

SRT
*Sequence model with Rates
and Times*

Martin Linder, UU

Tom Britton, SU

Örjan Åkerborg, KTH

Jens Lagergren, KTH,

Bengt Sennblad

What?

- Substitution rate evolution
 - Molecular clock (Zuckerkandl & Pauling, 1962)
 - Relaxed clock
- Divergence times
- Sequence evolution

Why?

- Avoid Molecular clock assumption
- Allows estimation of:
 - Divergence times
 - Substitution rates
- Biologically relevant priors for edge lengths

Syst. Biol. 54(3):455–470, 2005
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150500945313

Branch-Length Prior Influences Bayesian Posterior Probability of Phylogeny

ZIHENG YANG¹ AND BRUCE RANNALA²

¹*Department of Biology, University College London, Darwin Building, Gower Street, London WC1E 6BT, United Kingdom; E-mail: z.yang@ucl.ac.uk*

²*Genome Center and Section of Evolution and Ecology, University of California Davis, One Shields Avenue, Davis, California 95616, USA*

Abstract.— The Bayesian method for estimating species phylogenies from molecular sequence data provides an attractive alternative to maximum likelihood with nonparametric bootstrap due to the easy interpretation of posterior probabilities for trees and to availability of efficient computational algorithms. However, for many data sets it produces extremely high posterior probabilities, sometimes for apparently incorrect clades. Here we use both computer simulation and empirical data analysis to examine the effect of the prior model for internal branch lengths. We found that posterior probabilities for trees and clades are sensitive to the prior for internal branch lengths, and priors assuming long internal branches cause high posterior probabilities for trees. In particular, uniform priors with high upper bounds bias Bayesian clade probabilities in favor of extreme values. We discuss possible remedies to the problem, including empirical and full Bayesian methods and subjective procedures suggested in Bayesian hypothesis testing. Our results also suggest that the bootstrap proportion and Bayesian posterior probability are different measures of accuracy, and that the bootstrap proportion, if interpreted as the probability that the clade is true, can be either too liberal or too conservative. [Fair-balance paradox; Lindley's paradox; model selection; molecular phylogenetics; posterior probabilities; prior; star tree paradox.]

When?

- Gillespie, J. H. The causes of molecular evolution Oxford University Press, 1991
- Autocorrelated models
 - Thorne, Kishino and co-workers, 1998,2001,2002
 - Huelsenbeck et al 2001
 - Yang and co-workers 2001
- Independent and identically distributed (iid)
 - Linder 2003, and submitted, Drummond 2006, Yang & Rannala 2006, Lepage 2007

How

- A model for rate variation over the tree
 - AutoCorrelated rates
 - $r_u \sim \text{LogN}(r_{p(u)}, \nu)$
 - Independent and Identically Distributed rates
 - $r_u \sim f(\text{mean}, \text{variance})$
 - f - Gamma, LogN, InvG
- Model for divergence time
 - BD process, Dirichlet, Uniform, Coalescence
- Substitution model
 - JC69, HKY85, GTR, JTT etc.

