PhyloBayes
- what, why, and how

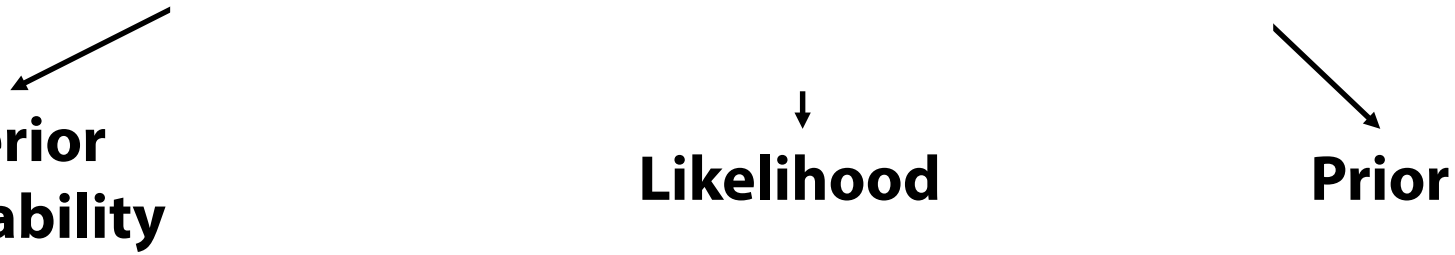
Xiaofan Zhou
Integrative Microbiology Research Centre
South China Agricultural University
2019-01-25

PhyloBayes

a Bayesian phylogenetic software based on *mixture* models

Bayesian phylogenetics

$$P(M, D)$$



Among site variation in sequence evolution

- Evolutionary rates

- Substitution model
  - Exchange rate matrix
  - Equilibrium frequency profile (**PhyloBayes-CAT!!!**)

Components of substitution model

Rate matrix of the GTR model

From	To A	To T	To C	To G
A		$a\pi_T$	$b\pi_C$	$c\pi_G$
T	$a\pi_A$		$d\pi_C$	$e\pi_G$
C	$b\pi_A$	$d\pi_T$		$f\pi_G$
G	$c\pi_A$	$e\pi_T$	$f\pi_C$	

"exchangeability"

From	To A	To T	To C	To G
A		a	b	c
T	a		d	e
C	b	d		f
G	c	e	f	

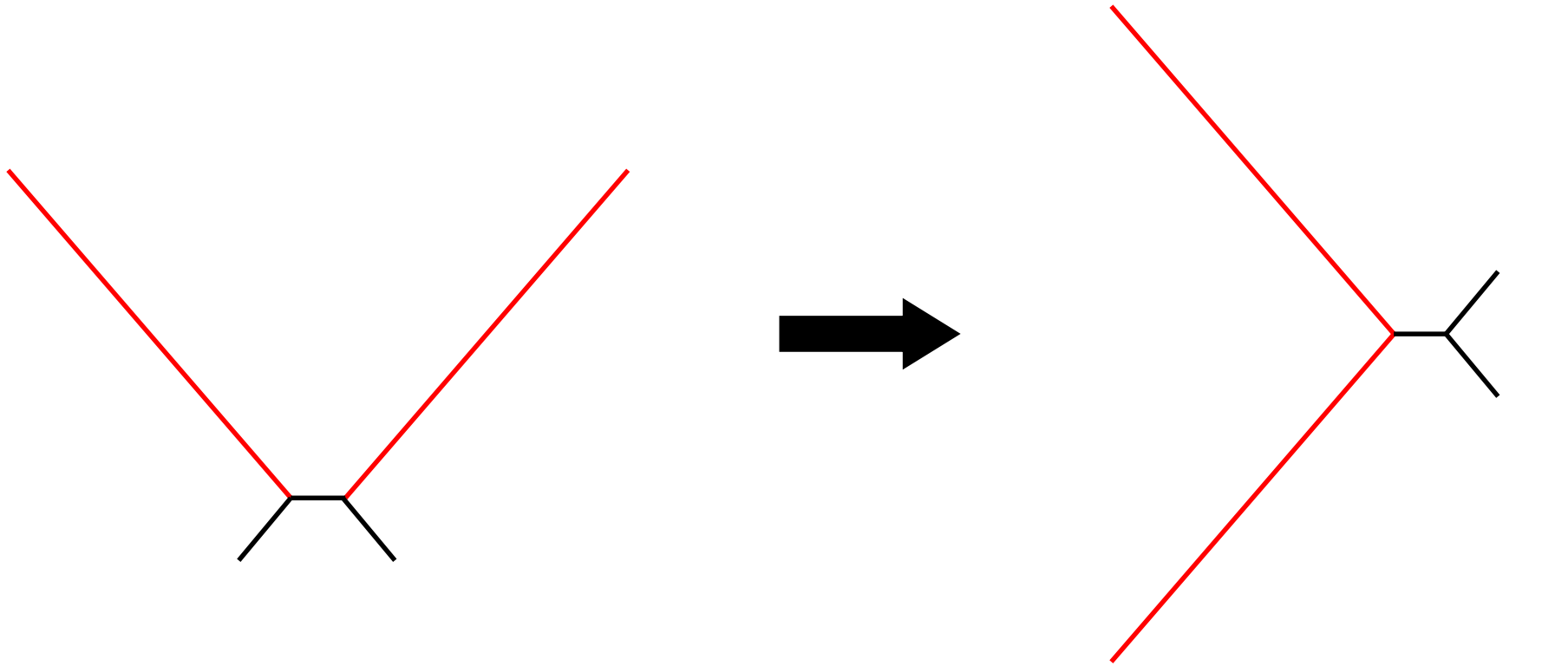
F81 (Poisson) model:
$a = b = c = d = e = f = 1$

