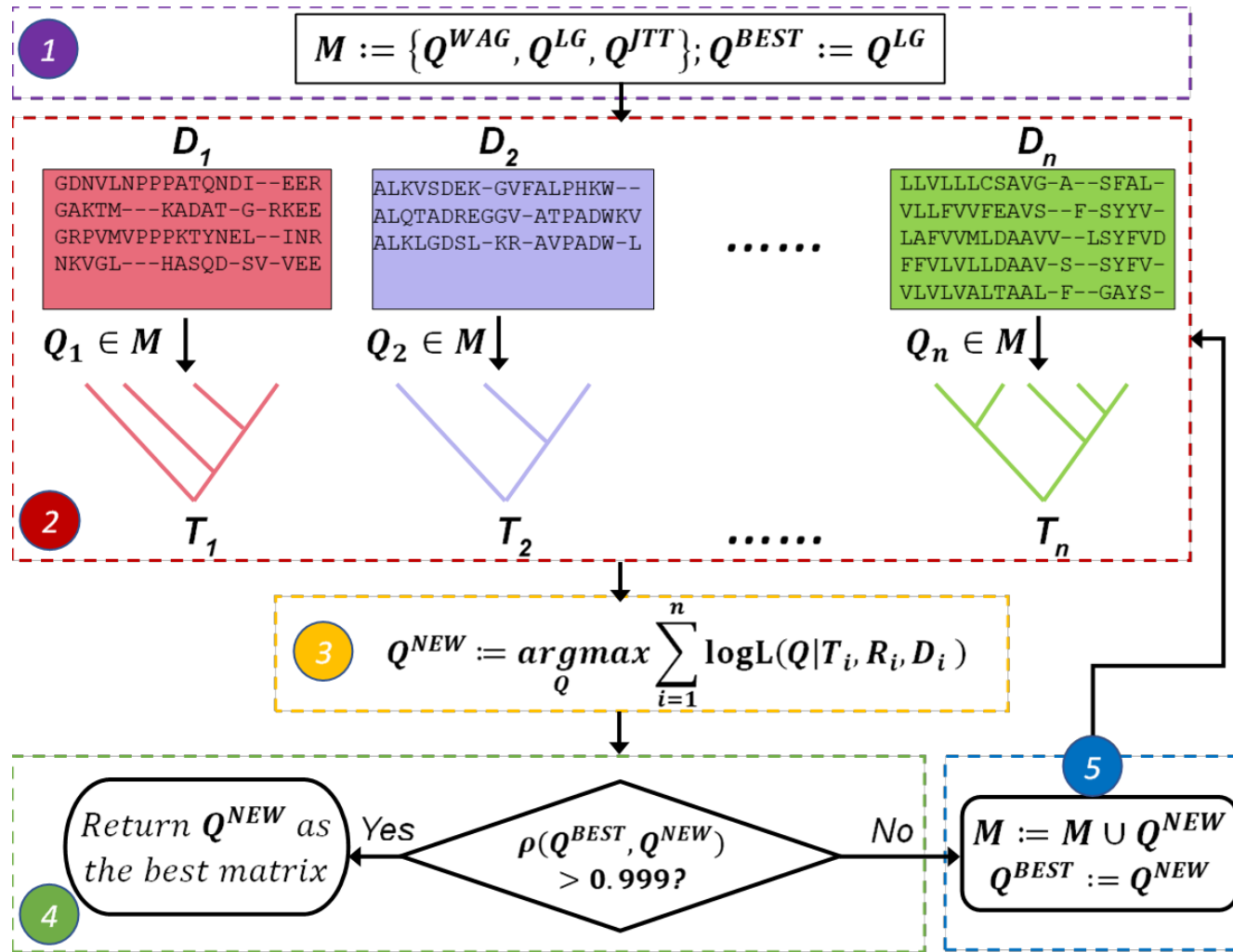
QMaker: Fast and accurate method to estimate empirical models of protein evolution [🔗](#)

B.Q. Minh, C. Cao Dang, L.S. Vinh, R. Lanfear

Systematic Biology 70:1046–1060 (2021)

Altmetric 6 Citations 67

QMaker: Model training



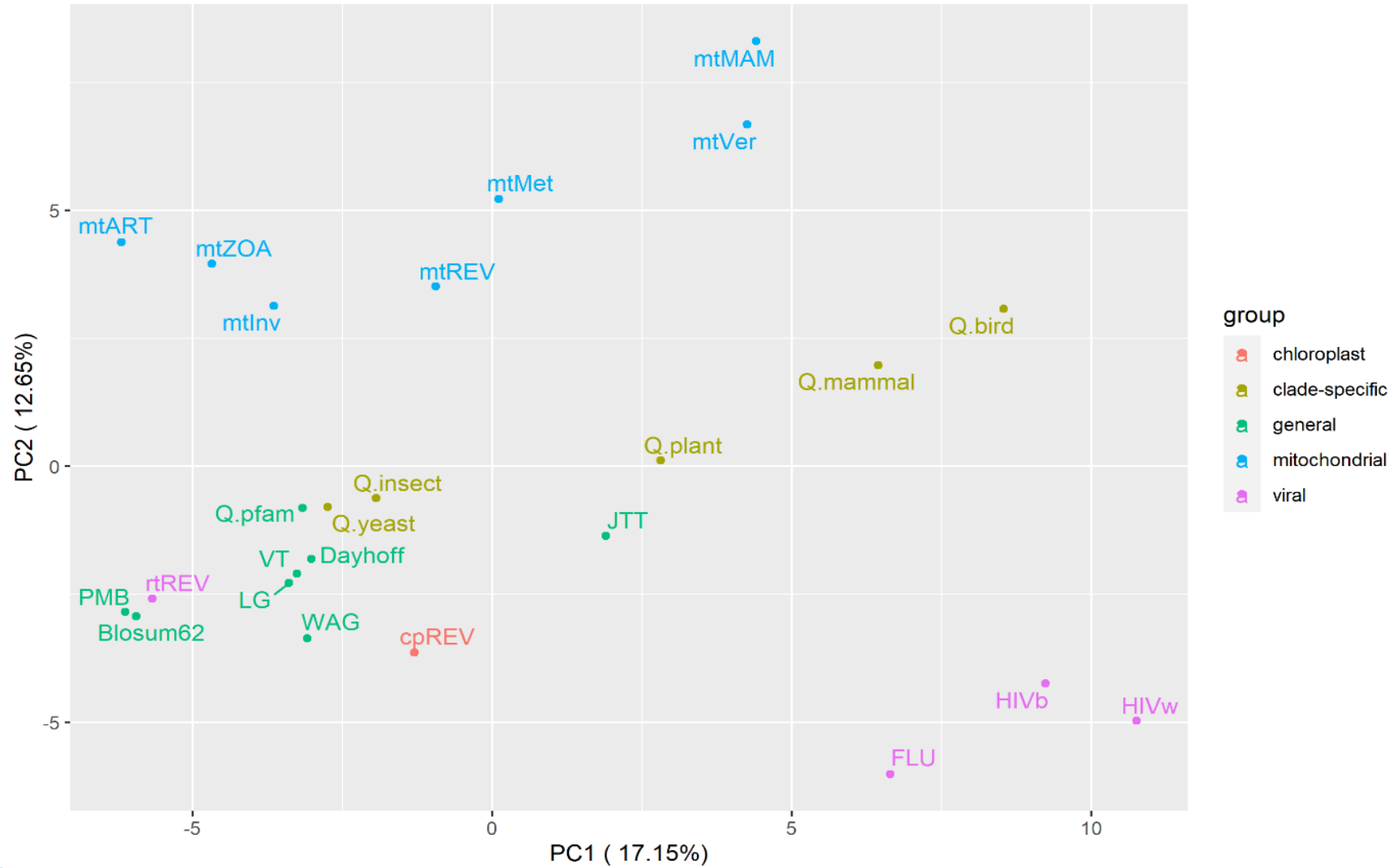
QMaker: Fast and accurate method to estimate empirical models of protein evolution [🔗](#)

