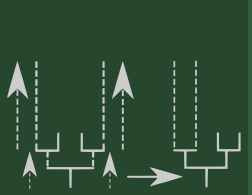


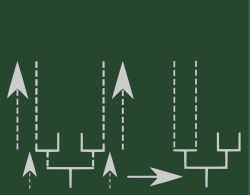
Maximum likelihood (ML) inconsistency

in terms of long branch attraction

Patrick Kück, Wolfgang Wägele

Bioinformatics
Forschungsmuseum Alexander Koenig

22nd February

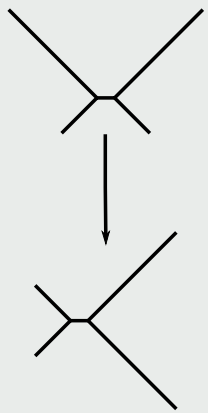


Topological Bias

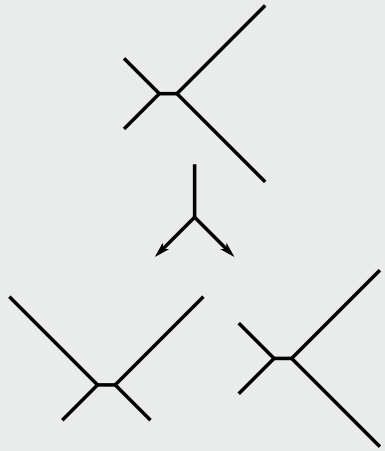
Felsenstein 1978

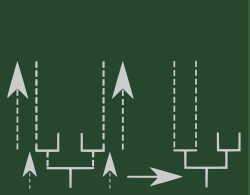
- Phylogenetic methods become inconsistent when model assumptions are **strongly** violated

Felsenstein Zone



Farris Zone

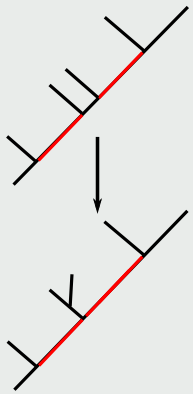




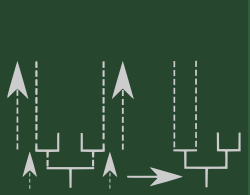
3 classes of LBA

Wägele & Mayer 2007

LBA I (Symplesiomorphy Trap)



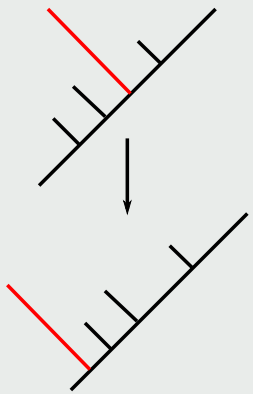
- plesiomorphy trap
- synapomorphies substituted
- increase of plesiomorphies / apomorphies ratio
- paraphyletic grouping
- use of more taxa



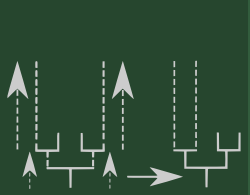
3 classes of LBA

Wägele & Mayer 2007

LBA II (Signal Erosion)



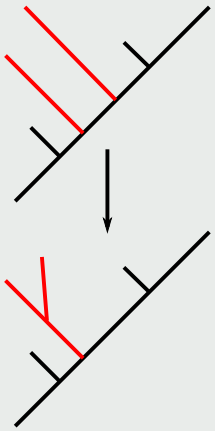
- phylogenetic signal erosion
- shared plesiomorphies with distantly related taxa
- synapomorphies substituted
- long branch slips down
- use of genes with lower rates



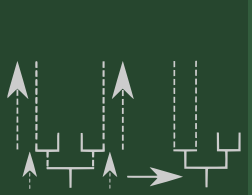
3 classes of LBA

Wägele & Mayer 2007

LBA III (Parallel Substitutions)

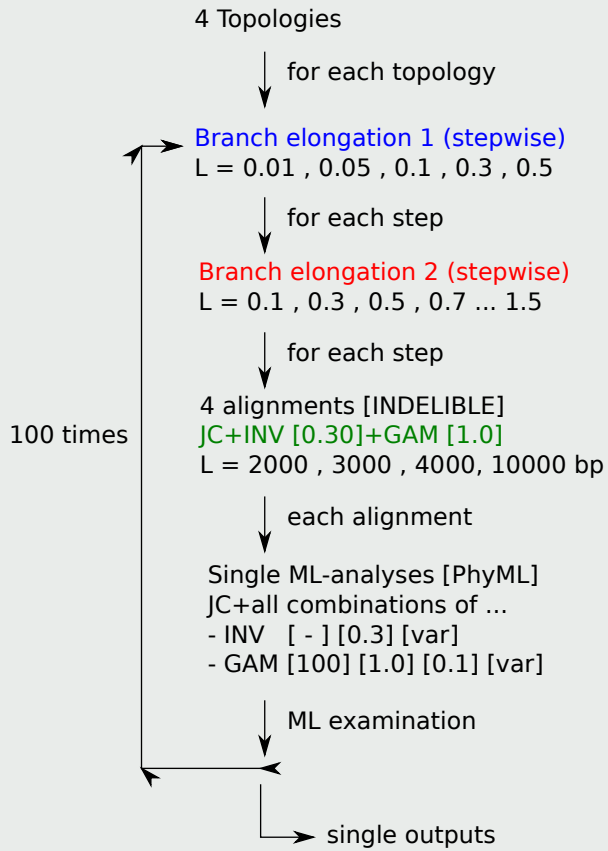


- identical character states evolve independently
- high number of convergent characters
- branches can form nonsense clades

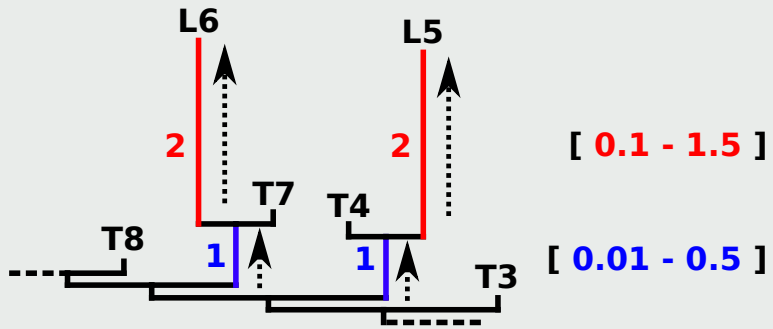


Long-Branch-Testpipeline (LoBraTe)