Equilibrium frequencies

$[\pi_A, \pi_T, \pi_C, \pi_G]$

Dirichlet process for site assignment

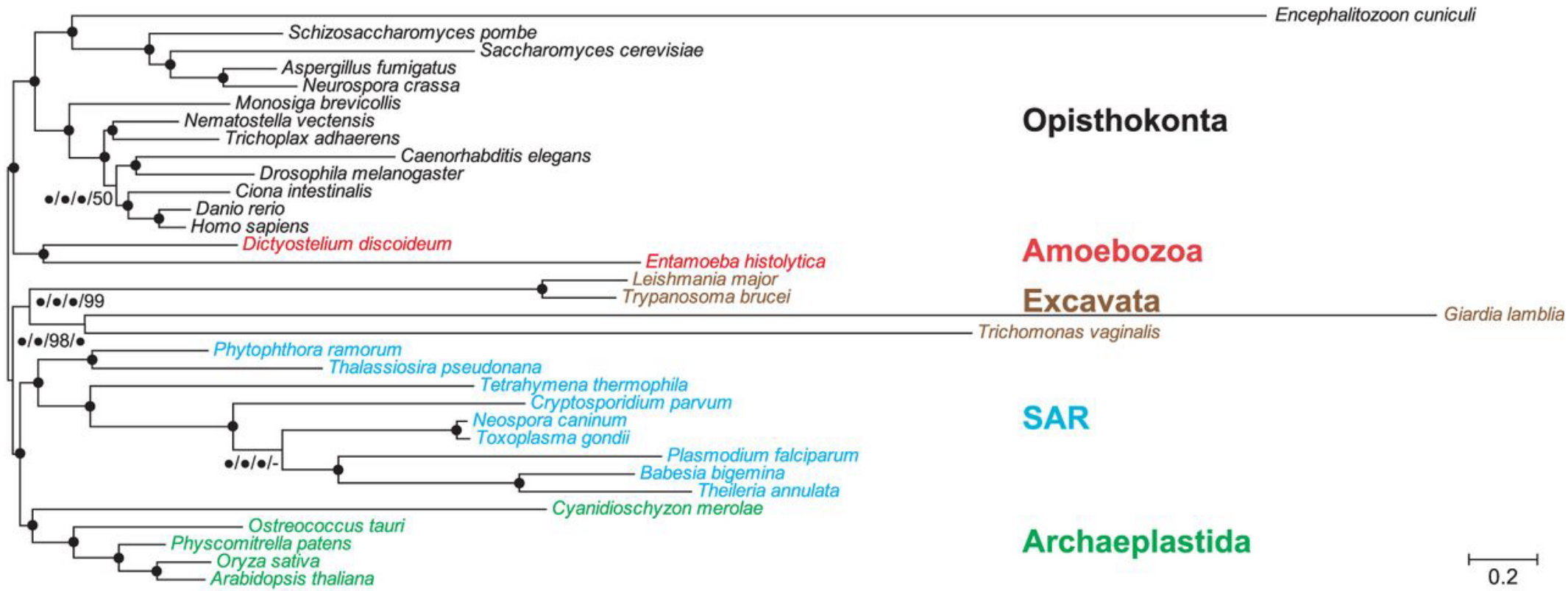
Long-Branch Attraction



True tree

Inferred tree

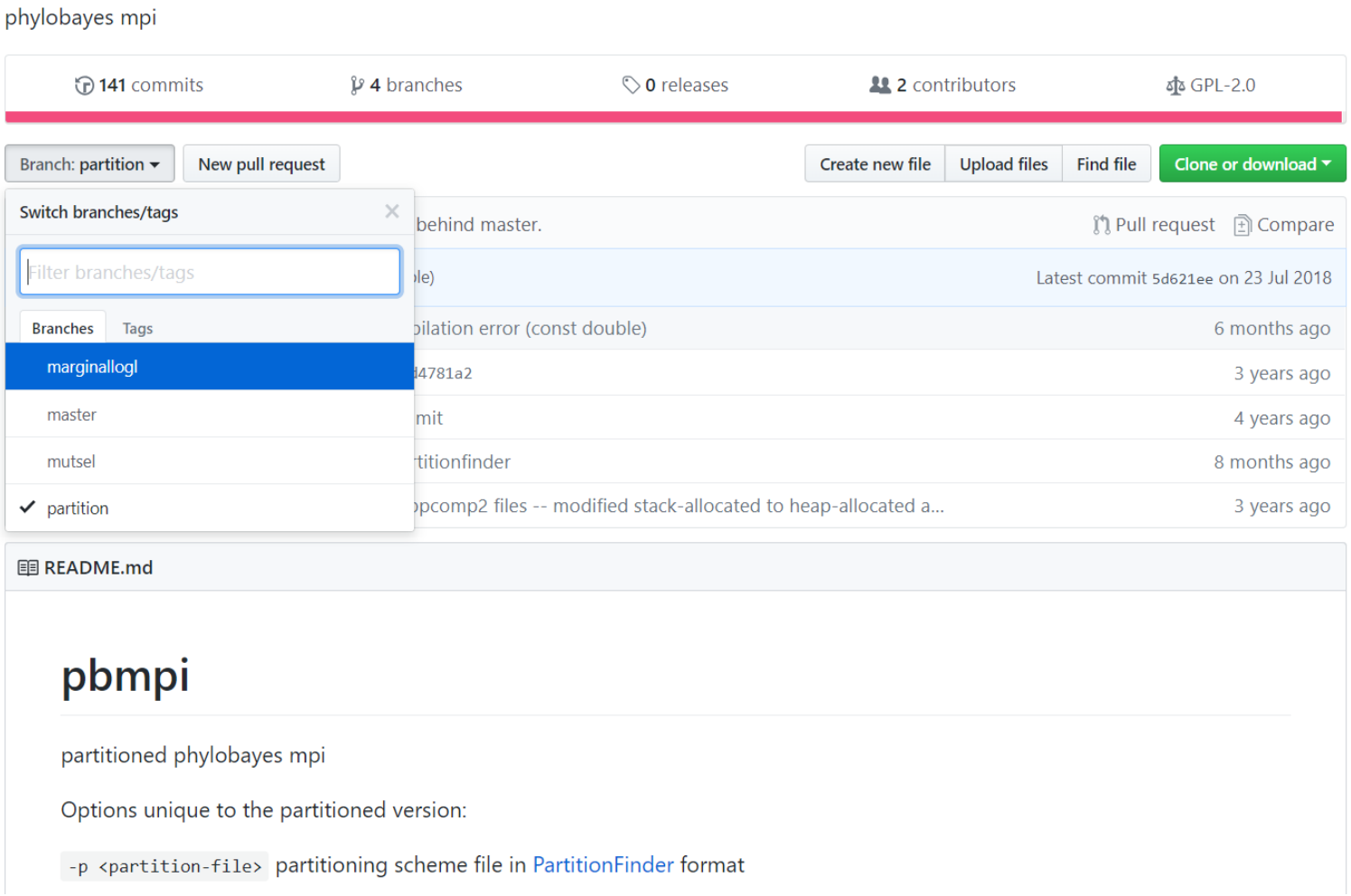
PhyloBayes-CAT can suppress LBA



Models available in PhyloBayes

Exchangeability matrix	Equilibrium frequency	Evolutionary Rate	Data partition
• Poisson • Empirical model - JTT - LG … • GTR • Mixture model	• Site homogeneous - similar to "+F" in ML • Site heterogeneous - CAT - empirical mixture models (e.g. C10-C60)	• Gamma distribution - discrete - continuous • CAT model of rate	• Partially linked branch-length • Partition-specific substitution matrix not supported (?)

Partitioned PhyloBayes-MPI



Limitations of PhyloBayes

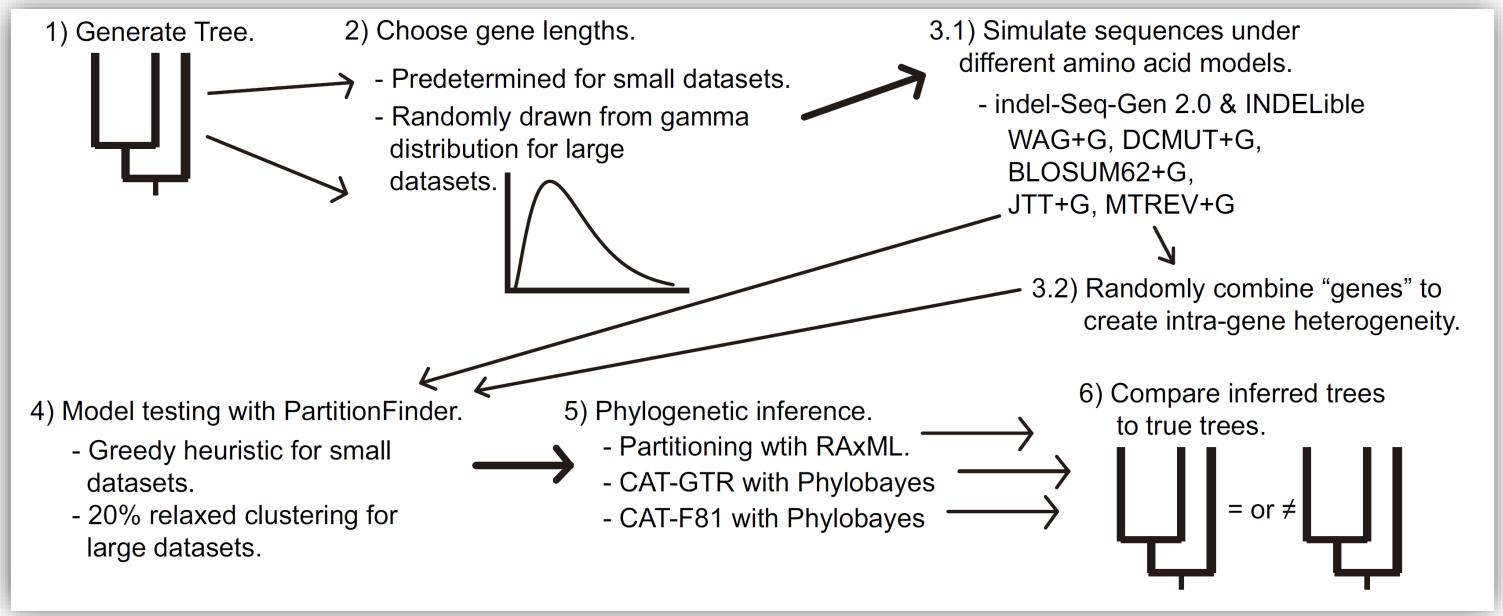
- CAT+GTR model is ***very*** slow……
  - Even with PhyloBayes-MPI
- Hard to converge
  - Particularly CAT+GTR
- CAT-Poisson is often used instead of CAT-GTR
- Results of unconverged analyses were reported