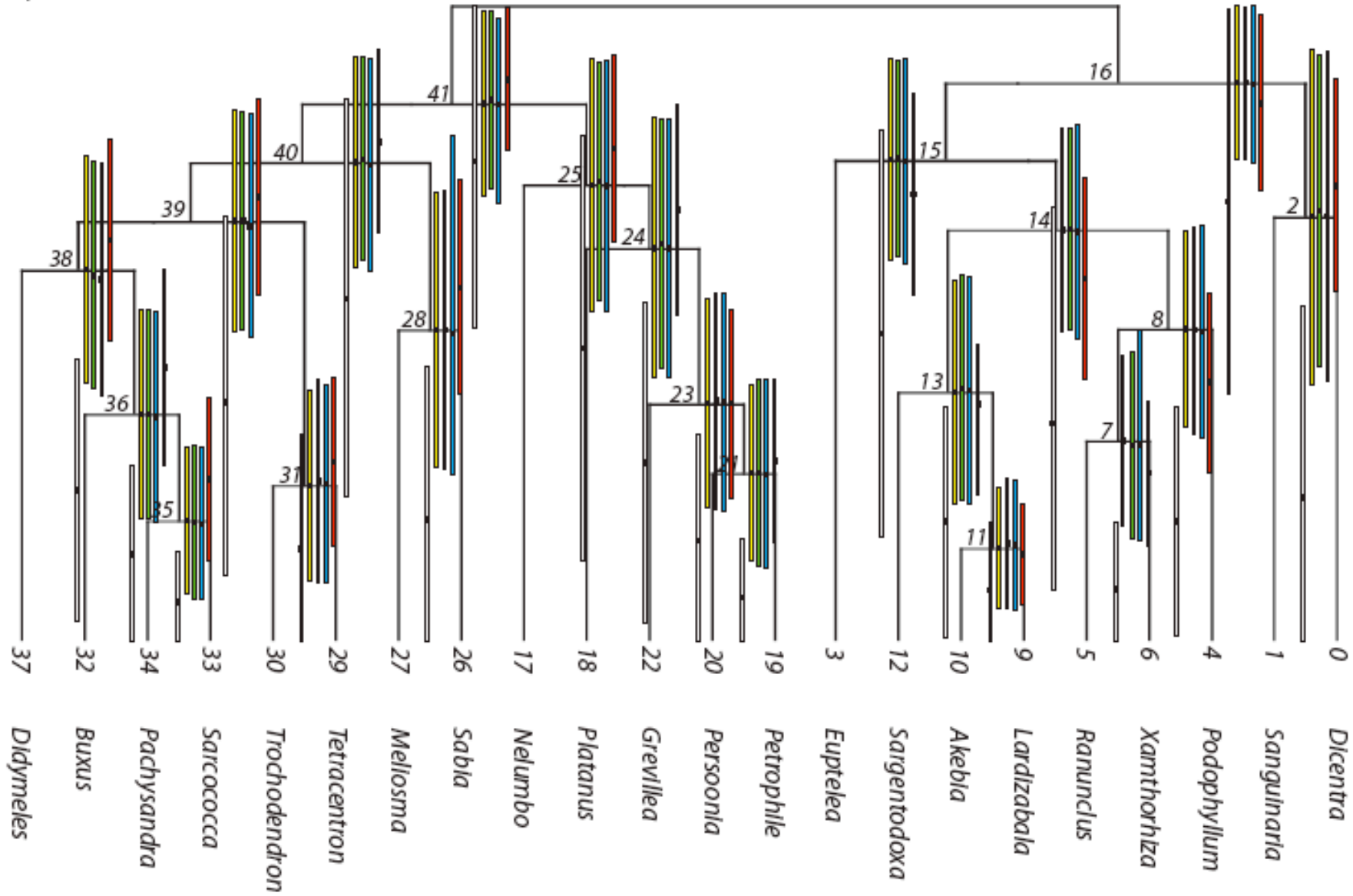
Which?

- Comparison of rate models (Linder 2003, submitted)
 - No difference in performance AC vs iid

Data	<i>iid/Γ</i>		<i>iid/IG</i>		<i>iid/LN</i>		<i>AC</i>		<i>clock</i>
	$\log \bar{L}_m$	$\log \frac{\bar{L}_m}{L_m}$	$\log \bar{L}_m$	$\log \frac{\bar{L}_m}{L_m}$	$\log \bar{L}_m$	$\log \frac{\bar{L}_m}{L_m}$	$\log \bar{L}_m$	$\log \frac{\bar{L}_m}{L_m}$	$\log \frac{\bar{L}_m}{L_m}$
Non-informative time prior									
eudicot	-4945.25	0.06	-4945.19	0	-4945.20	0.01	-4945.48	0.29	48.11
simian	-12186.5	0.8	-12186.6	0.9	-12186.6	0.9	-12185.7	0	11.4
hominid	-4305.22	0	-4305.24	0.02	–	–	-4305.37	0.15	1.98
Calibrated time prior									
simian	-12163.6	0.9	–	–	–	–	-12162.7	0	28.1