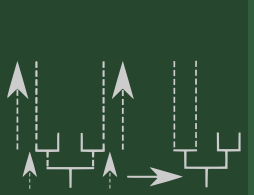
Setup of LoBraTe



Branch Elongation 1&2

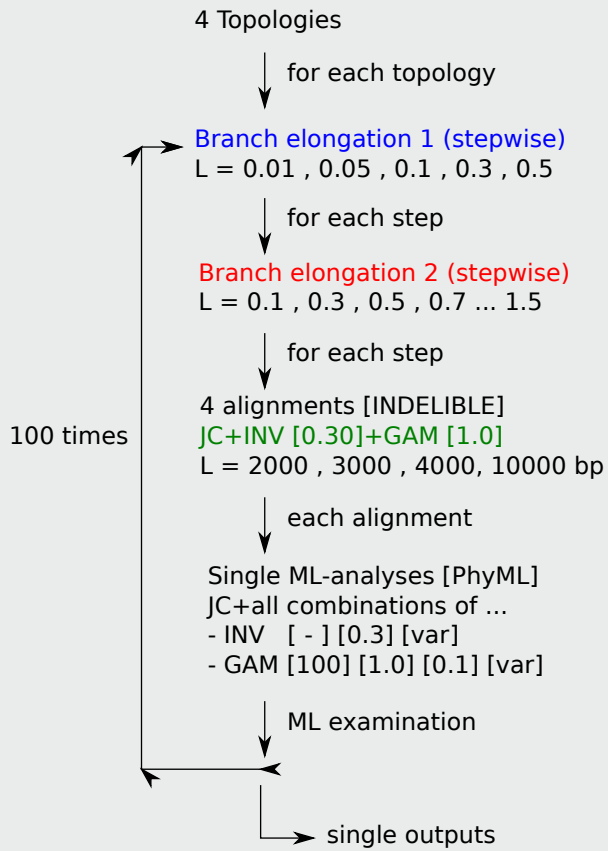


- 768,000 single data sets

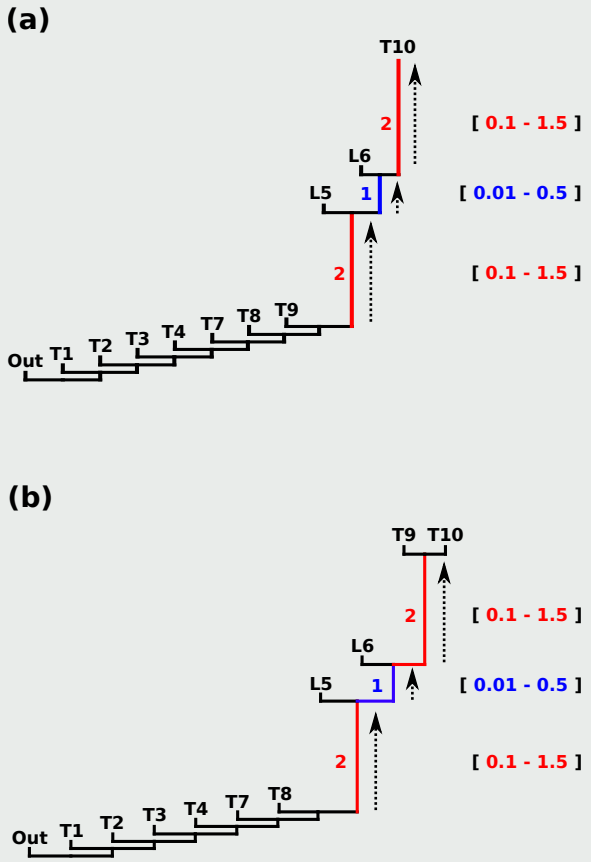


Long-Branch-Testpipeline (LoBraTe)

Setup of LoBraTe

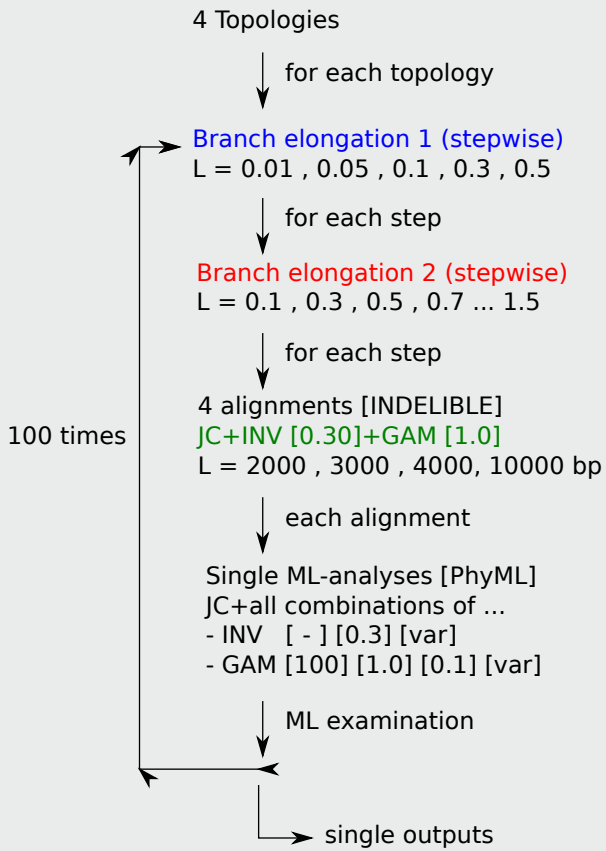


Topologies for LBA I

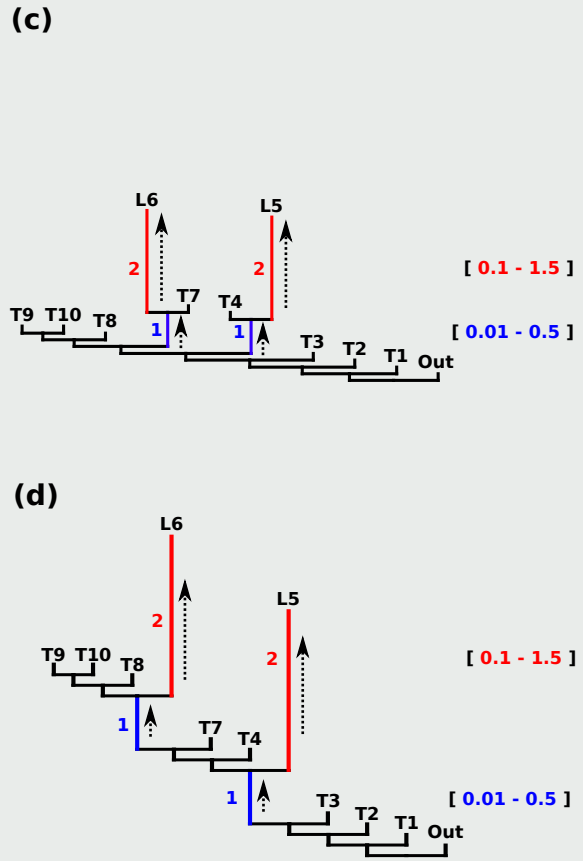


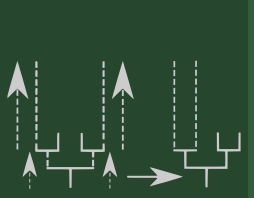
Long-Branch-Testpipeline (LoBraTe)