Syst. Biol. 66(2):232–255, 2017

DOI:10.1093/sysbio/syw084
Advance Access publication September 14, 2016

Who Let the CAT Out of the Bag? Accurately Dealing with Substitutional Heterogeneity in Phylogenomic Analyses

NATHAN V. WHELAN* AND KENNETH M. HALANYCH



"..... CAT-F81 consistently performed worse than other models *(partitioning and CAT-GTR)* in inferring the correct branching patterns ....."

Alternative solution in ML

- Empirical mixture models in ML
  - C10 to C60

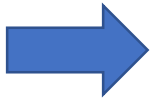
```
[ ----------------------------------------------------------
   CAT-C10 profile mixture model of Le, Gascuel & Lartillot (2008)
 ---------------------------------------------------------- ]

frequency C10pi1 = 0.4082573125 0.0081783015 0.0096285438 0.0069870889 0.0349388179 0.0075279735 0.0097846653 0.1221613215 0.0039151830 0.0125784287 0.0158338663 0.0059670150 0.00813132
frequency C10pi2 = 0.1027763487 0.0418664491 0.0213272051 0.0155943616 0.0149663448 0.0440685478 0.0419667447 0.0138805792 0.0158864807 0.1066076641 0.1131944125 0.0436343681 0.04378003
frequency C10pi3 = 0.0351766018 0.0019678632 0.0016591476 0.0006768741 0.0078706538 0.0016559557 0.0019686768 0.0022420602 0.0012878339 0.3515819591 0.1278183107 0.0018856550 0.02426317
frequency C10pi4 = 0.0408513927 0.0269887074 0.2185648186 0.2333814790 0.0037602852 0.0380451418 0.0901238869 0.1158332065 0.0373197176 0.0025523644 0.0052164616 0.0485017266 0.00225717
frequency C10pi5 = 0.0185492661 0.0062362395 0.0024895723 0.0009775062 0.0070416514 0.0083539447 0.0024891617 0.0028952913 0.0040103982 0.1632422345 0.4443079409 0.0043570878 0.12028156
frequency C10pi6 = 0.1106750119 0.0352190043 0.0405186210 0.1636437899 0.0014834855 0.0877962201 0.2638456592 0.0325228293 0.0163803600 0.0068334902 0.0140679579 0.0677158208 0.00489881
frequency C10pi7 = 0.0522657662 0.0668294648 0.0714836849 0.0297745257 0.0143324928 0.0736540298 0.0388386669 0.0228101108 0.1551638111 0.0187406149 0.0653779932 0.0439469345 0.02071891
frequency C10pi8 = 0.0116587342 0.0050990142 0.0064011054 0.0021742457 0.0105340743 0.0040203734 0.0024251112 0.0034709143 0.0366787049 0.0187185330 0.0676489746 0.0026694717 0.01435348
frequency C10pi9 = 0.0627195947 0.2038782162 0.0428629162 0.0236193294 0.0052662886 0.1098111767 0.0686284994 0.0256174957 0.0332612124 0.0128968249 0.0305627740 0.2270839355 0.01240369
frequency C10pi10 = 0.1145518598 0.0324008908 0.0750614981 0.0416192189 0.0098549497 0.0339624663 0.0364907910 0.0503817581 0.0165233329 0.0092949460 0.0139153707 0.0423026886 0.0082240
model C10 = POISSON+G4+FMIX{C10pi1:1:0.1191344178,C10pi2:1:0.0874372456,C10pi3:1:0.1037105070,C10pi4:1:0.0922584809,C10pi5:1:0.1070492801,C10pi6:1:0.1329945166,C10pi7:1:0.0538028458,C10
```

Alternative solution in ML

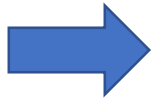
- Empirical mixture models in ML
  - C10 to C60