B.Q. Minh, C. Cao Dang, L.S. Vinh, R. Lanfear

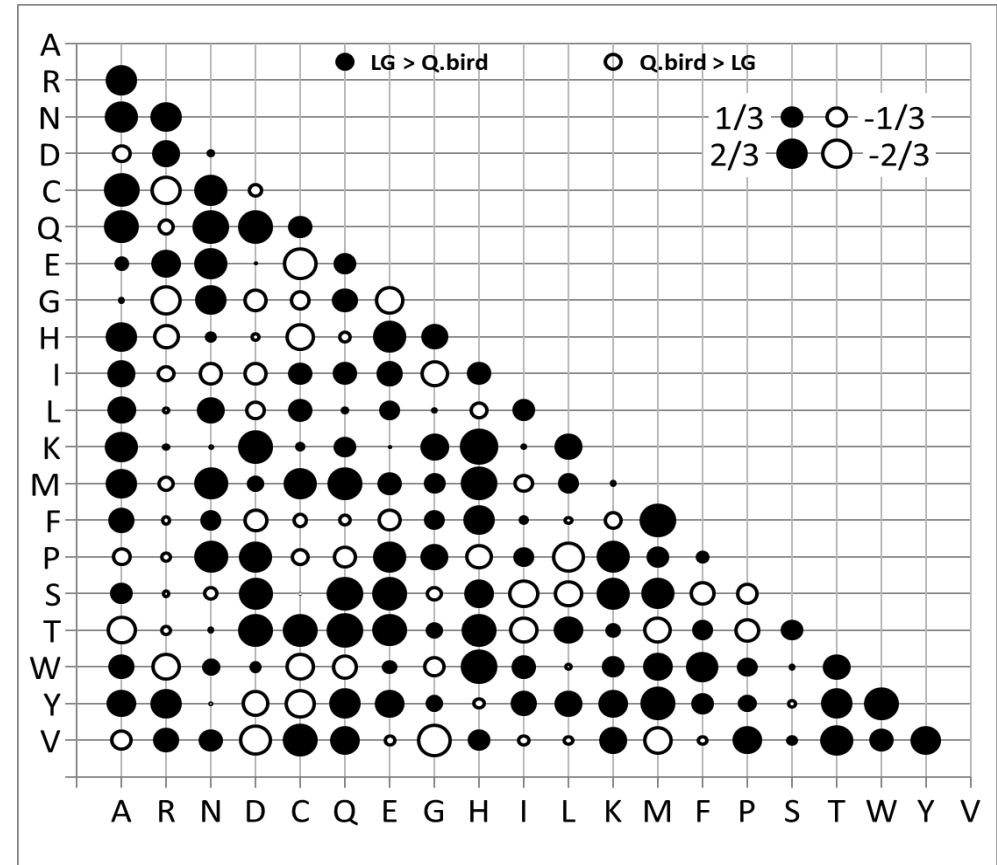
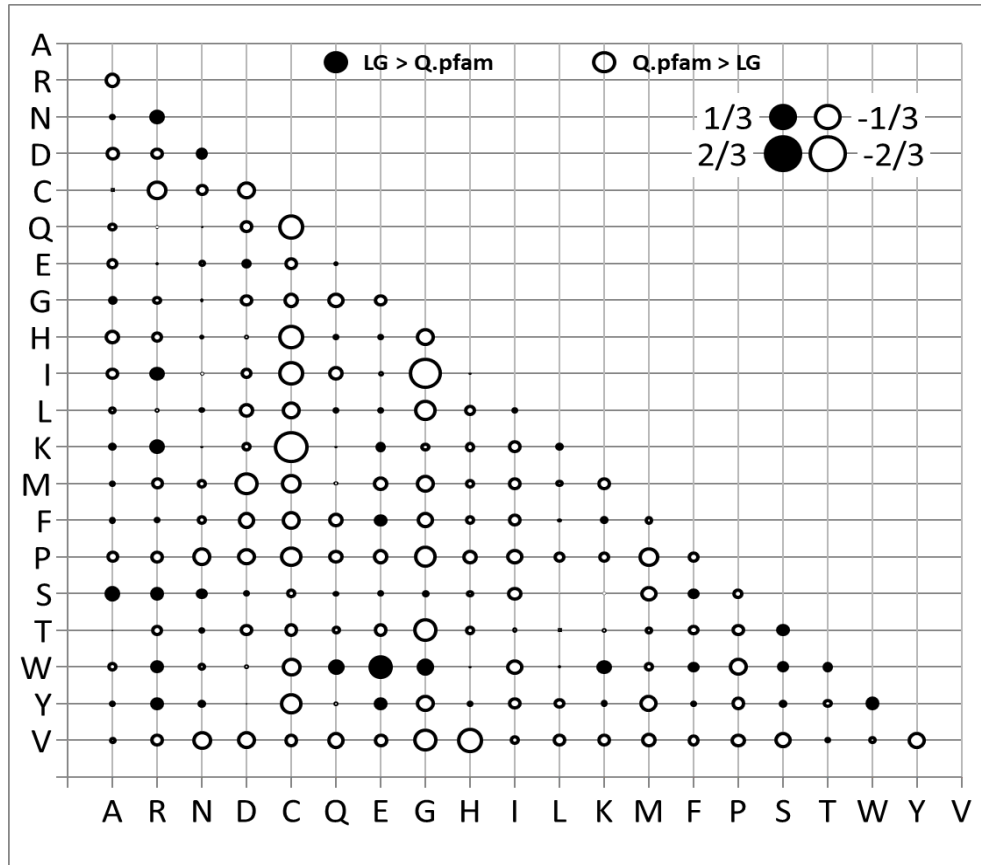
Systematic Biology 70:1046–1060 (2021)