Setup of LoBraTe



Topologies for LBA II&III



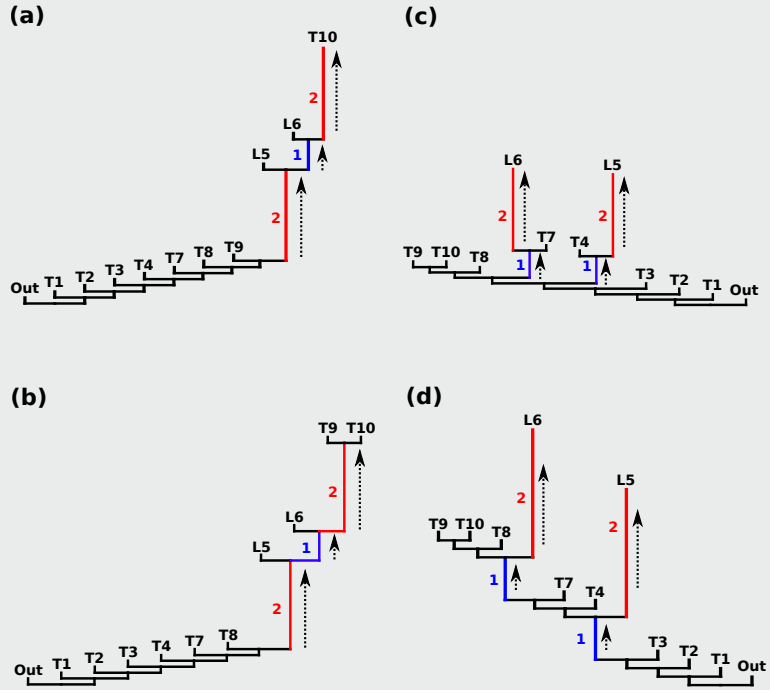


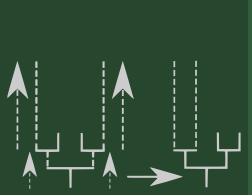
Model Setup

PHYML+TBR

Jukes-Cantor		
Γ		ρ_{inv}
$\alpha[0.1]$		
$\alpha[0.1]$	+	$\theta[0.3]$
$\alpha[0.1]$	+	$\theta[var]$
$\alpha[1.0]$		
$\alpha[\mathbf{1.0}]$	+	$\theta[\mathbf{0.3}]$
$\alpha[1.0]$	+	$\theta[var]$
$\alpha[100]$		
$\alpha[100]$	+	$\theta[0.3]$
$\alpha[100]$	+	$\theta[var]$
$\alpha[var]$		
$\alpha[var]$	+	$\theta[0.3]$
$\alpha[var]$	+	$\theta[var]$

Topologies



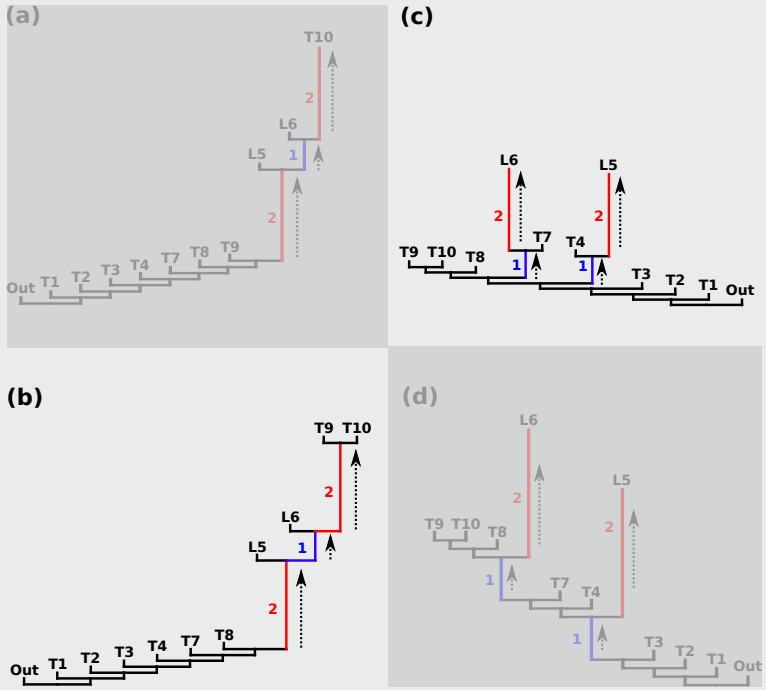


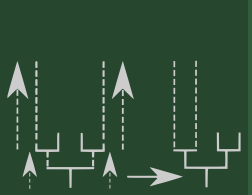
Model Setup

PHYML+TBR

	Jukes-Cantor	
Γ		ρ_{inv}
$\alpha[1.0]$	+	$\theta[0.3]$
$\alpha[100]$		
$\alpha[var]$		
$\alpha[var]$	+	$\theta[var]$

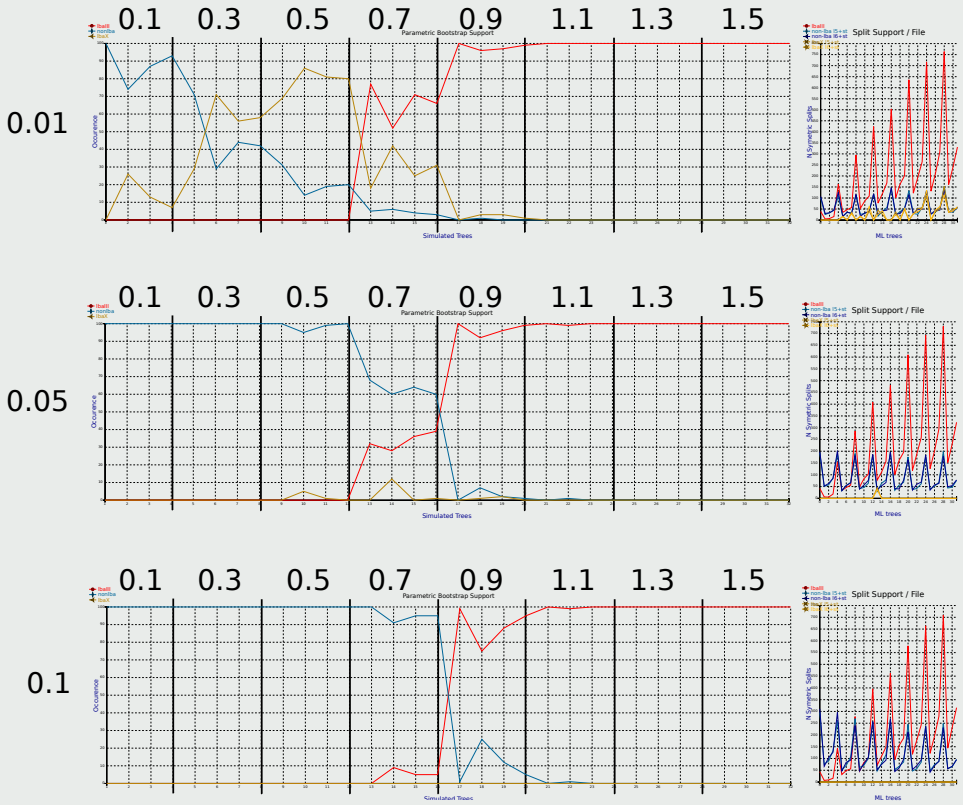
Topologies

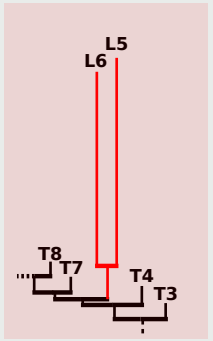


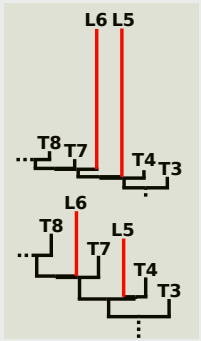


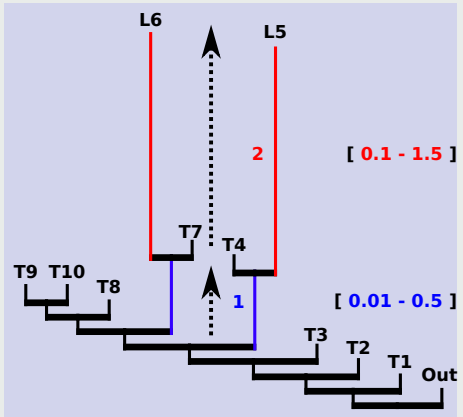
LoBraTe output

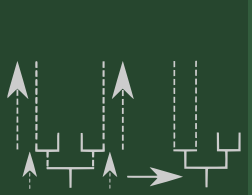
parametric bootstrap|splits (LBA II&III) JC+α[100]