- Fast implementation in IQ-TREE, further accelerated using PMSF

Tree, Alignment, Model (e.g., C10)



$$P(j|x) = \frac{w_j \times P(x|j)}{\sum_j w_j \times P(x|j)}$$



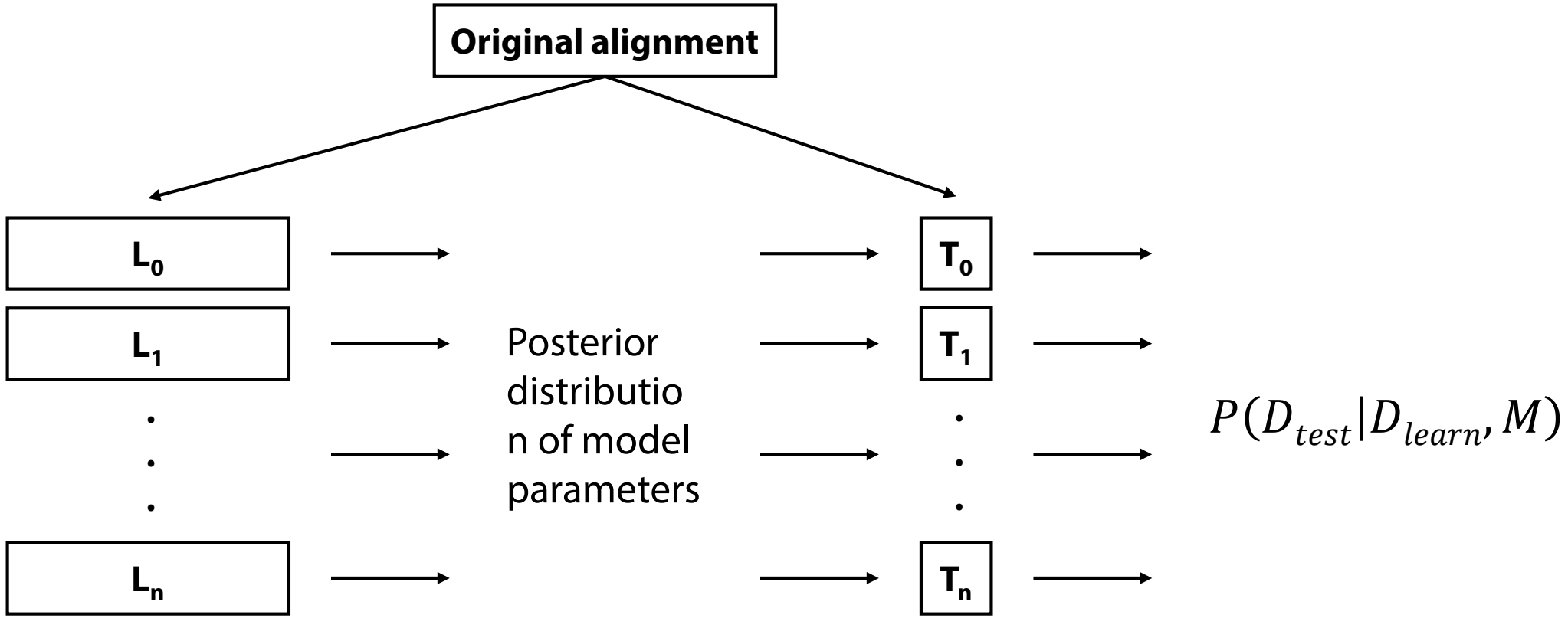
$$f_a(x) = \sum_j f_{aj} \times P(j|x)$$

Wang et al. (2017) Syst. Biol.

Model comparison in PhyloBayes

- Bayes factor
- Cross-Validation
  - How well can the models predict future data?
- Posterior predictive analysis
  - How well does simulated data approximate observed data?