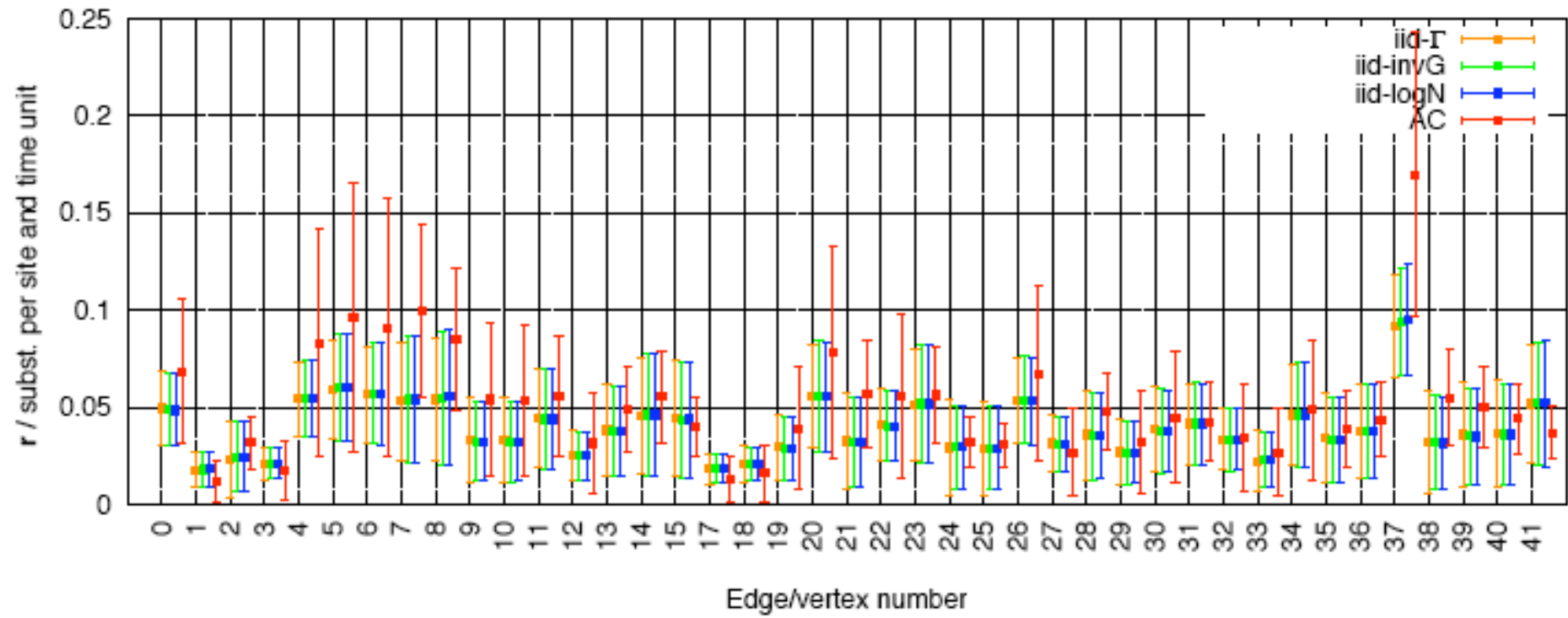
rbcL -- time estimates

a)



rbcL--rate estimates

b)

