A more realistic approach to simulating heterotachy and its effect on phylogenetic accuracy

Christoph Mayer

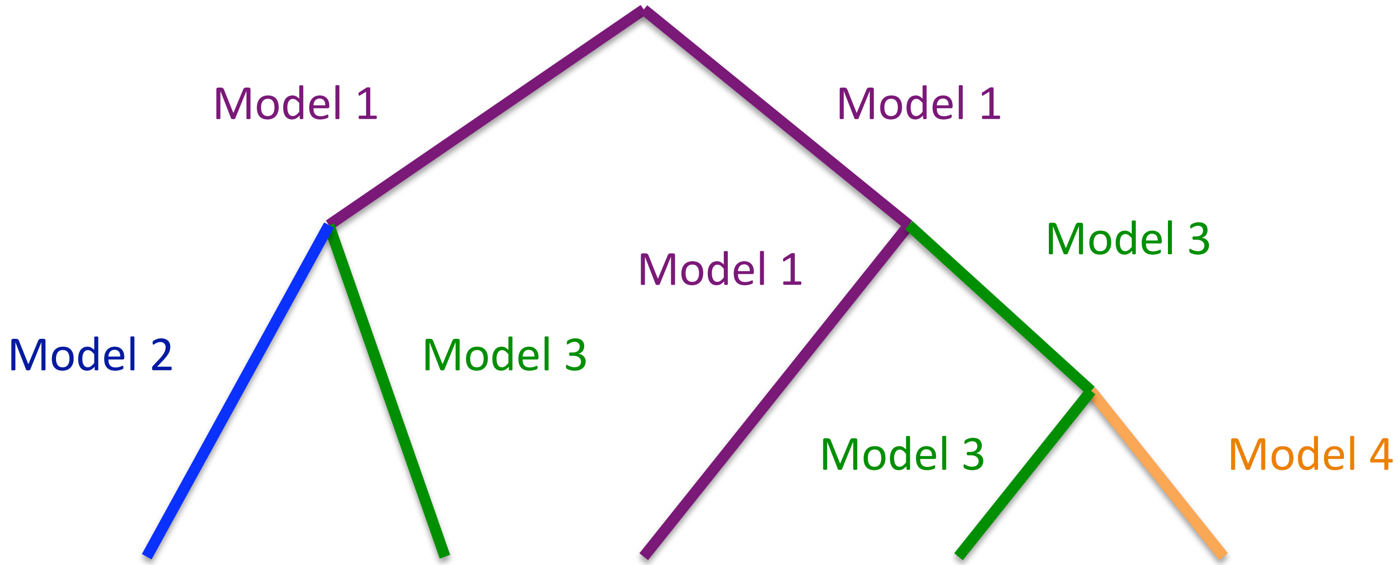
Stefan Richter

Ruhr Universität Bochum, Germany

MIEP-08

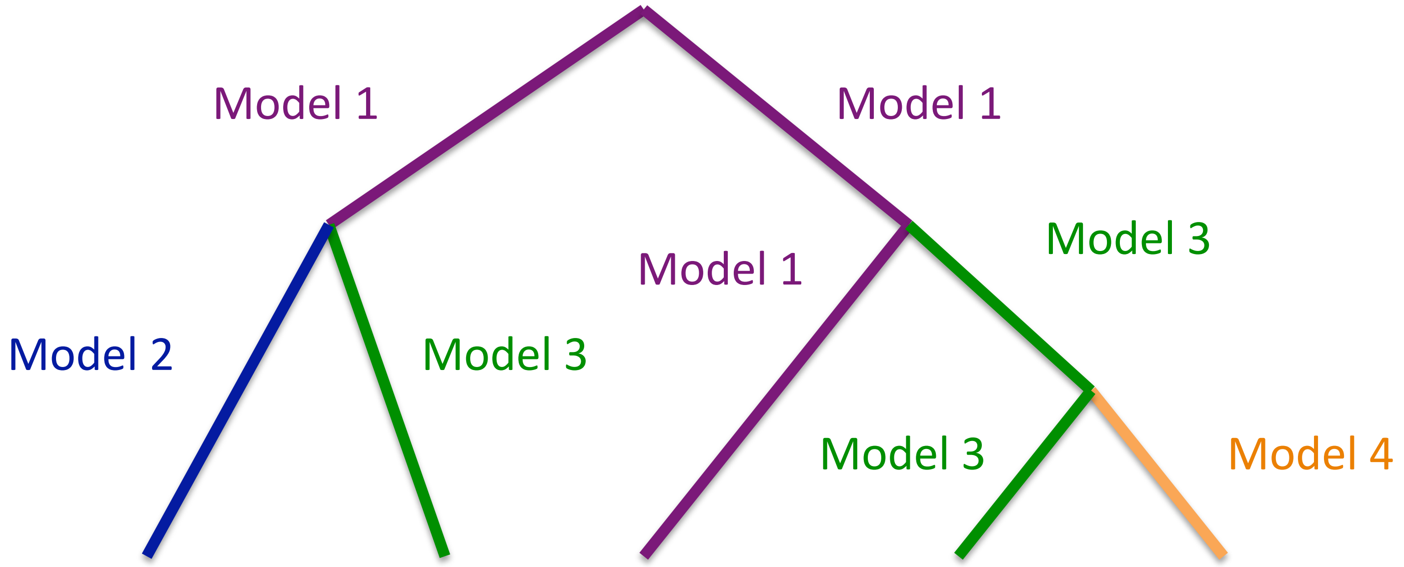
Simulating data sets with multiple models