Cross Validation



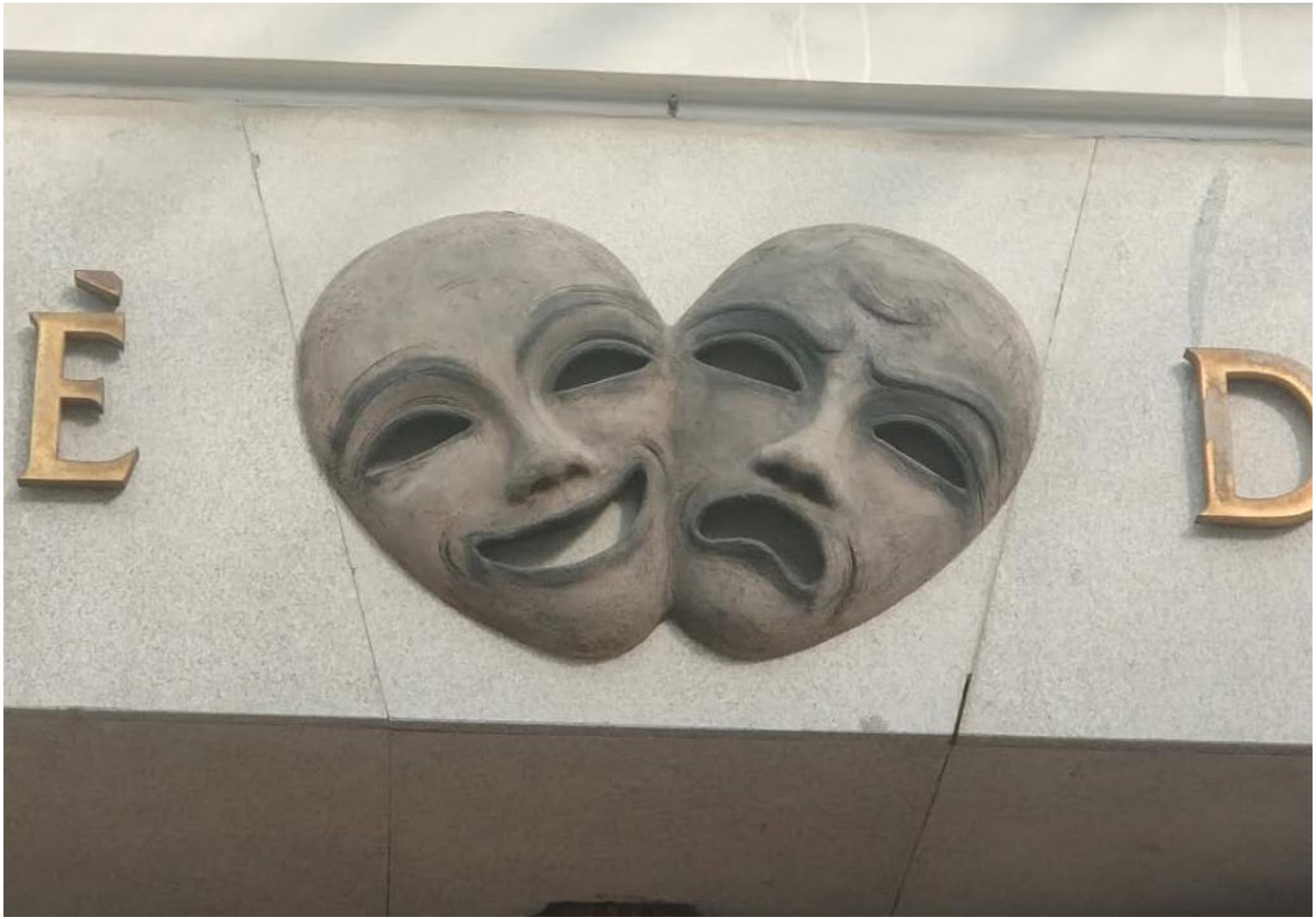
- $P(D_{test}|D_{learn}, Ma)\ vs.\ P(Dtes_t|Dl_{earn}, Mb)$

Posterior predictive analysis

- Compare observed data vs. simulate data (using trained model)

- Resemble parametric bootstrap in ML
  - Parameters sampled from posterior distributions instead of being fixed

- Available tests:
  - Biochemical specificity
  - Compositional homogeneity
  - Saturation


É
D

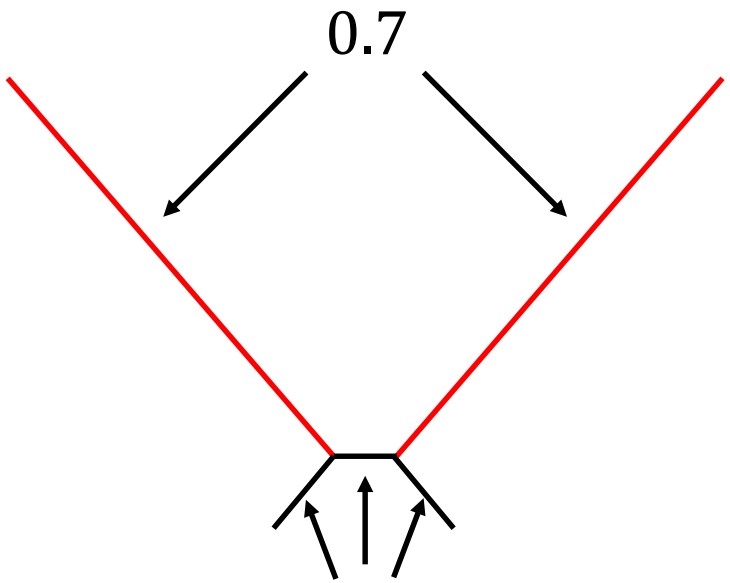
Tutorial

1. Run PhyloBayes analyses on a toy dataset with four taxa:
   - Check for convergence (e.g., TRACER)
   - Compare inferred tree with true tree (e.g., phylo.io)
   - Try different models (e.g., F81, GTR, CAT+F81, CAT+GTR…)