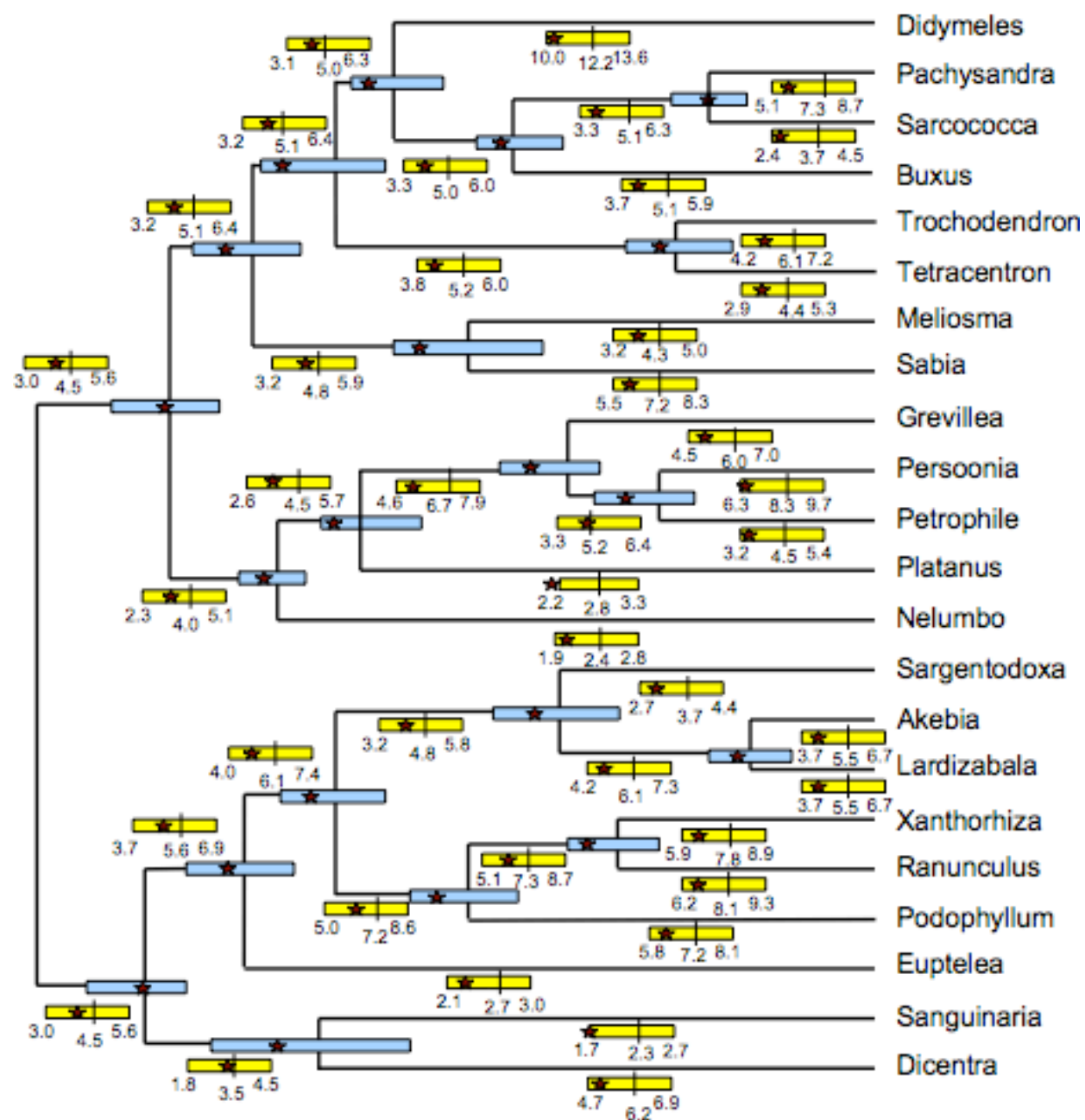


Which?

- Comparison of rate models (Linder 2003, submitted)
 - No difference in performance AC vs iid
 - Estimate diff AC vs iid
 - No correlation between adjacent rates
- Lepage et al. 2007
 - Support for AC for most investigated data sets
 - Reasons for deviant results
 - Tree size, diff data, aa vs. DNA, diff Bayes factor approx.

How long?

- Faster implementation (Åkerborg et al. 2008)
 - MapDP
 - Discretization of divergence time space
 - Maximum A Posteriori (MAP) framework
 - *rbcL* estimates similar



How long?