We developed a simulation program which allows simulating data sets along a given tree with different substitution models along different branches of a tree



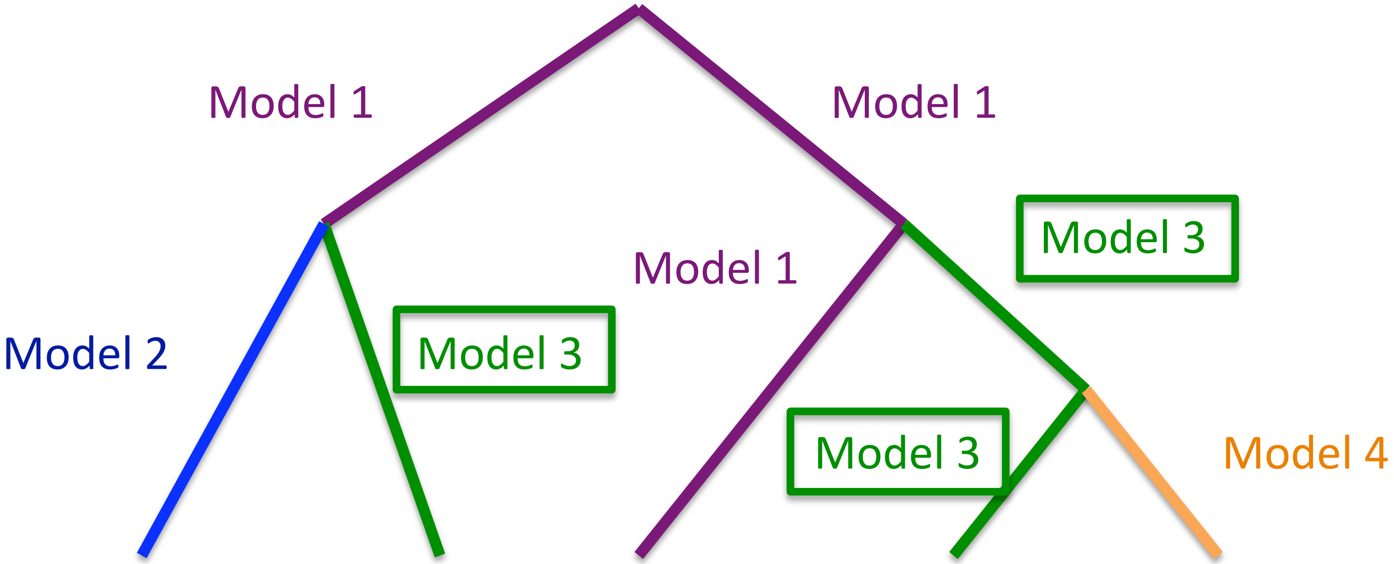
Simulating data sets with multiple models

We developed a simulation program which allows simulating data sets along a given tree with different substitution models along different branches of a tree