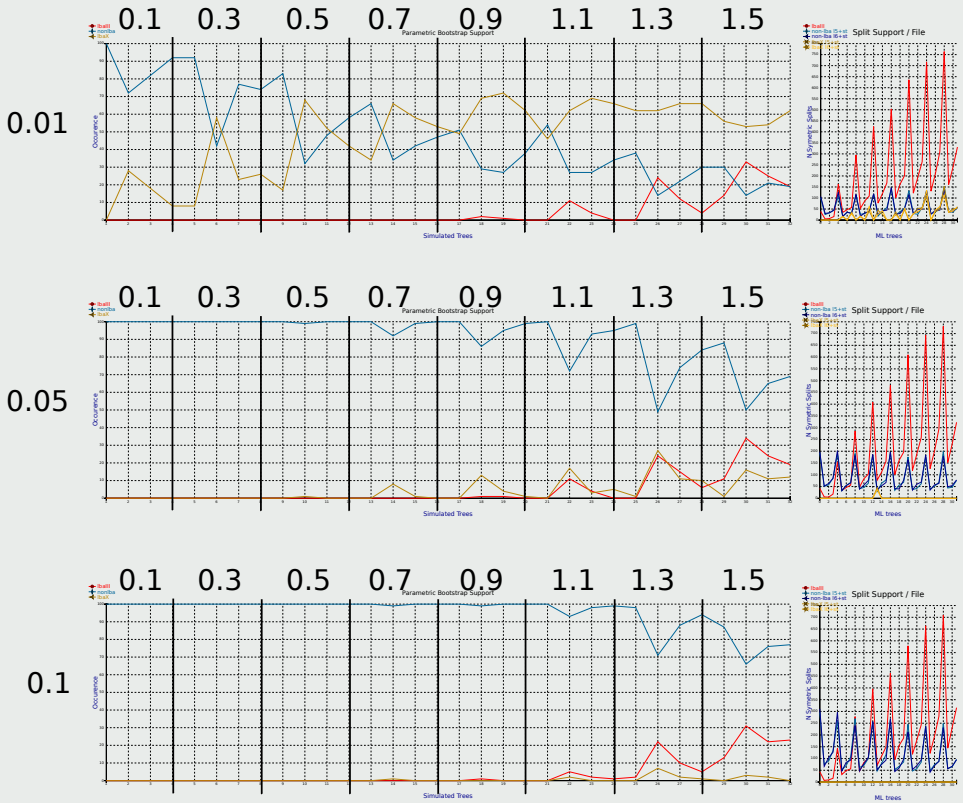


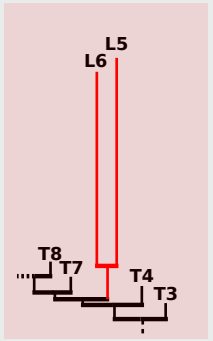


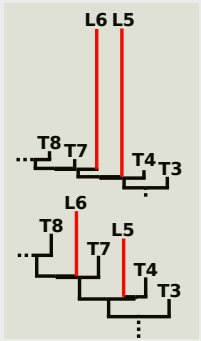


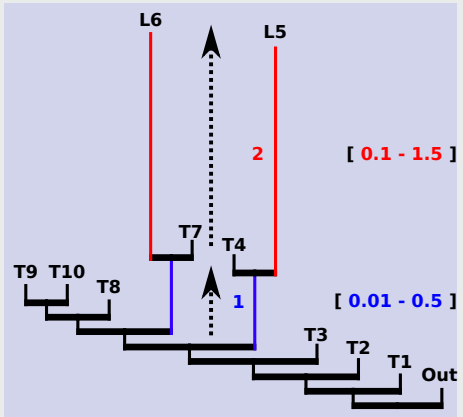
LoBraTe output

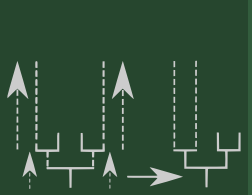
parametric bootstrap|splits (LBA II&III) JC+$\alpha[var]$





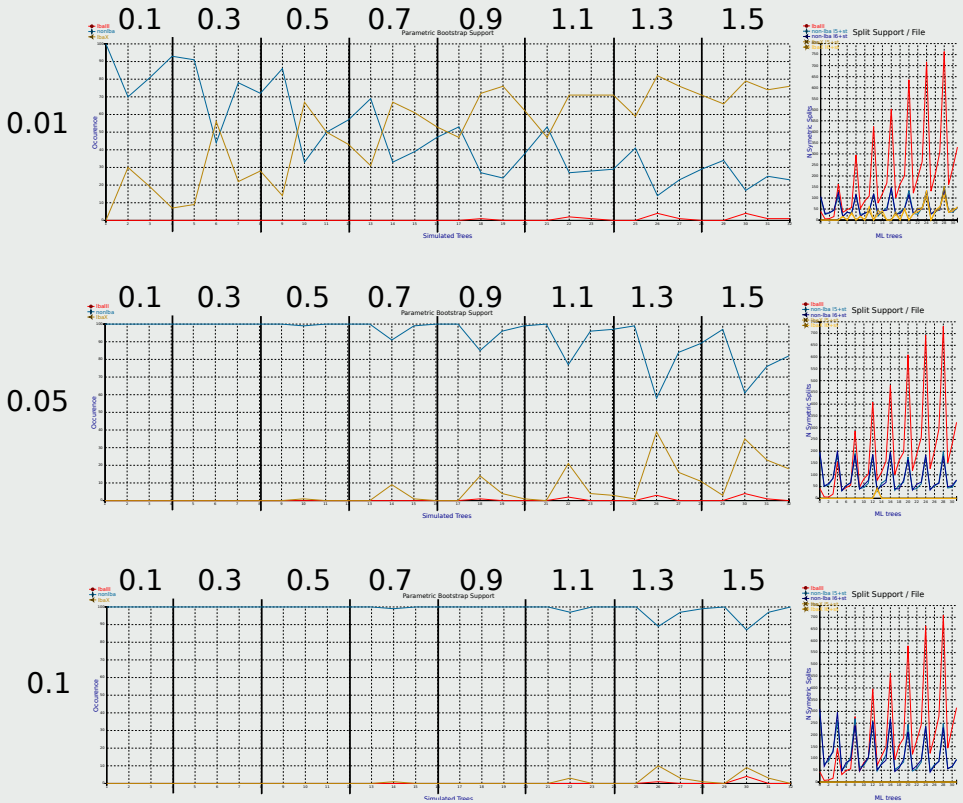


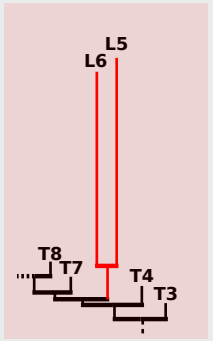


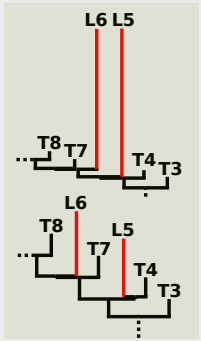


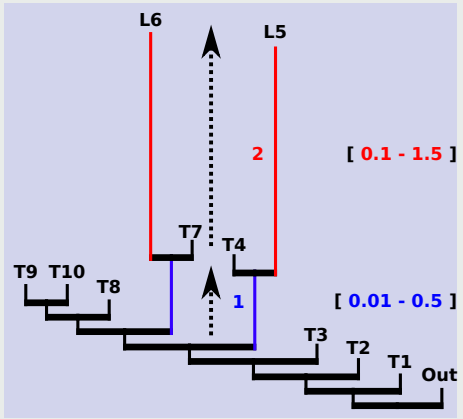
LoBraTe output

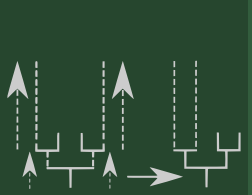
parametric bootstrap|splits (LBA II&III) JC+$\alpha[var]+\theta[var]$