Altmetric 6 Citations 67

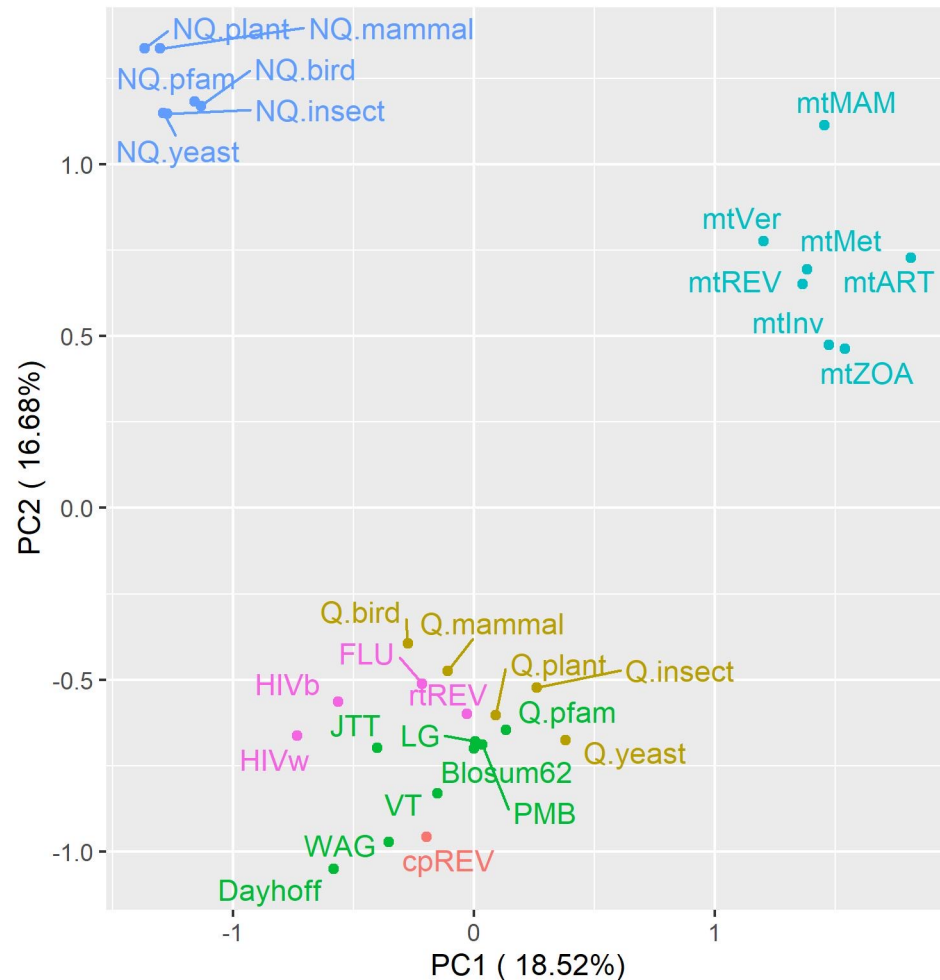
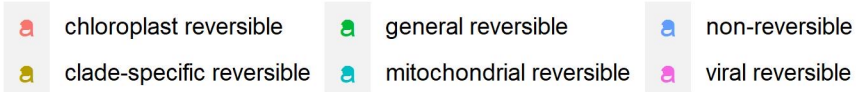
Relationship between Q matrices



LG vs Q.pfam and Q.bird



From reversible to non-reversible models



nQMaker: Estimating Time Nonreversible Amino Acid Substitution Models

C.C. Dang, B.Q. Minh, H. McShea, J. Masel, J.E. James, L.S. Vinh, R. Lanfear

Systematic Biology 71:1110-1123 (2022)



Efficient Likelihood Computations with Nonreversible Models of Evolution

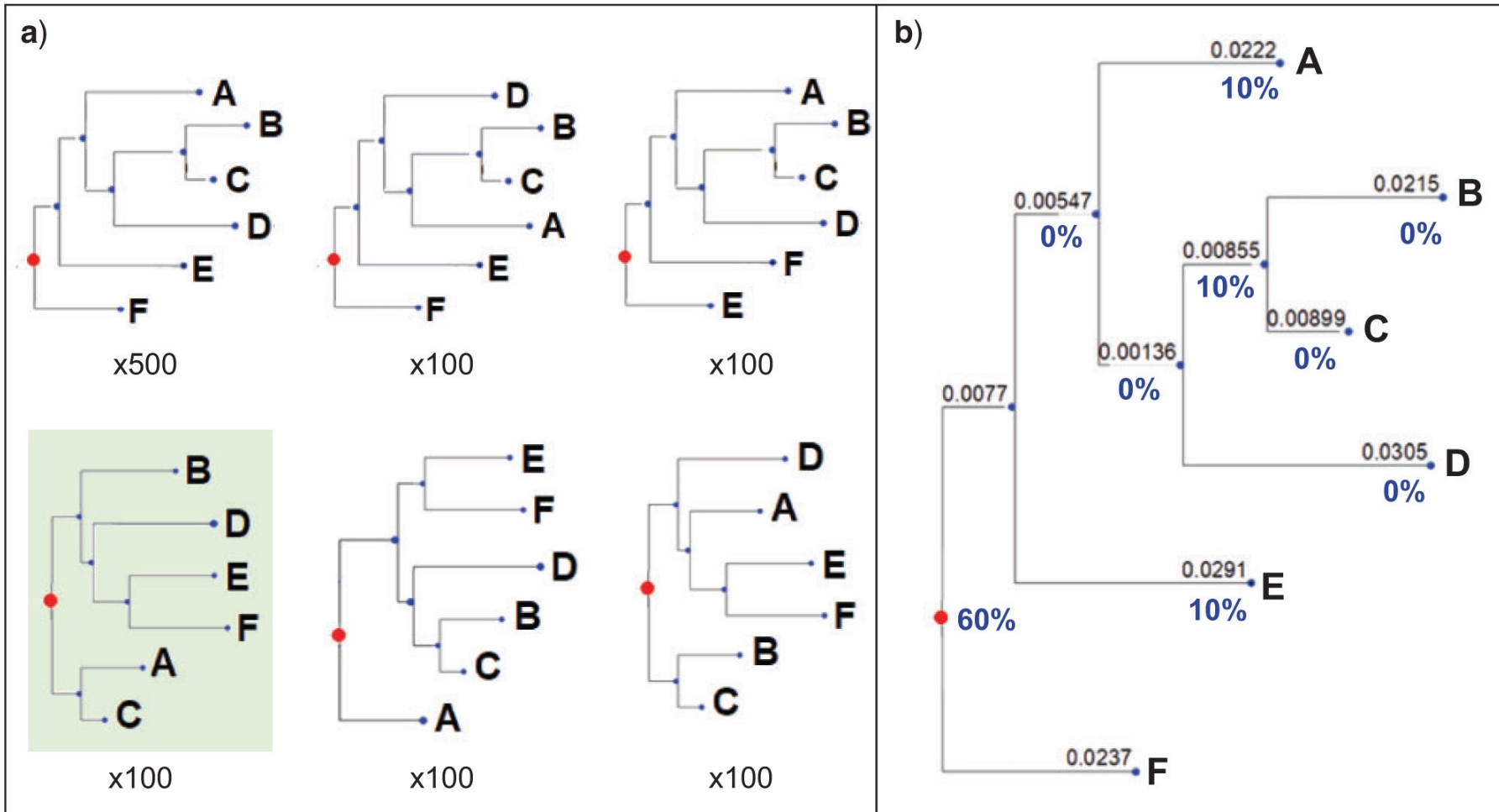
Bastien Boussau , Manolo Gouy

Systematic Biology, Volume 55, Issue 5, October 2006, Pages 756-768,
<https://doi.org/10.1080/10635150600975218>