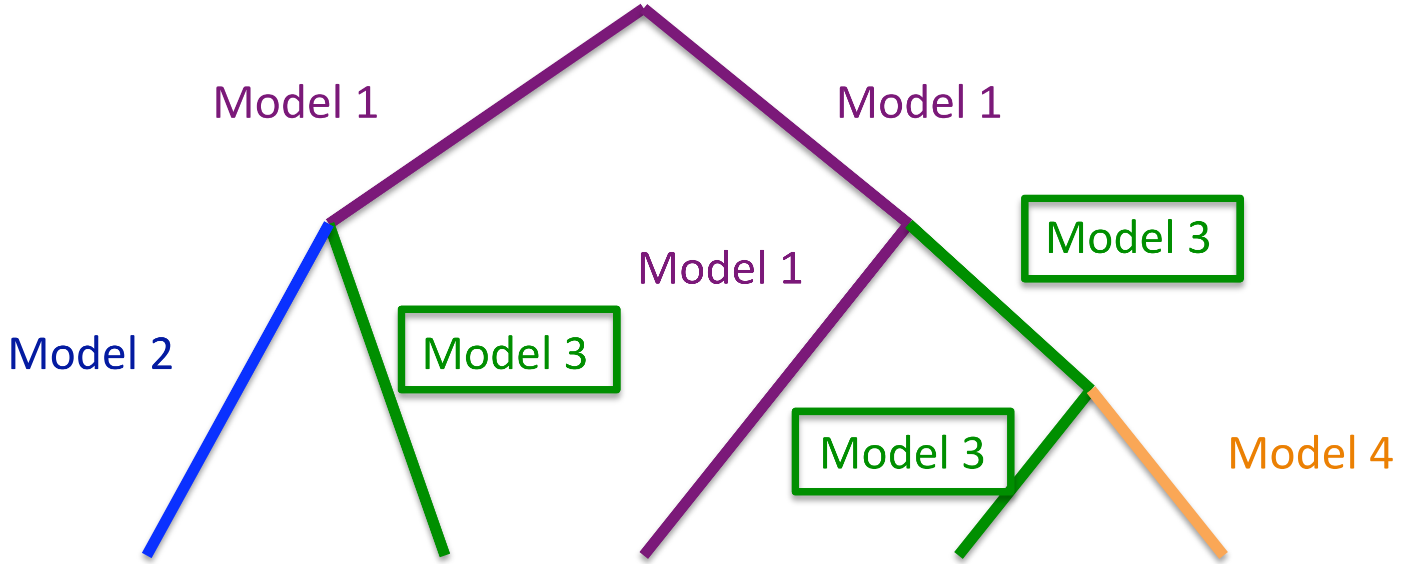


Substitution model: Basic model + Parameters + G + I

Simulating data sets with multiple models

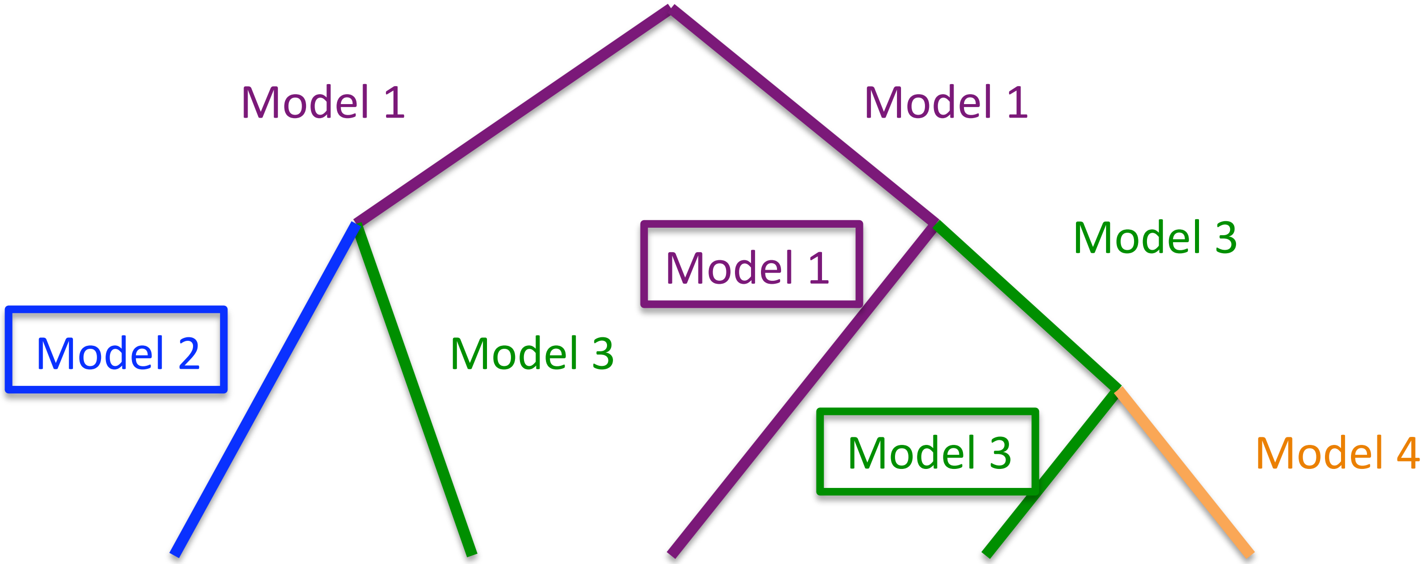