- Faster implementation (Åkerborg, submitted)
 - Map-PD
 - Discretization of divergence time space
 - Maximum A Posteriori (MAP) framework
 - *rbcL* estimates similar
 - Gene tree reconstruction
 - mtDNA, data Yang & Yoder, 2003

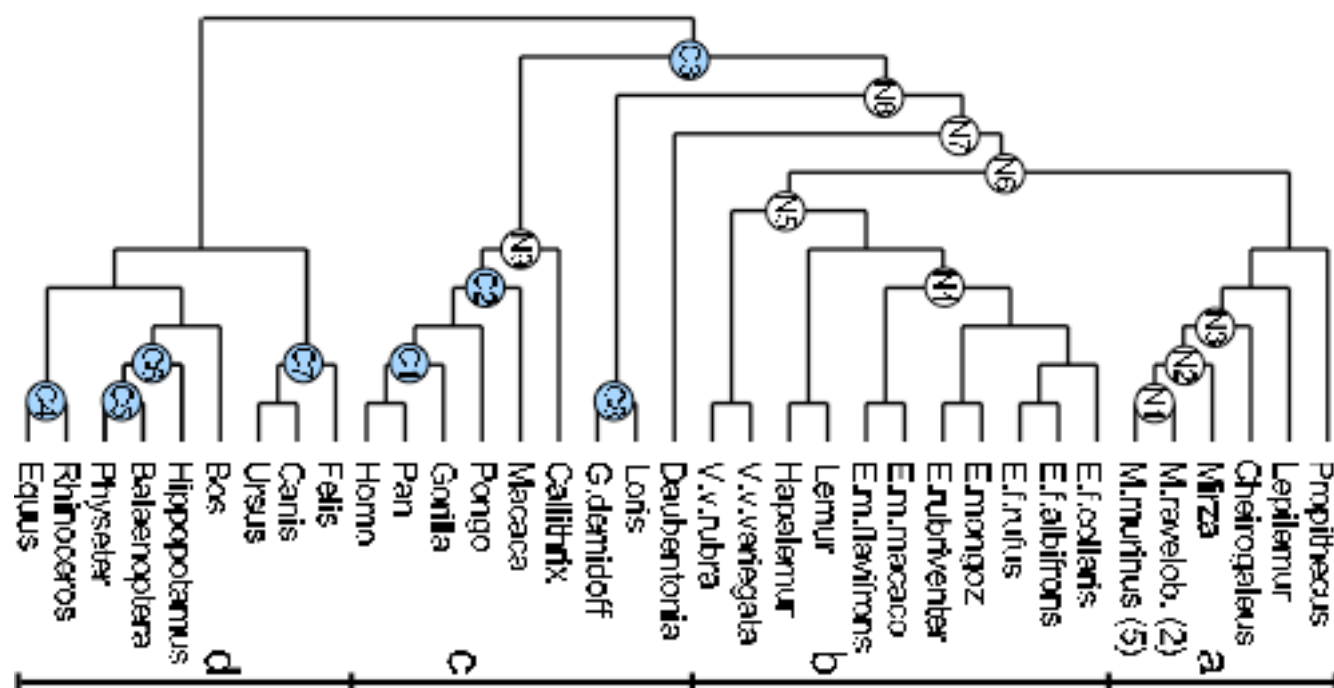


Table 3: Phylogeny inference performed on a mtDNA dataset. 100 times MAP phylogeny inference on step by step larger proportions of the tree shown in Figure 5.

Tree	Known ¹ tree	best	Inferred ² median	worst	Success ³ %
a	-8843.8	-8816.8	-8829.0	-8953.7	72
a + b	-13381.9	-13345.0	-13369.6	-13472.5	65
a + b + c	-22054.2	-21977.7	-22025.0	-22339.3	73
a + b + c + d	-30327.7	-30235.7	-30315.5	-30658.7	58

¹Reference log-likelihood value obtained with the tree topology in Figure 5.