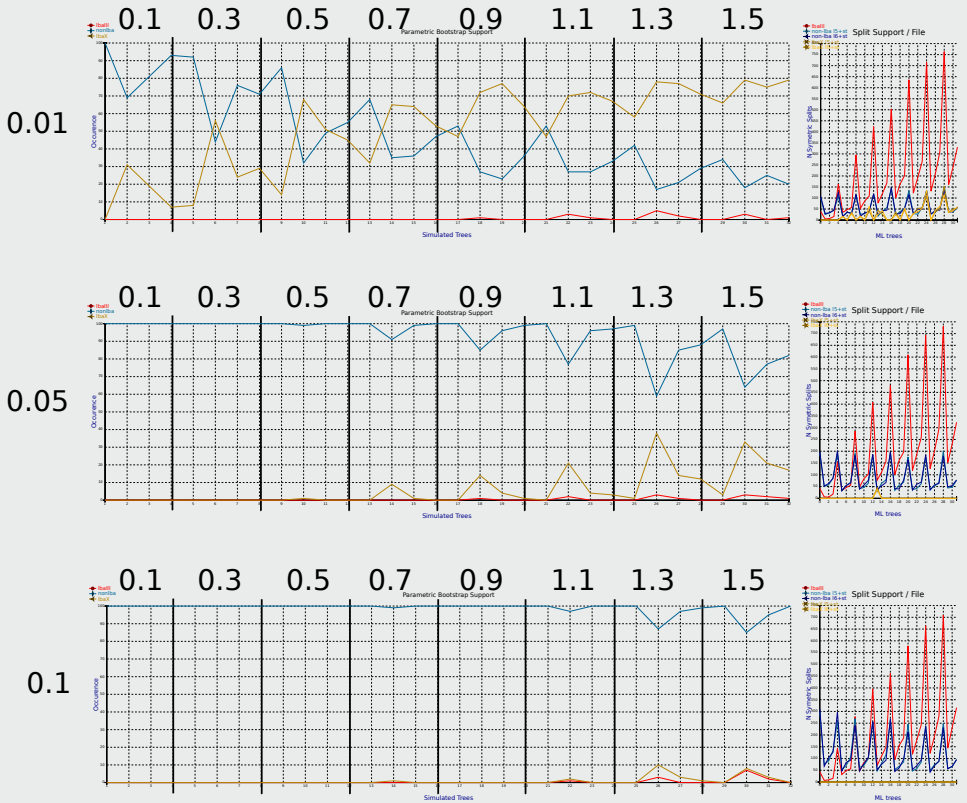


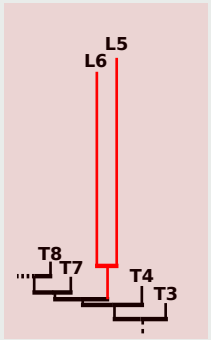


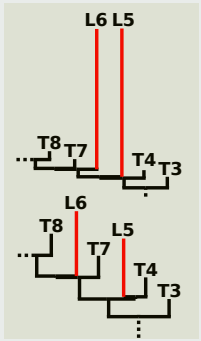


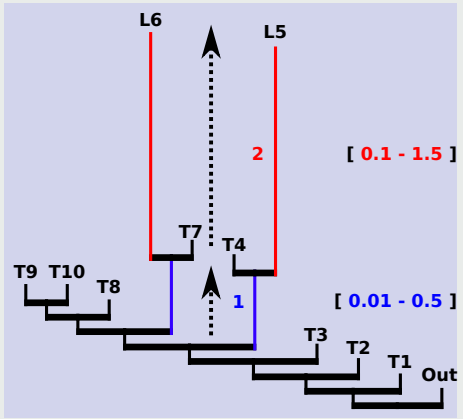
LoBraTe output

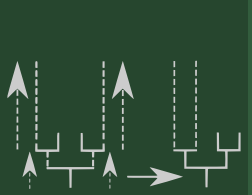
parametric bootstrap|splits (LBA II&III) JC+$\alpha[1.0]+\theta[0.3]$