2. Perform model comparison
   - Cross-Validation
   - Posterior predictive analysis

3. Compare with empirical mixture models in ML (IQ-TREE)

The dataset was simulated under LBA

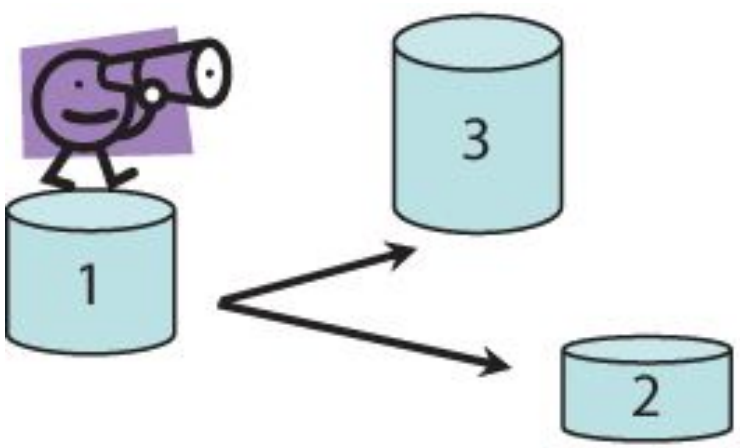


True tree

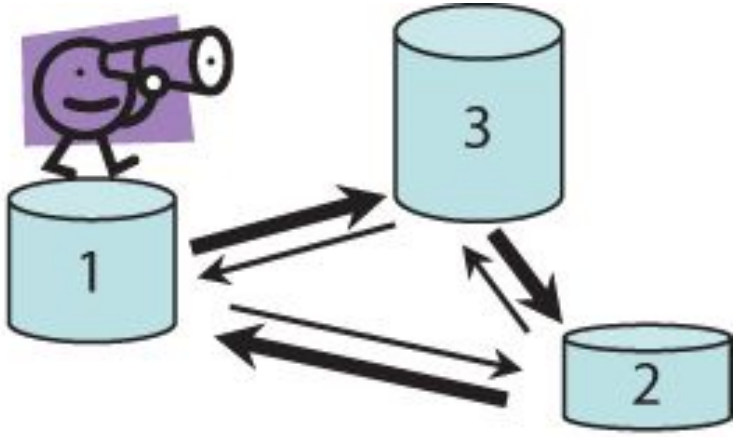
- 4 taxa, ~21,000 amino acid sites
- Simulated under LG+C60+F
- n ranges from 1 to 10 to model different levels of LBA artefact

Wang et al. (2018) Syst. Biol.

Markov chain Monte Carlo



(**a**) Symmetrical moves



(**b**) Asymmetrical moves

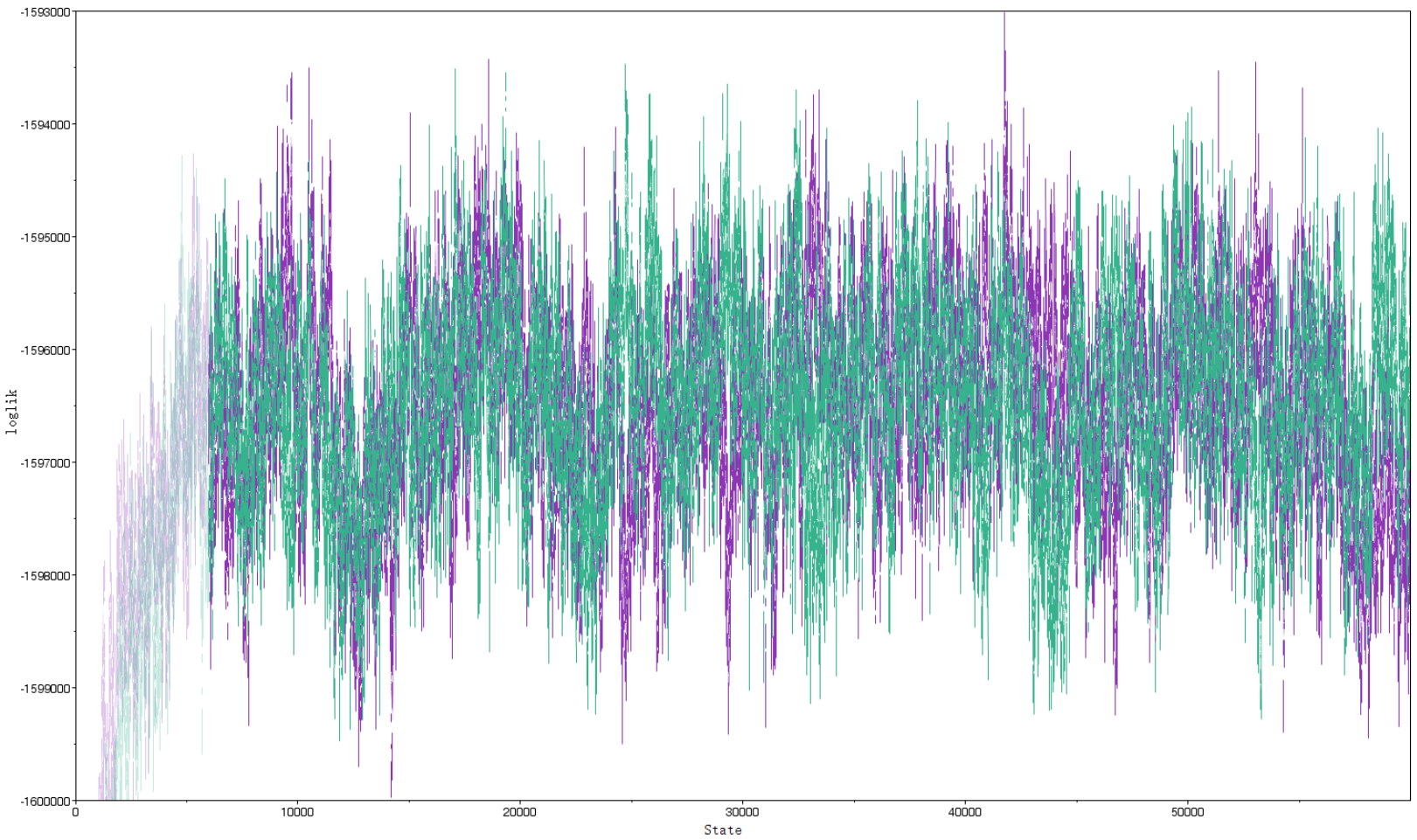
When does it finish?