Simulating data sets with multiple models

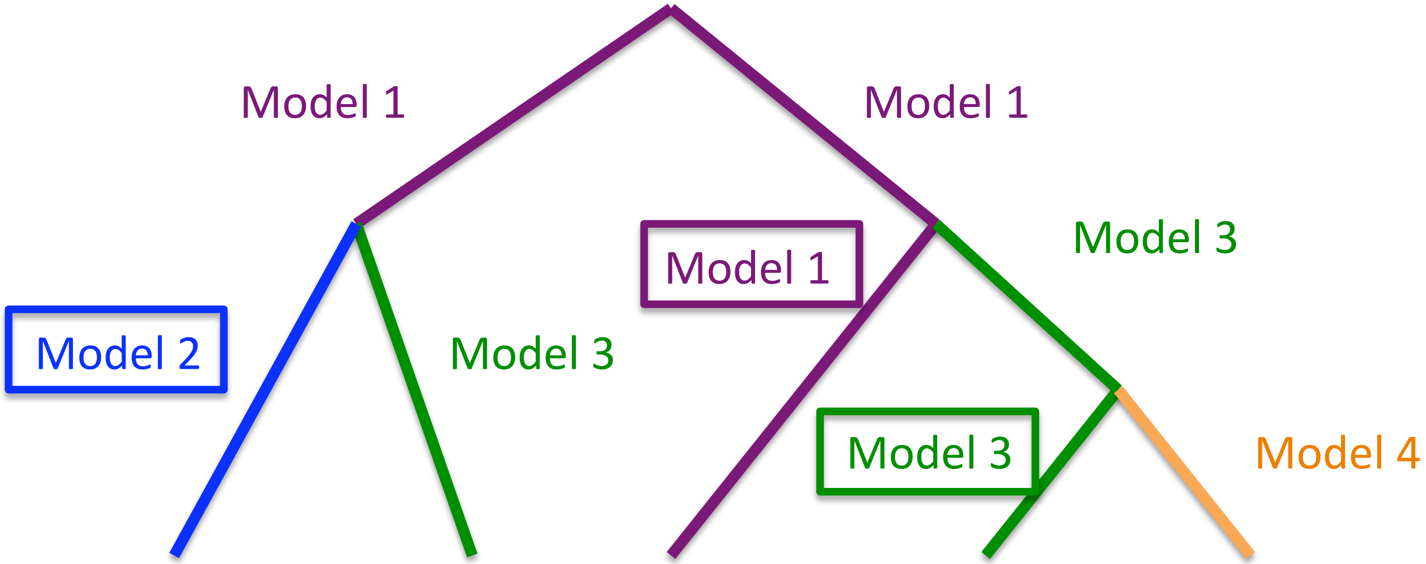


Models with same name share site-rates drawn from a gamma distribution + invariant sites