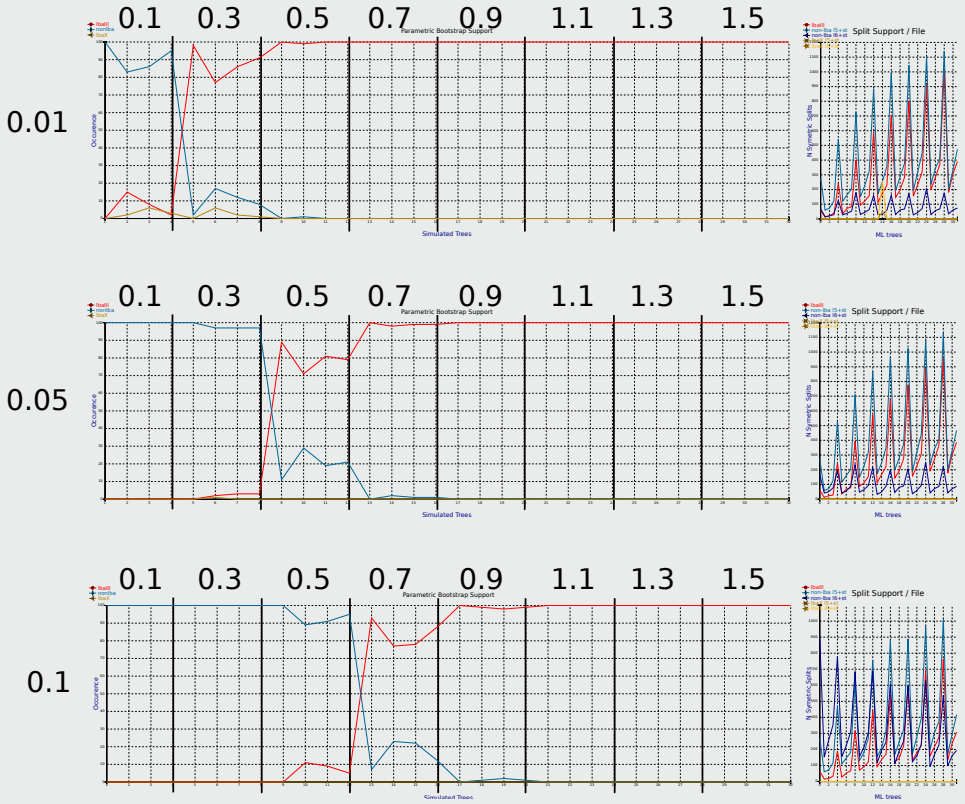


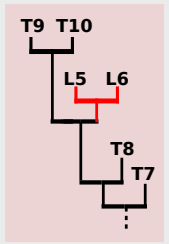


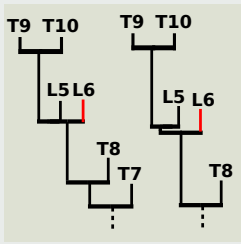


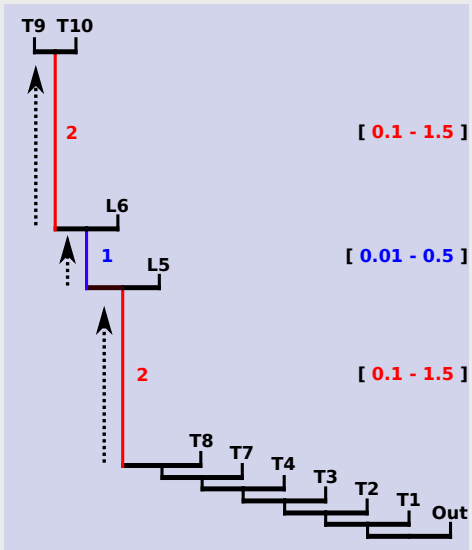
LoBraTe output

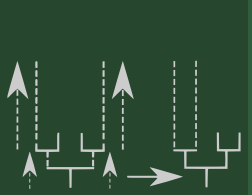
parametric bootstrap|splits (LBA I) JC+α[100]





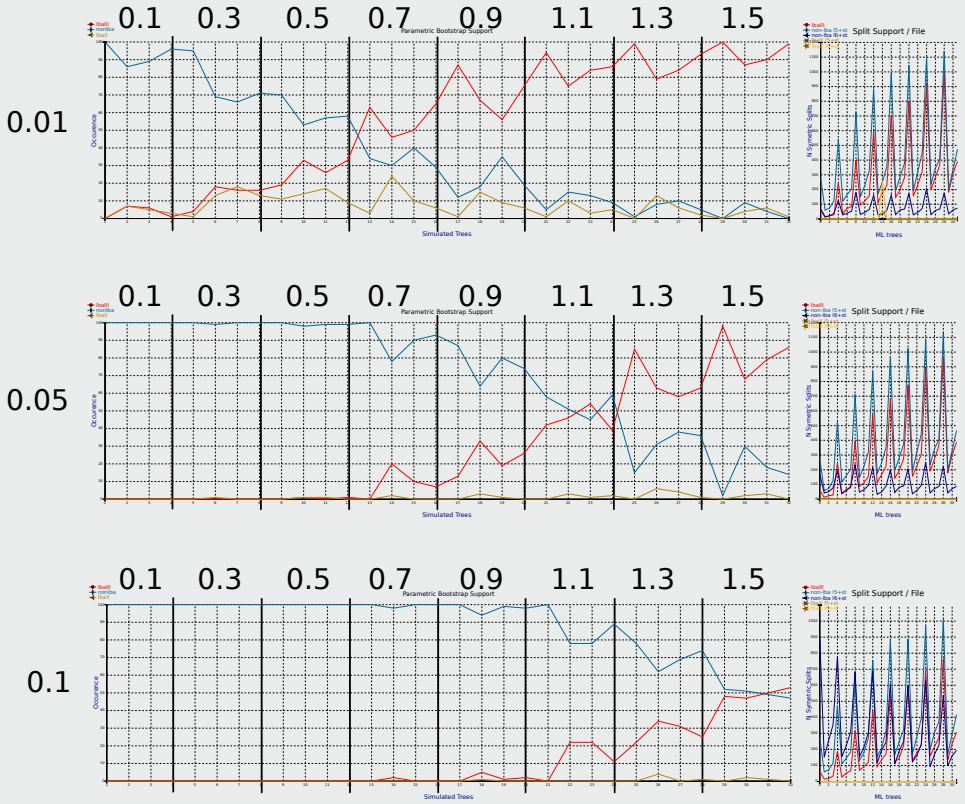


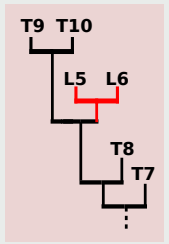


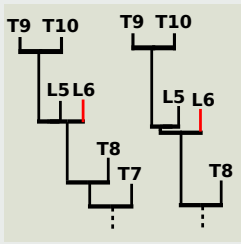


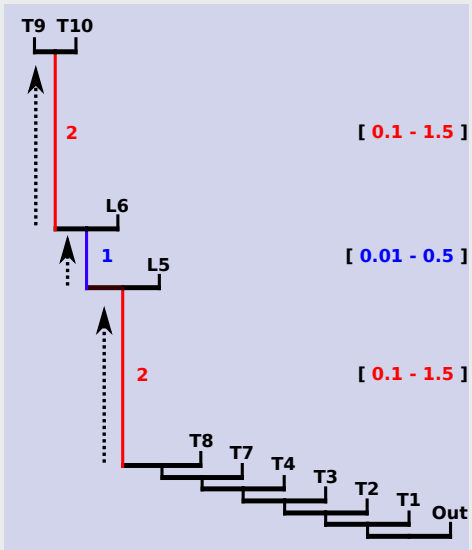
LoBraTe output

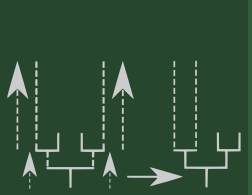
parametric bootstrap|splits (LBA I) JC+$\alpha[var]$