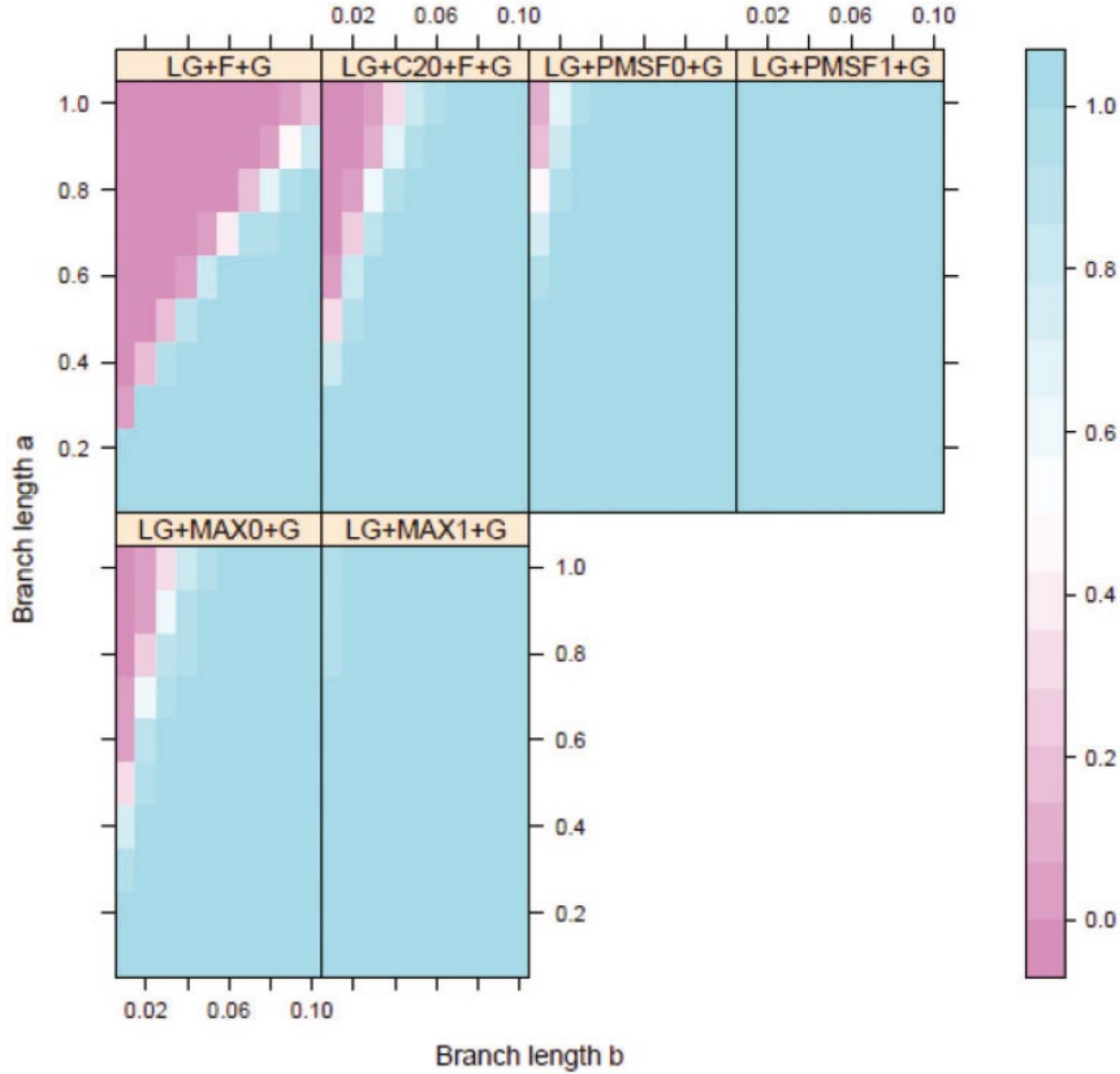
- Check for convergence between runs

- Visually:
  - TRACER (http://beast.community/tracer)
  - AWTY (http://king2.scs.fsu.edu/CEBProjects/awty/awty_start.php)
  - Graphylo (https://github.com/wrf/graphphylo)

- bpcomp (in PhyloBayes)

When does it finish?




LG+F+G
LG+C20+F+G
LG+PMSF0+G
LG+PMSF1+G
LG+MAX0+G
LG+MAX1+G
0.02
0.06
0.10
0.2
0.4
0.6
0.8
1.0
0.0
Branch length a
Branch length b