Non-reversible models allow to compute rootstrap supports

Rooted bootstrap trees

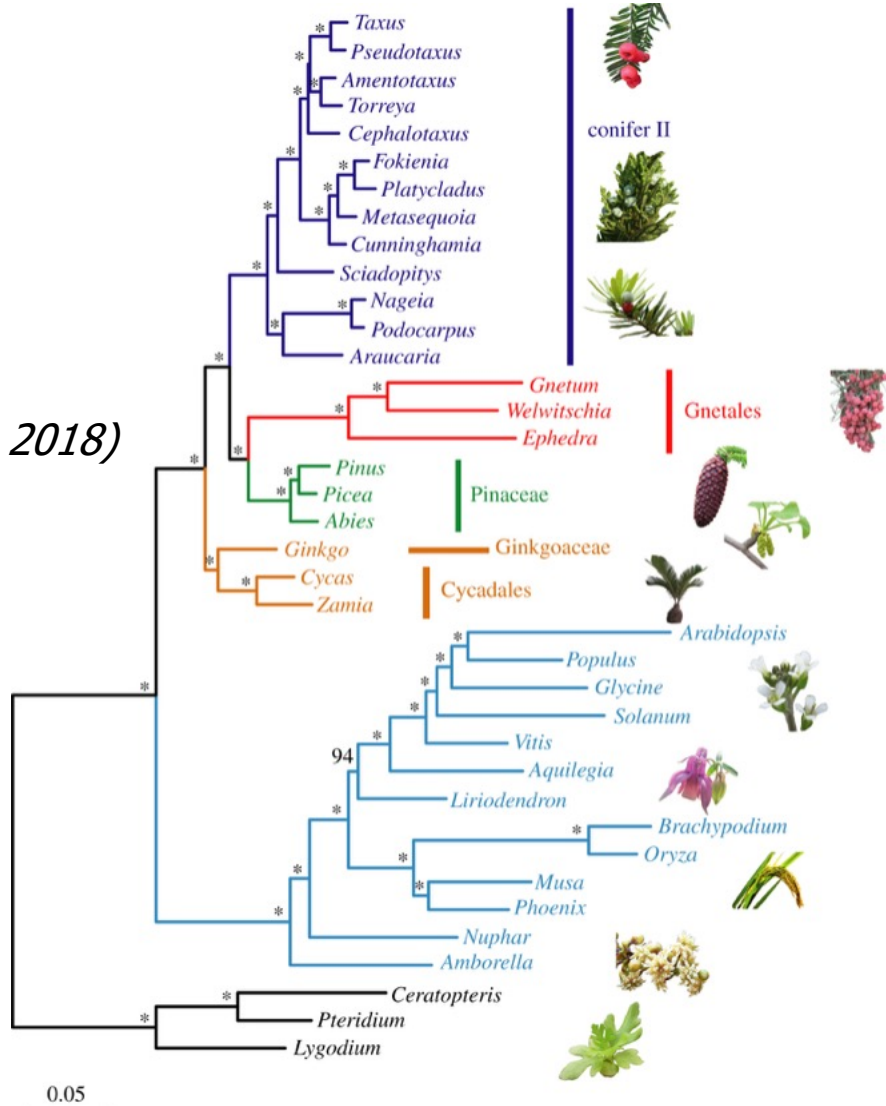
Rootstrap supports



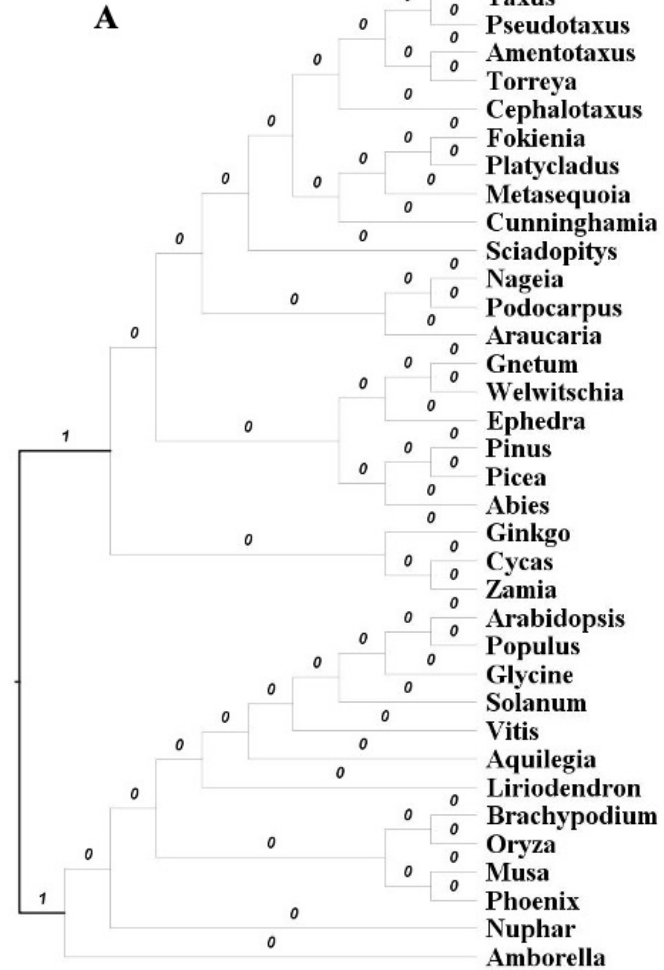
Suha-Naser Khmour

Rooting plant tree without the need of outgroup

(Ran et al. 2018)



Plant rooted tree inferred with NQ.plant



Shown on each branch:
Rootstrap support

Assumptions in model based phylogenetics

- Alignment sites have evolved independently (*i.i.d.*)
- Evolution is memory-less (Markov process)
- Amino acid frequencies are at equilibrium (stationarity)
- Amino acid substitution rates are same in both directions (reversibility)
- There is one Q matrix (homogeneity)
- There is one tree (treelikeness)

Assumptions in model based phylogenetics

- Alignment sites have evolved independently (*i.i.d.*)
- Evolution is memory-less (Markov process)
- Amino acid frequencies are at equilibrium (stationarity)
- ~~○ Amino acid substitution rates are same in both directions (reversibility)~~
- There is one Q matrix (homogeneity)
- There is one tree (treelikeness)

Assumptions in model based phylogenetics

- **Alignment sites have evolved independently (*i.i.d.*)**
- Evolution is memory-less (Markov process)
- Amino acid frequencies are at equilibrium (stationarity)
- ~~○ Amino acid substitution rates are same in both directions (reversibility)~~
- **There is one Q matrix (homogeneity)**
- **There is one tree (treelikeness)**



Thomas Wong
4:00 – 4:45 PM

Profile mixture model

Site-likelihood for site i given an exchangeability matrix S :

$$L_i(S) = \sum_{j=1}^k w_j L(S, F_j | \text{site}_i)$$

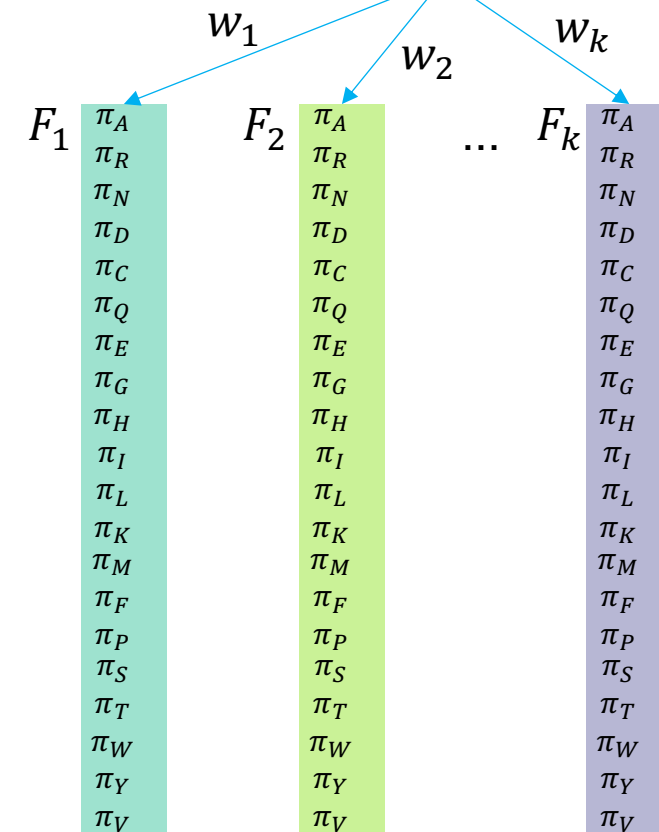
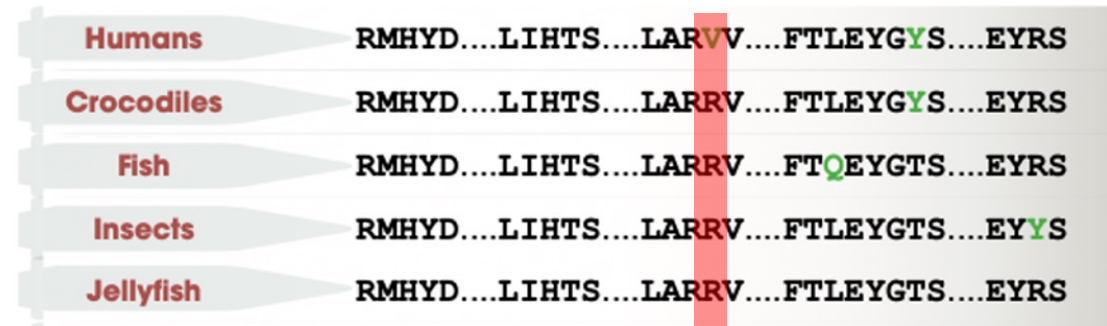
Mixture log-likelihood over all sites:

$$\ell(S) = \sum_{i=1}^N \log L_i(S)$$

Find S that maximizes the log-likelihood:

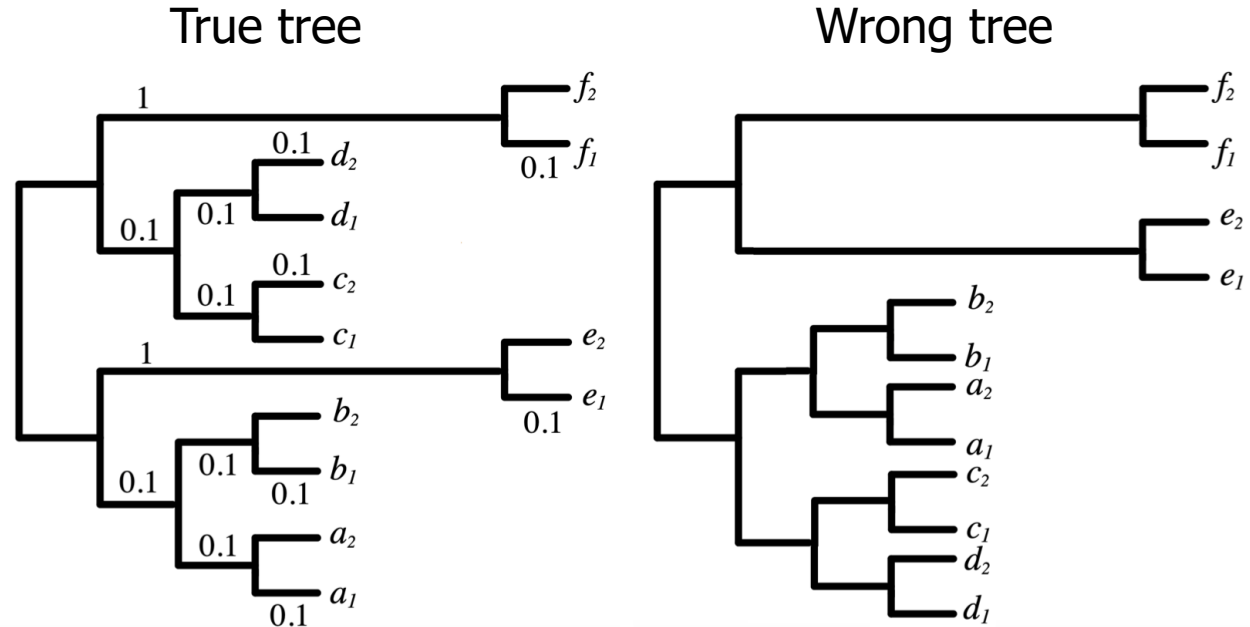
$$S_{max} = \underset{S}{\operatorname{argmax}} \ell(S)$$

Protein Sequence Alignment



Does it help for long branch attraction

Hector Banos

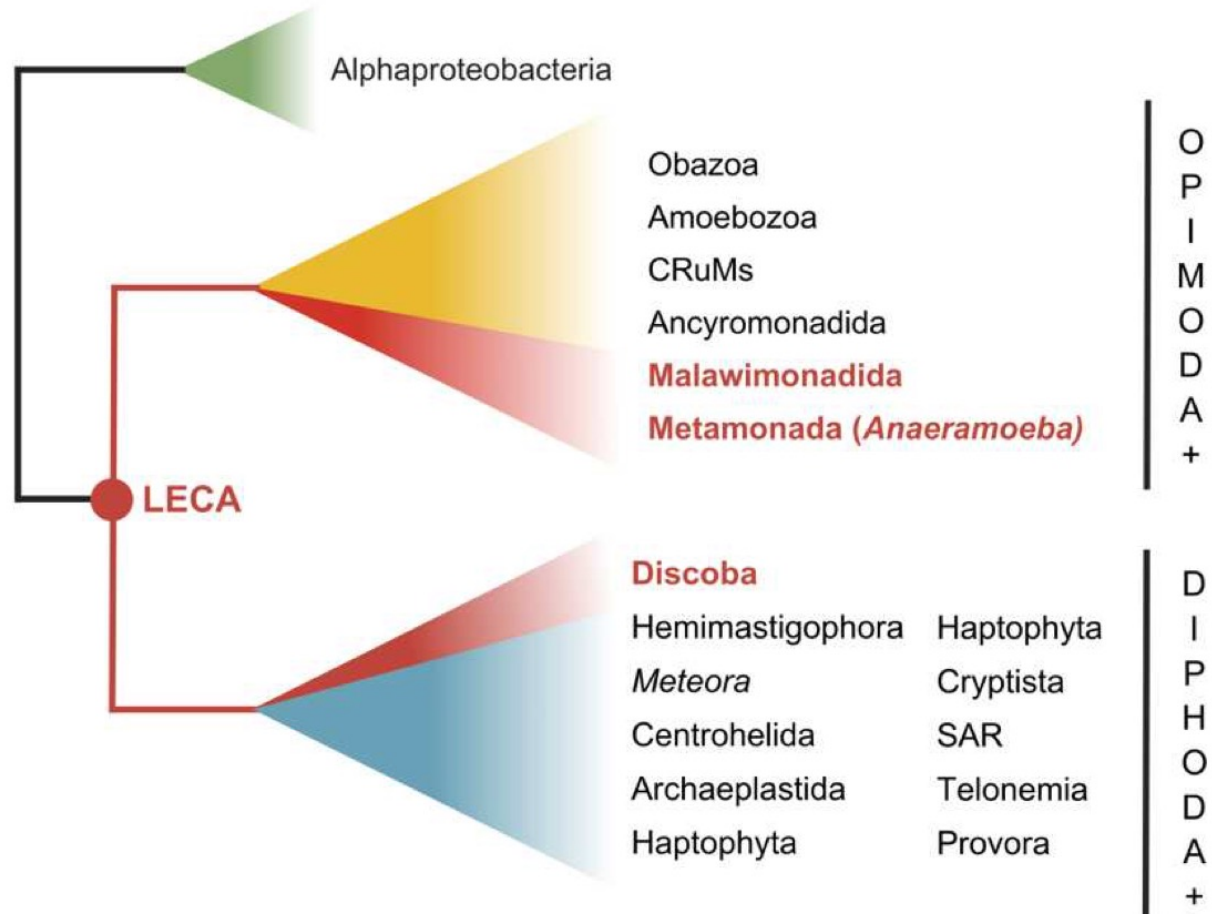


Frequency of inferring true trees

Data simulated under
POISSON+C10+G4

Model	N=500	N=10,000
POISSON+C10+F+G4	64%	96%
LG+C10+F+G4	46%	58%
GTR20+C10+F+G4	64%	83%

Application to rooting eukaryote Tree of Life



Andrew Roger

[A robustly rooted tree of eukaryotes reveals their excavate ancestry](#)