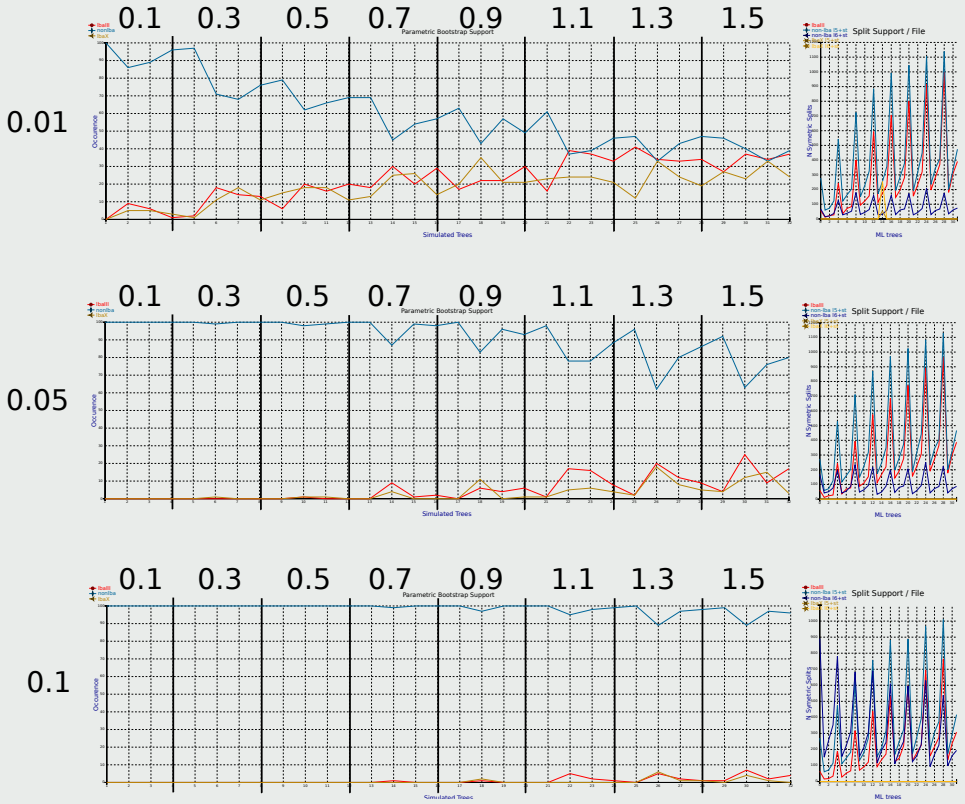


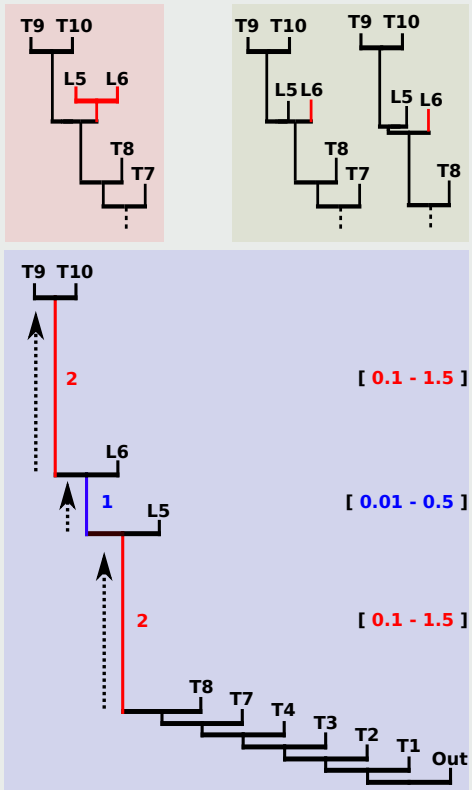


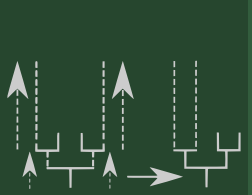


LoBraTe output

parametric bootstrap|splits (LBA I) JC+$\alpha[var] + \theta[var]$

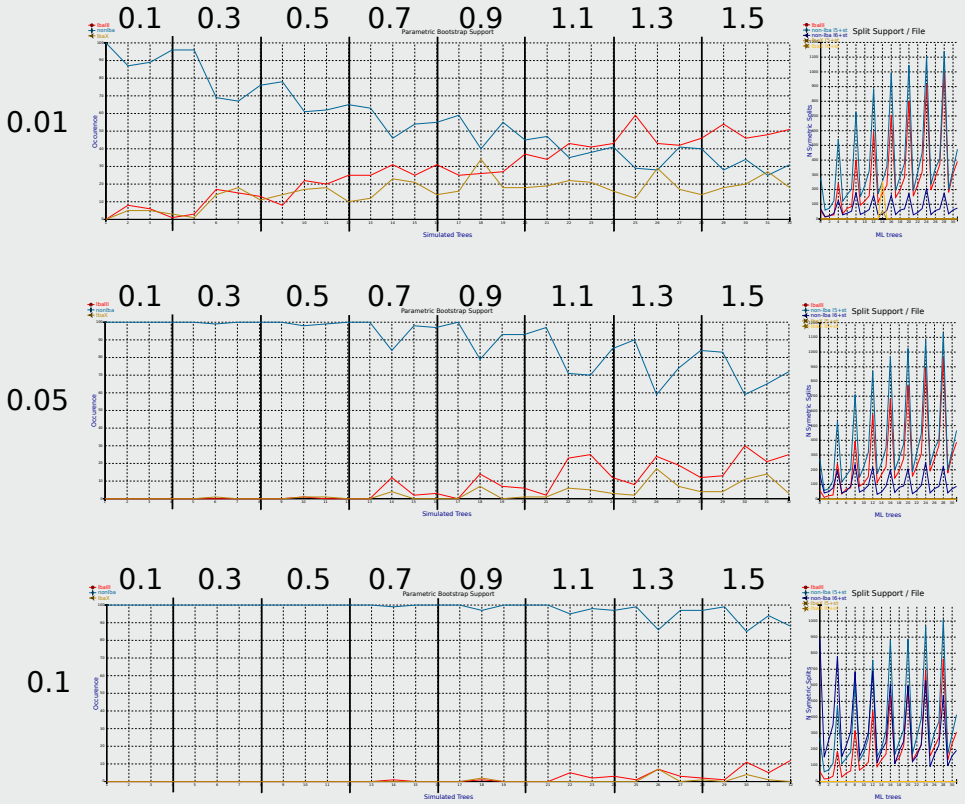


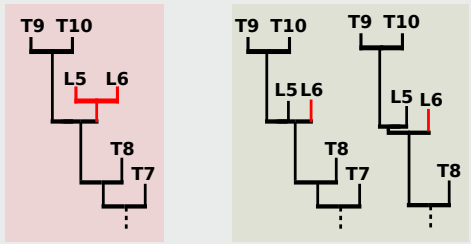


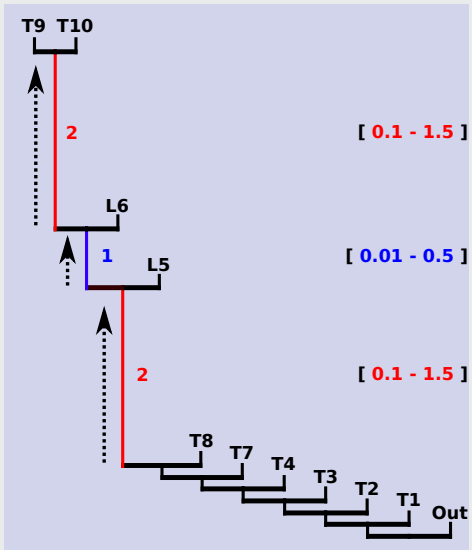


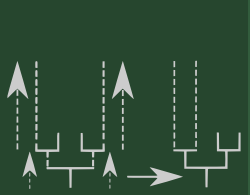
LoBraTe output

parametric bootstrap|splits (LBA I) JC+$\alpha[1.0]+\theta[0.3]$



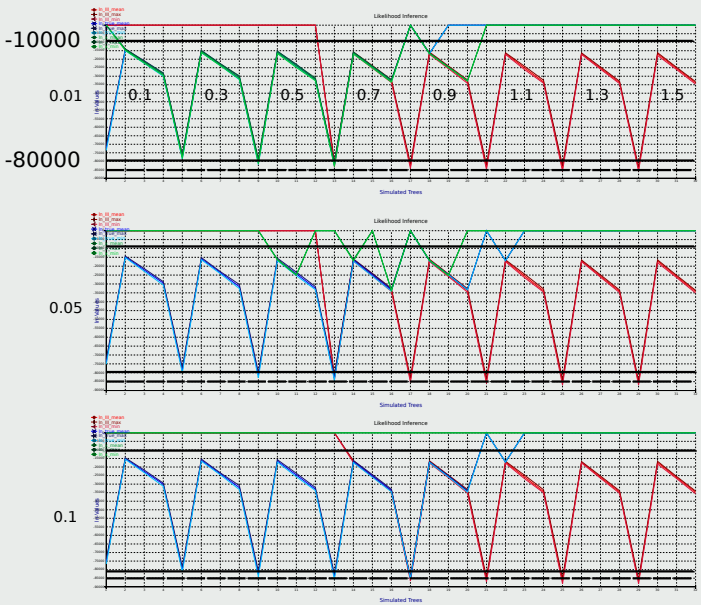




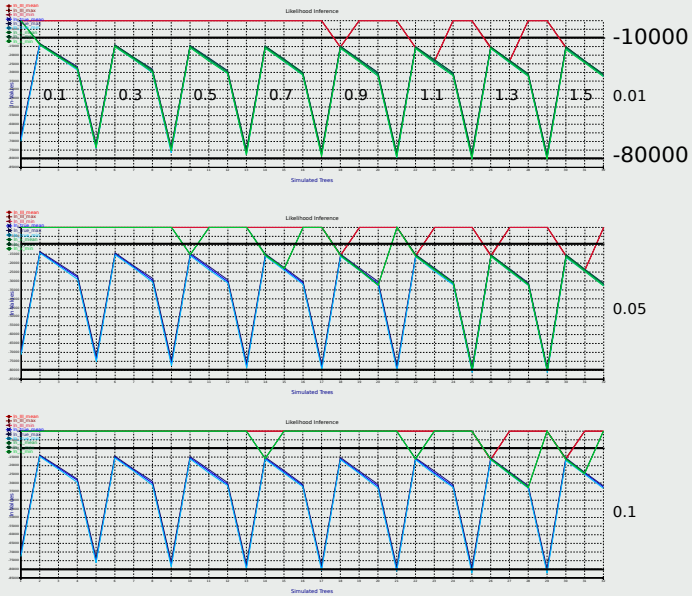


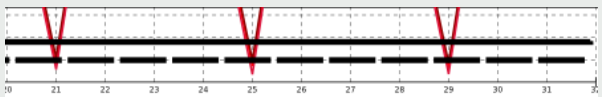
LoBraTe output: Mean Likelihood Values

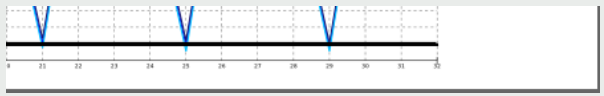
LBA II&III JC+$\alpha[100]$

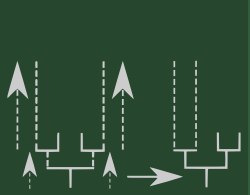


JC+$\alpha[var] + \theta[var]$



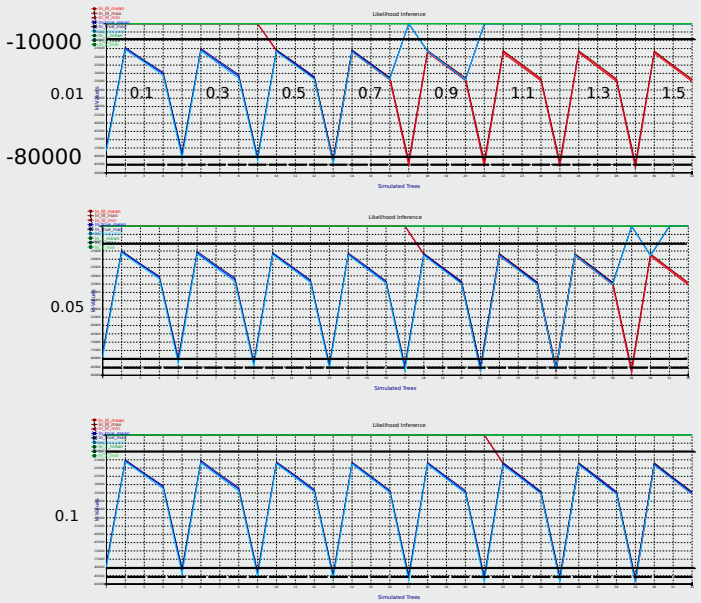


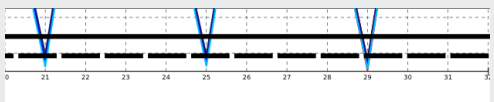




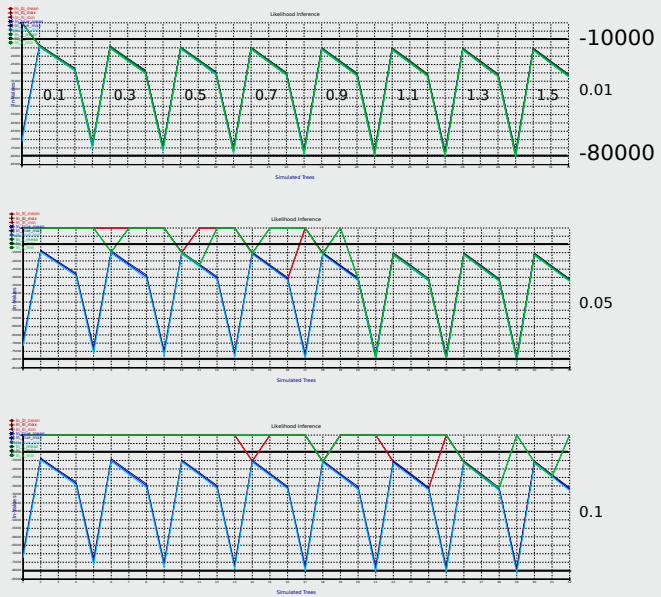
LoBraTe output: Mean Likelihood Values

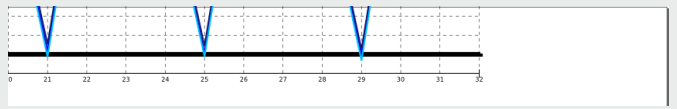
LBA I JC+$\alpha[100]$

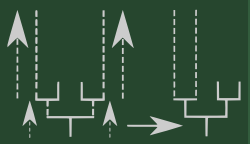




JC+$\alpha[var] + \theta[var]$



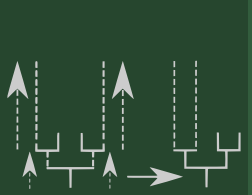