Simulating data sets with multiple models



Simulating data sets with multiple models

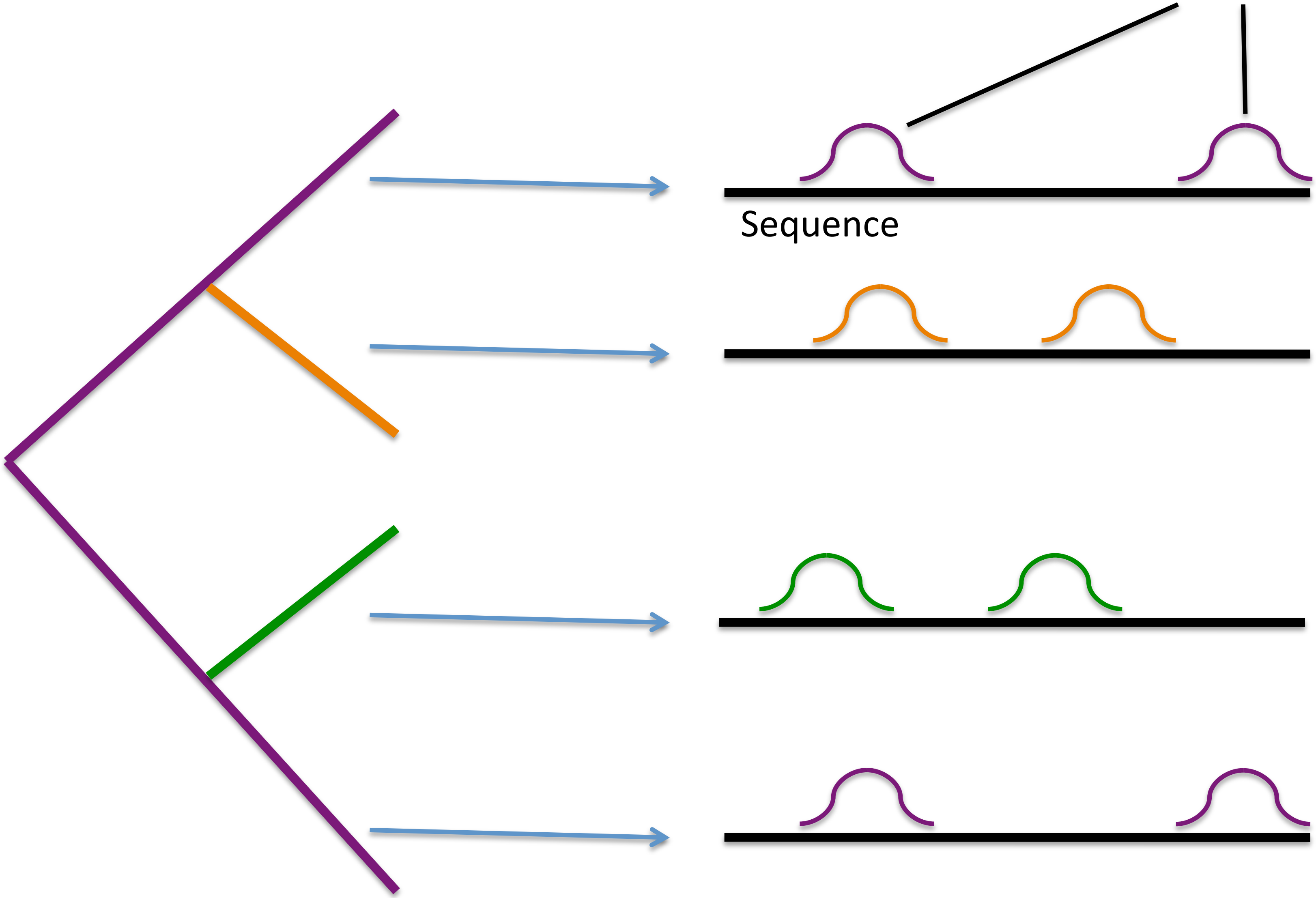


Models with different names have different site-rates drawn from a gamma distribution + different random invariant sites.

A proportion of sites can be specified that is inherited from a previously defined model.

Simulating data sets with multiple models

Effect of different site-rates along different branches: Different substitution hotspots



Our approach differs from previous approaches:

Phylogenetic mixtures: Different sites/partitions of alignment are simulated along different trees

Covarion models:

Tuffley and Steel (1998)	Site variation can be switched on or off governed by a Markov process
Galtier (2001)	Site-rates can switch among multiple evolutionary rates by a Markov process

- Proportion of sites in each rate category is constant across tree
- Rate at which sites switch is proportional to expected number of substitutions per site

Our approach differs from previous approaches:

Phylogenetic mixtures: Different sites/partitions of alignment are simulated along different trees

Covarion models:

Tuffley and Steel (1998)	Site variation can be switched on or off governed by a Markov process
Galtier (2001)	Site-rates can switch among multiple evolutionary rates by a Markov process

- Proportion of sites in each rate category is constant across tree
- Rate at which sites switch is proportional to expected number of substitutions per site

Our approach is more closely related to phylogenetic mixtures, but differs from it.