K. Williamson, L. Eme, H. Baños, C. McCarthy, E. Susko, R. Kamikawa, R. Orr, S. Muñoz-Gómez, [B.Q. Minh](#), A. Simpson, A. Roger

Research Square (2024)

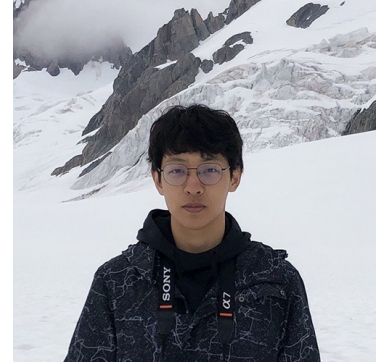
Tree inferred under LG+MEOW80+G4 with IQ-TREE 2

Part 3: Machine learning for protein model selection

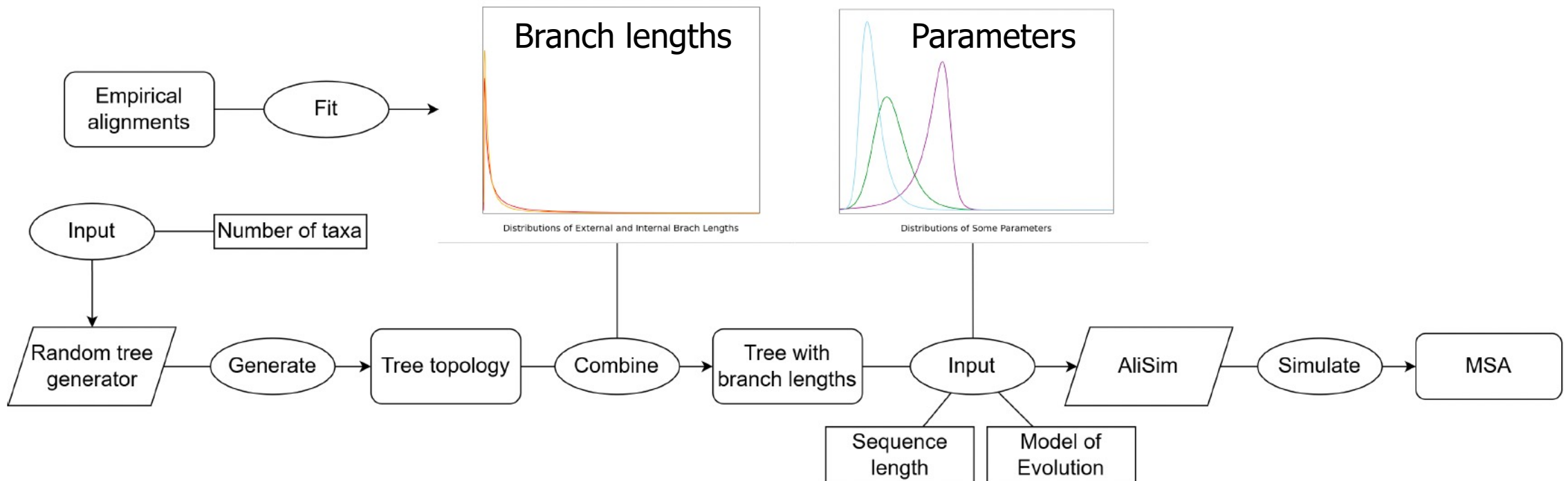
- ModelTeller: Random forest classifier for six DNA models (Abadi et al. 2020)
- ModelRevelator: Neural network classifier for six DNA models and regressor for alpha parameter of Gamma distribution of rate heterogeneity across sites (Burgstaller-Muehlbacher et al. 2023)
- ModelDetector: Neural network classifier for AA models (Nguyen & Vinh 2024) (Nov 16th)

Data for training

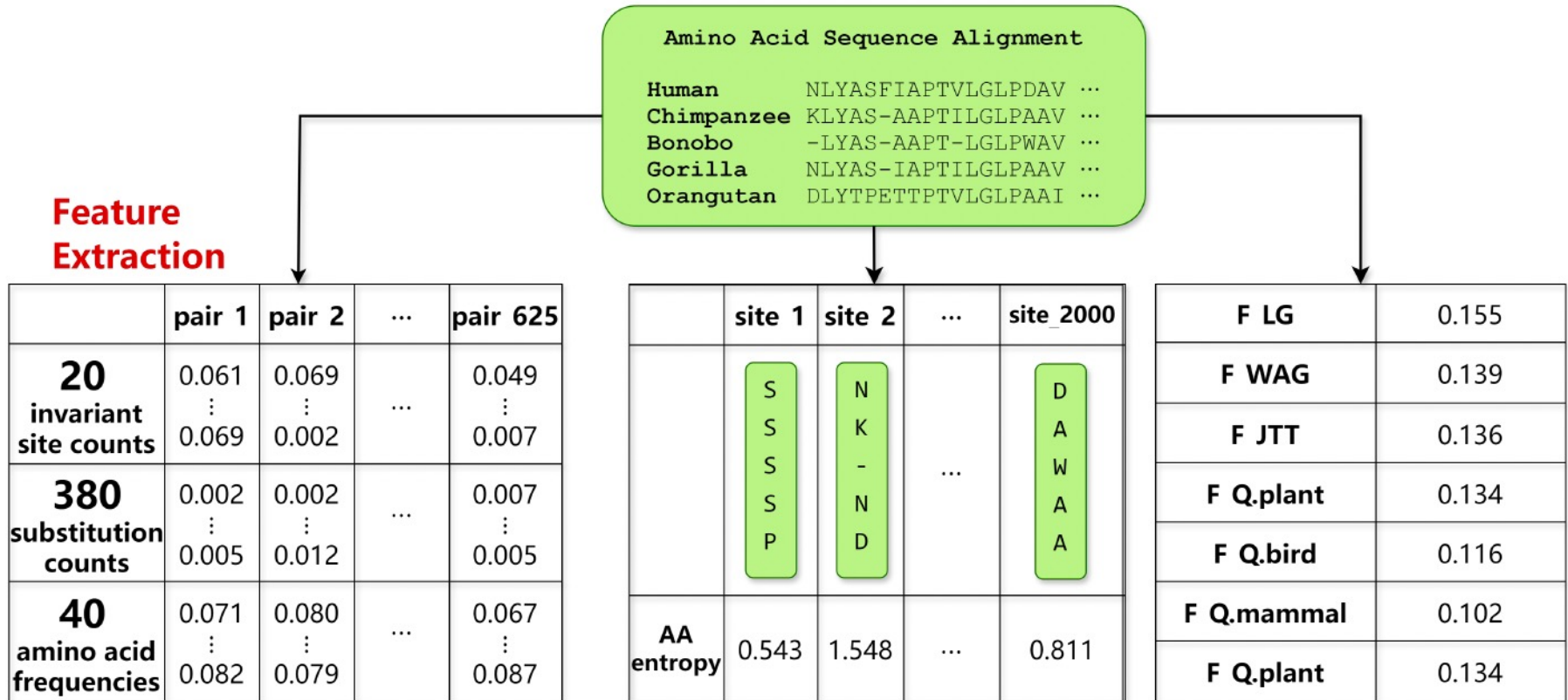
- Real data from EvoNAPS DB
 - 21,800 protein alignments (<https://github.com/Cibiv/EvoNAPS>)
 - LG, WAG, JTT, Q.pfam, Q.plant, Q.bird, Q.mammal are most represented (87%)
 - Gamma (G4), Free-rate models (R2, R3, R4) are most represented (90%)
- Simulated data