Conclusions

Among Site Rate Variation (ASRV)

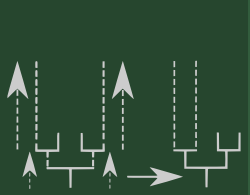
- ML methods are much more robust to LBA if among site rate variation is included
- ASRV including Γ and ρ_{inv} fit data better than Γ alone
- ASRV leads to improved likelihood scores
- Consistent parameter estimates possible with relative low number of taxa



Conclusions

Consistency of Maximum Likelihood (ML)

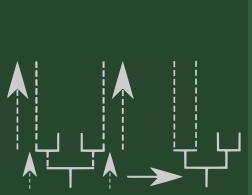
- LBA can afflict ML when models are not/weakly violated
- Large data are not immune to systematic error under correct models
- LBA trees show nearly identic likelihood scores than the true trees
- Near star-like subregions are not choosen at random from the possible resolved topologies



Conclusions

LBA I (symplesiomorphy trap)

- Class I long-branch effects can be shown in simulations
- Long internal branches did not lead to more robust ML estimates



Acknowledgement

Best thanks to the...

- ...Forschungsmuseum Koenig, Bonn
  - Bernhard Misof
  - Christoph Mayer
- ...Deutsche Forschungsgemeinschaft DFG
- ...Leibniz Gemeinschaft
- and especially to you !