Simulation setup:

The following simulation setup has been used:

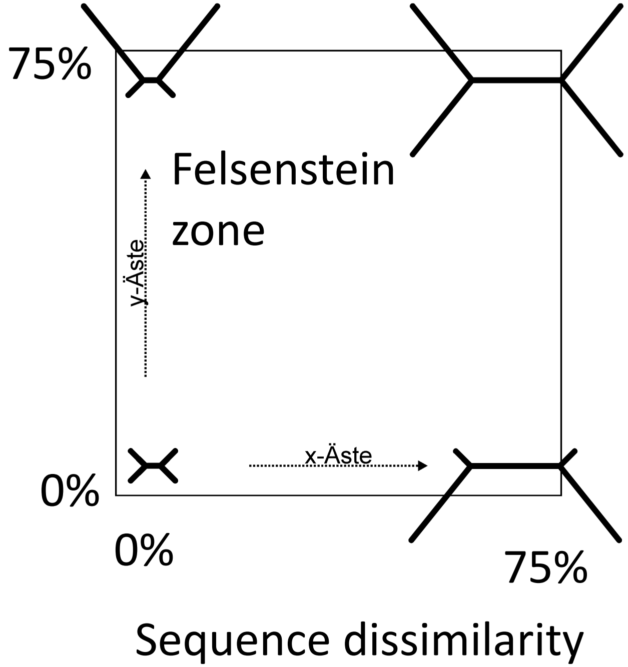
- data sets were simulated with a Markov process on 4-taxon trees
- on each branch we used a JC + G model to simulate evolution
- if not indicated otherwise, site rates where drawn randomly from a gamma distribution with alpha = 0.1
- heterotachy was simulated by using "different" models on different branches, were by differed model we mean that all site-rates were drawn independently. All equal models have the same site-rates.
- trees were reconstructed with PAUP* using ML and MP. For ML the JC+G model was specified and the parameter alpha was estimated (using 8 rate categories)

How to interpret the plots:

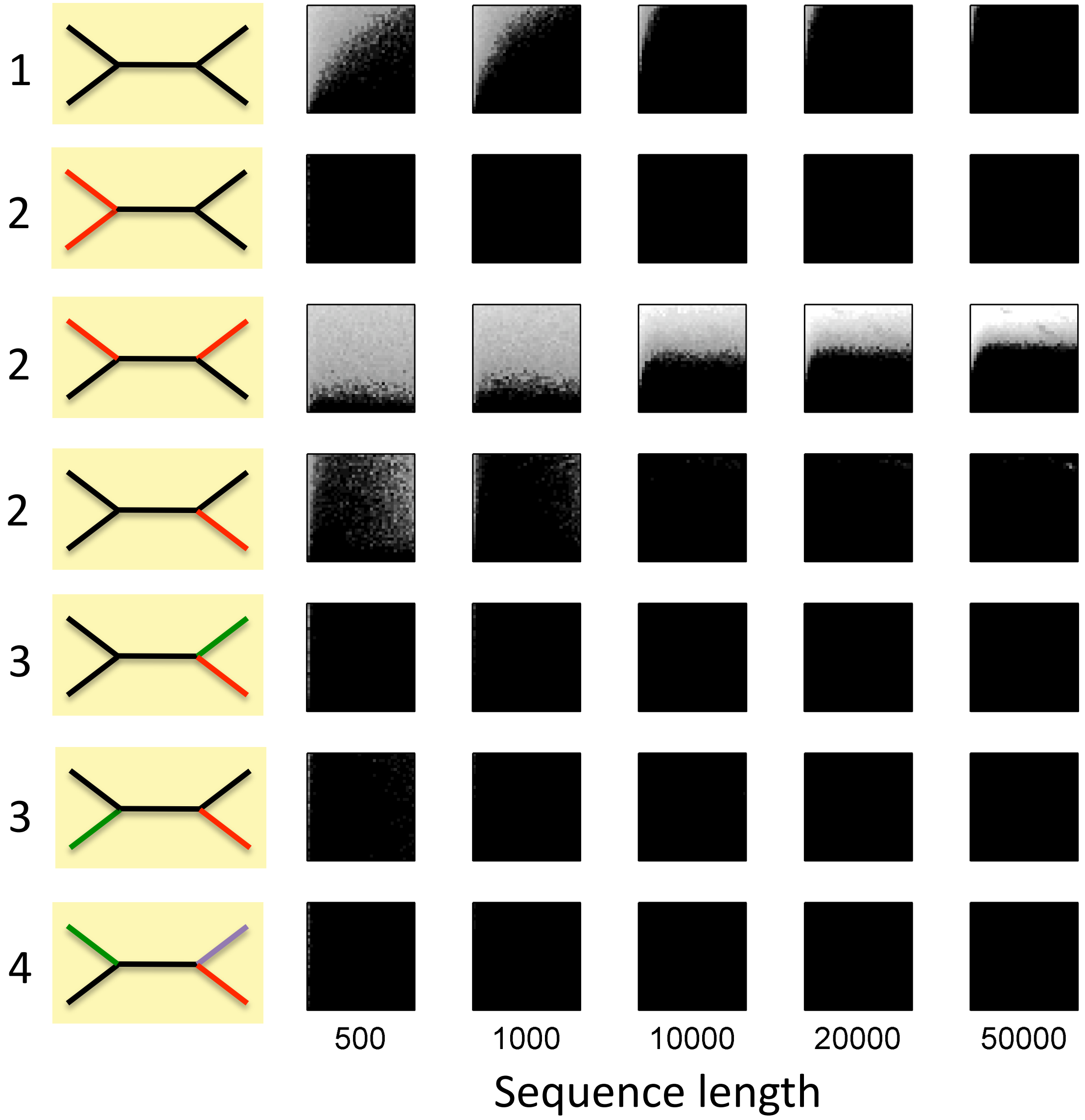
- in the plots a high reconstruction success is indicated by black, a low success by white areas.
- in the plots, branch lengths were varied from 1% to 73% sequence identity under the JC model in steps of 2% with 200 replicates at each point (analogous to Huelsenbeck 1995)