Yanghe Dong



Feature extractions



440 normalised statistics of 625 randomly selected pairs of sequences

Sorted amino acid entropies of 2000 randomly selected sites

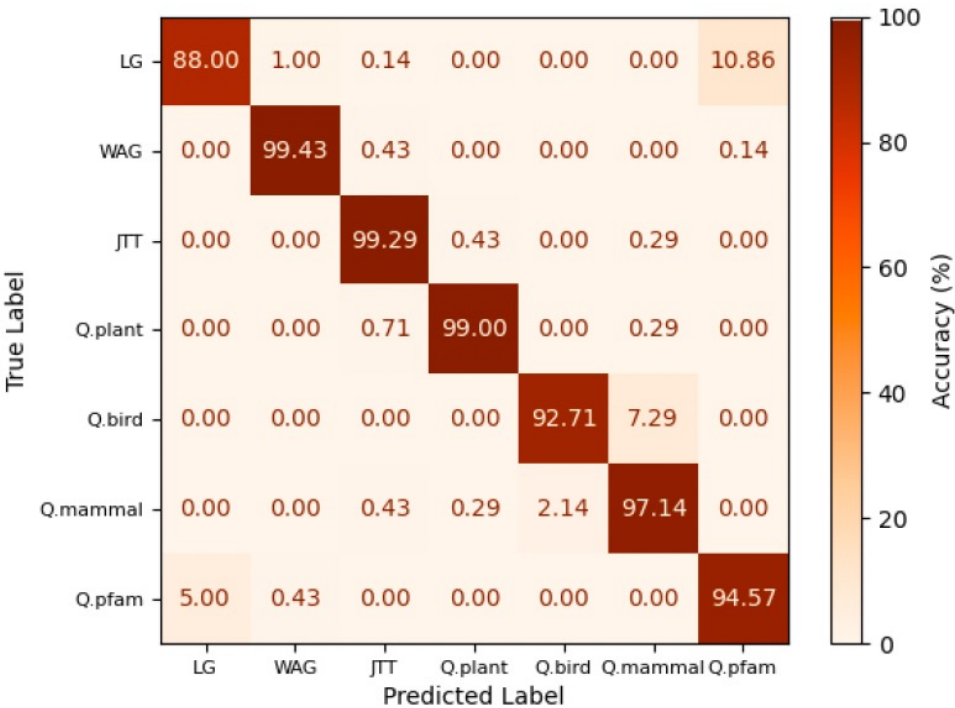
KL divergence of MSA amino acid frequencies from predefined frequencies of 7 models

Predict substitution models

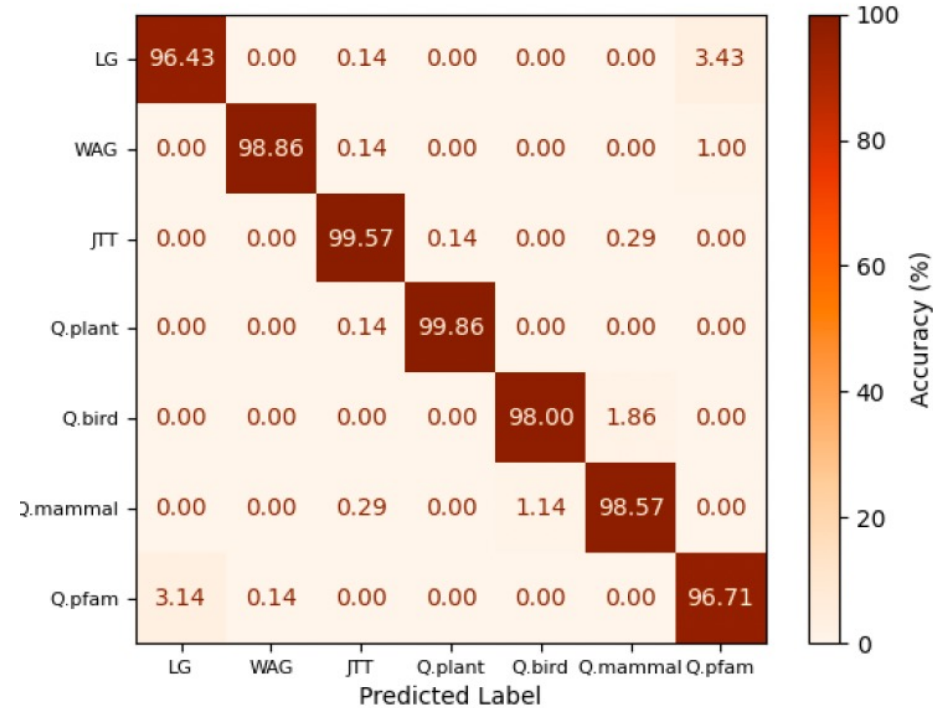
Predict +G, +R or not

Predict +F or not

Accuracy of predicting substitution models



Convolutional neural network



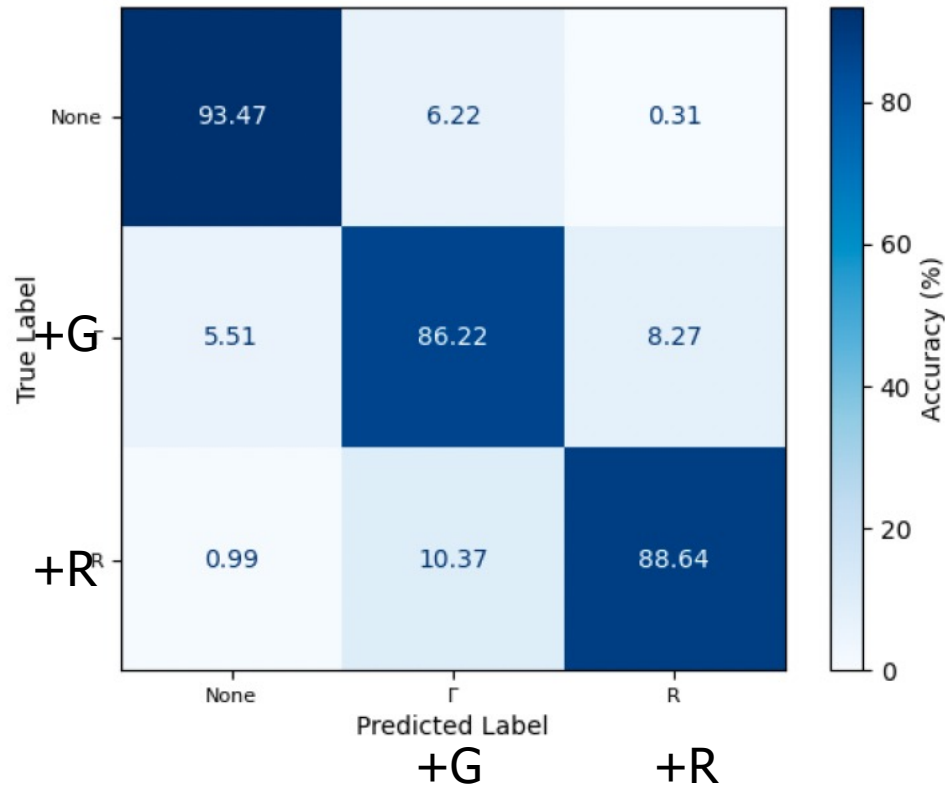
ModelFinder (maximum likelihood)