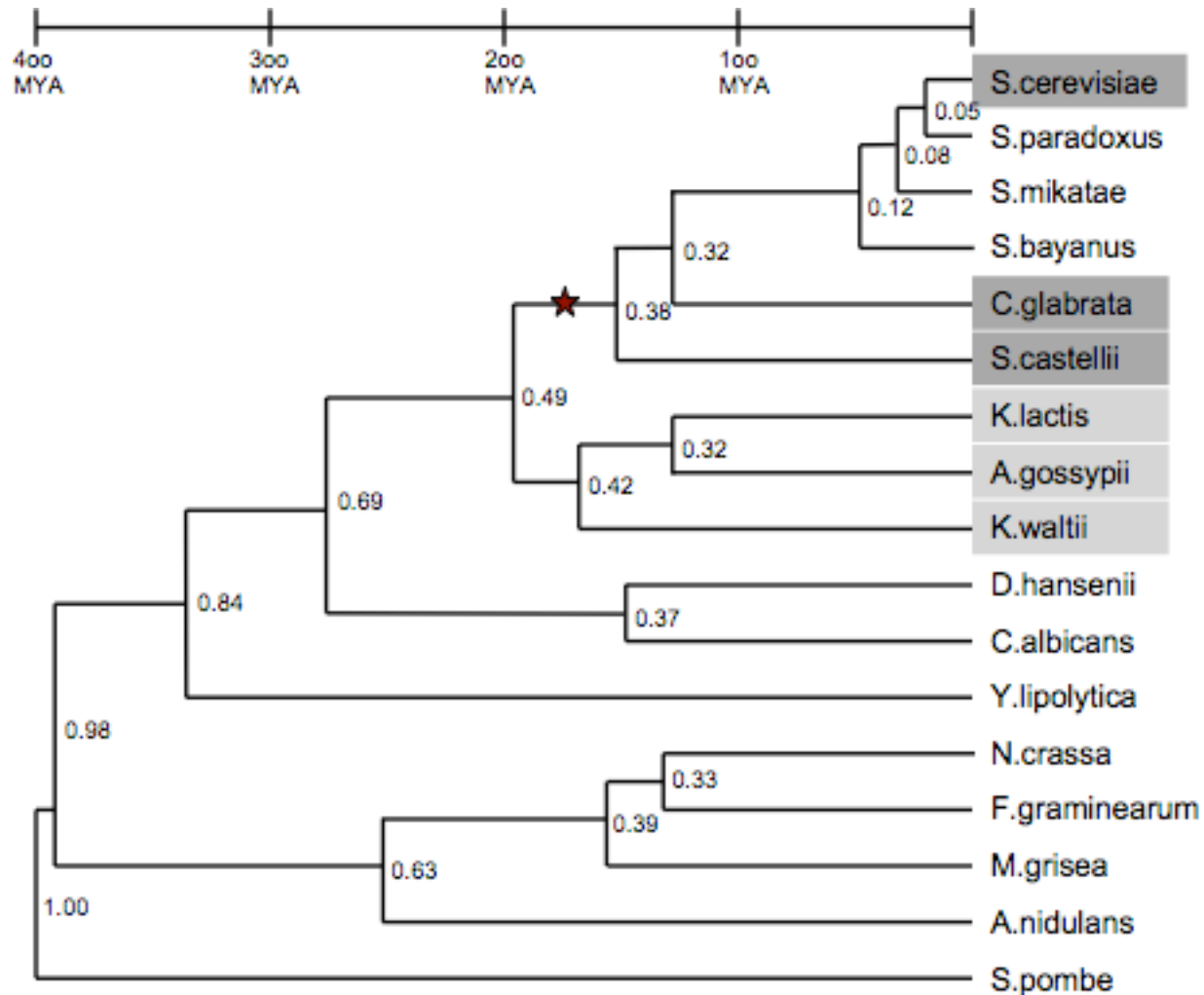
²Log-likelihood value obtained starting from a random tree topology.

³Percentage of runs finding a tree with likelihood at least as good as for the tree topology in Figure 5.

How long?

- Faster implementation (Åkerborg, submitted)
 - Map-PD
 - Discretization of divergence time space
 - Maximum A Posteriori (MAP) framework
 - *rbcL* estimates similar
 - Gene tree reconstruction
 - mtDNA, data Yang & Yoder, 2003
 - Ascomycete divergence time estimates
 - 1100 orthogroups from Wapinski et al. 2007

Ascomycete time estimates



Summary

- primeSRT
 - Integrated model
 - Relaxed clock, iid (or AC)
 - BD divergence time prior
 - Standard substitution models
 - Fast, discretized, algorithm
 - Divergence time estimation
 - Gene tree reconstruction