* Testing on real data:
Accuracy: 71%

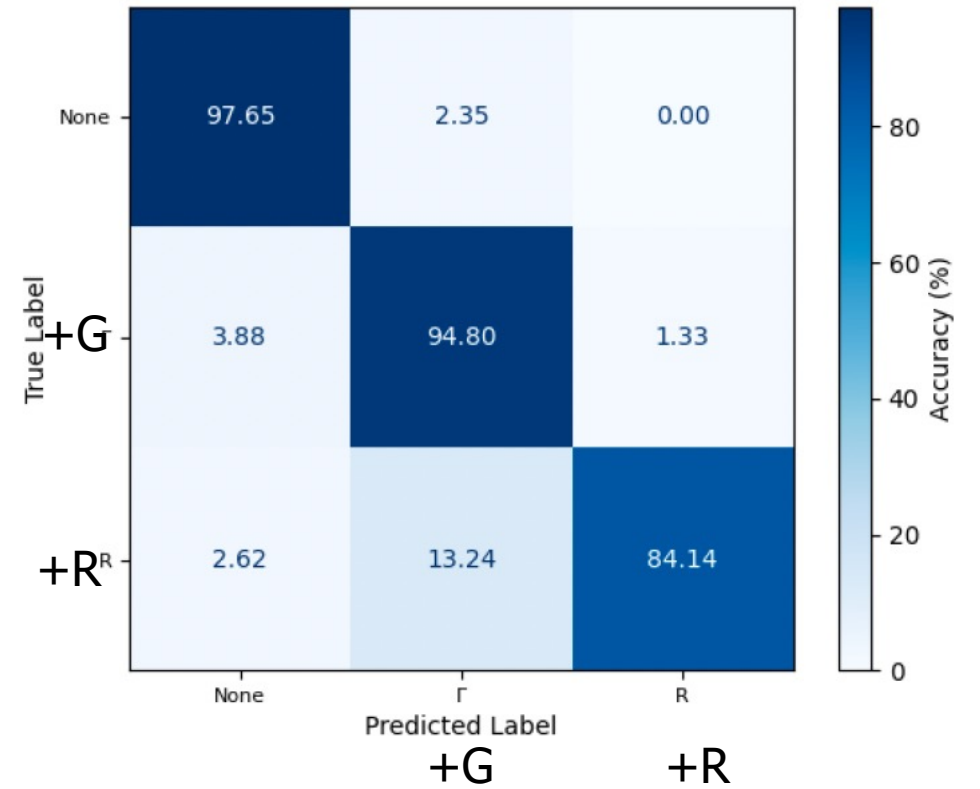
* Testing on simulated
data minicking real data:
78%

* Re-trained network on
real data:
accuracy 85%

Accuracy of predicting rate heterogeneity across sites

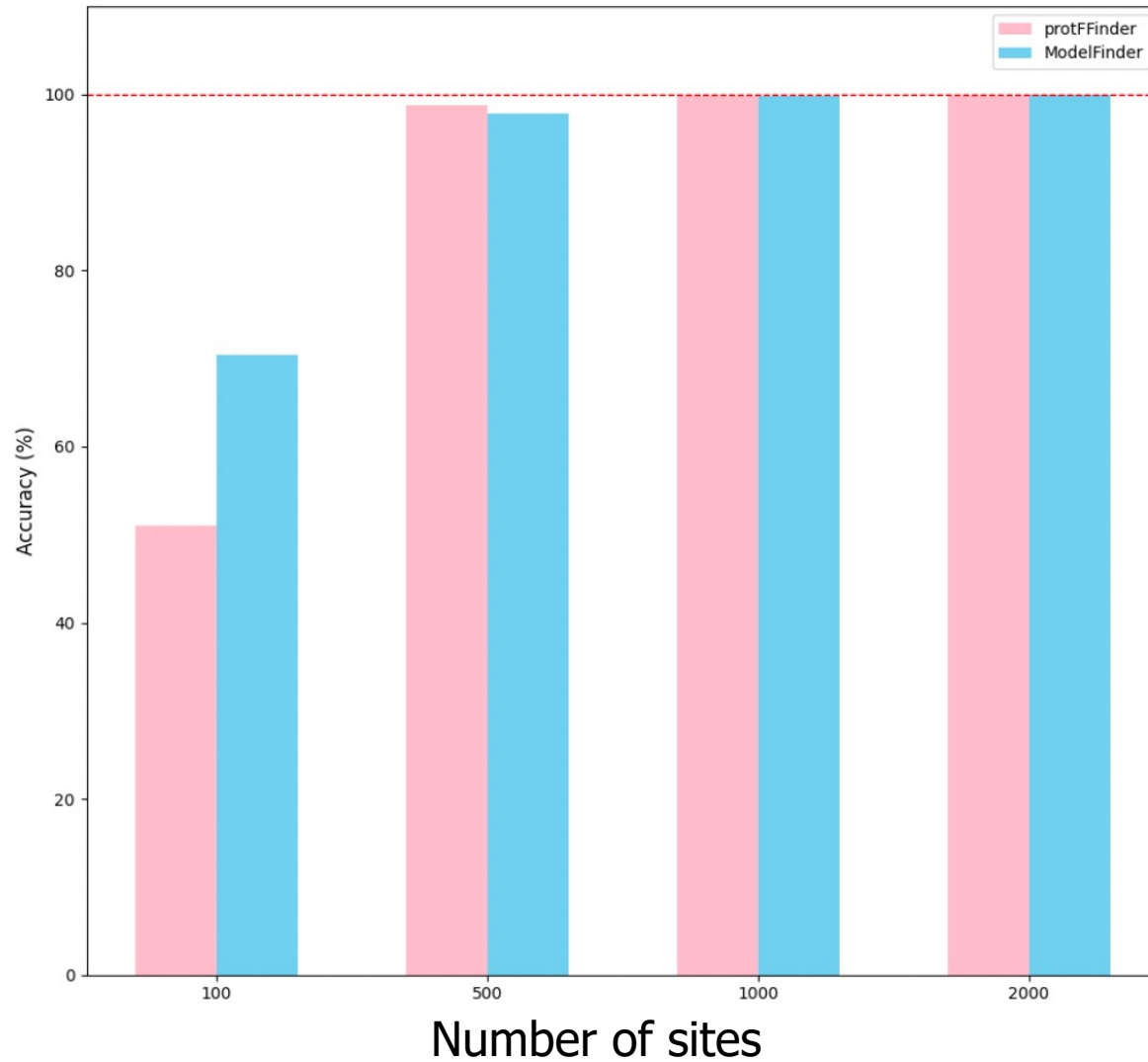


Convolutional neural network



ModelFinder (maximum likelihood)

Accuracy of +F/-F classifier



Summary

1. MAPLE and CMAPLE facilitate greener computing for molecular evolution
2. Complex models of sequence evolution facilitate deep protein phylogeny
3. Machine learning is promising despite gaps with maximum likelihood methods

Acknowledgements



**Minh's and
Lanfear's
lab ANU**