All models: JC + G, alpha = 0.1

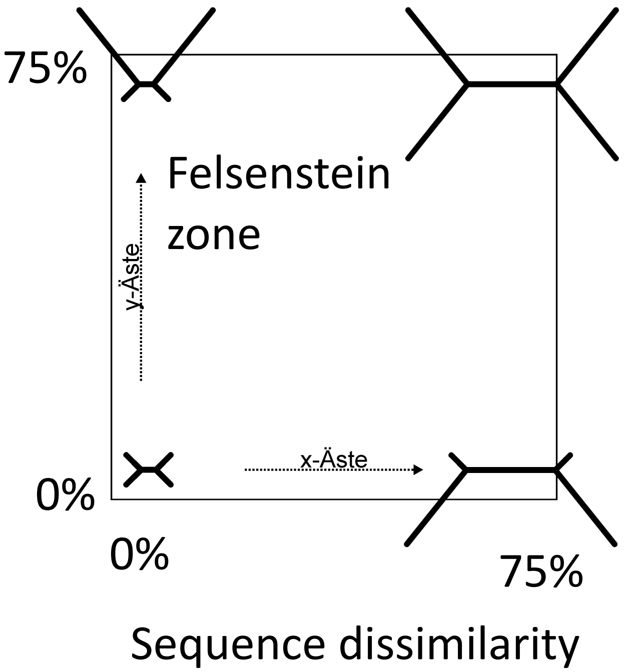
Tree shapes:



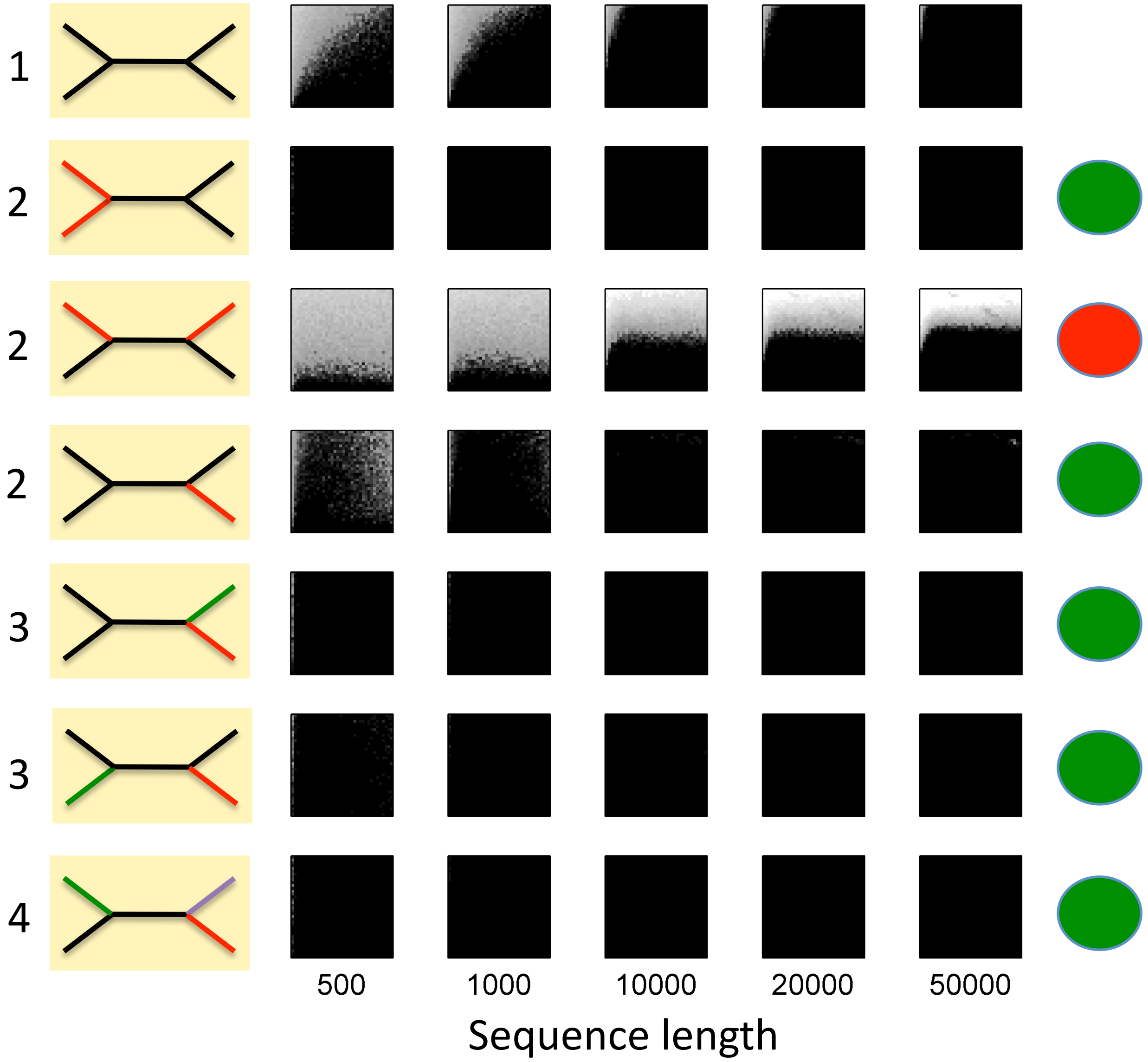
All models: JC + G, alpha = 0.1, Reconstruction: **ML**



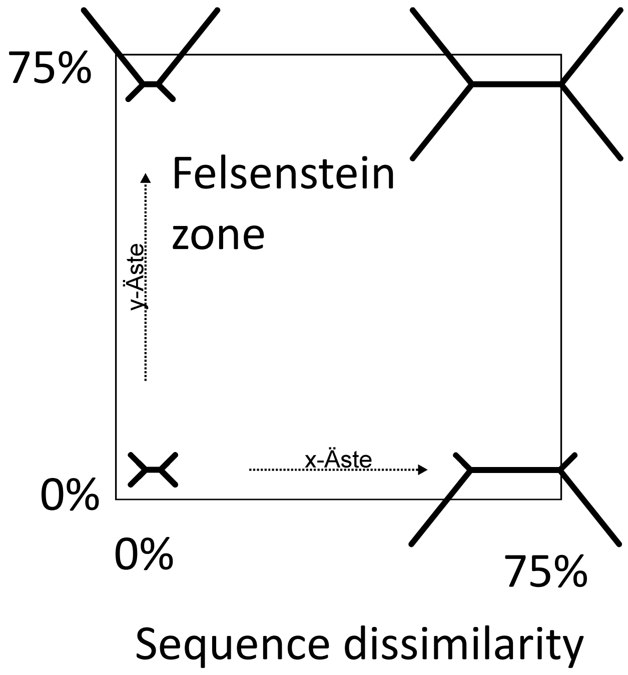
Tree shapes:



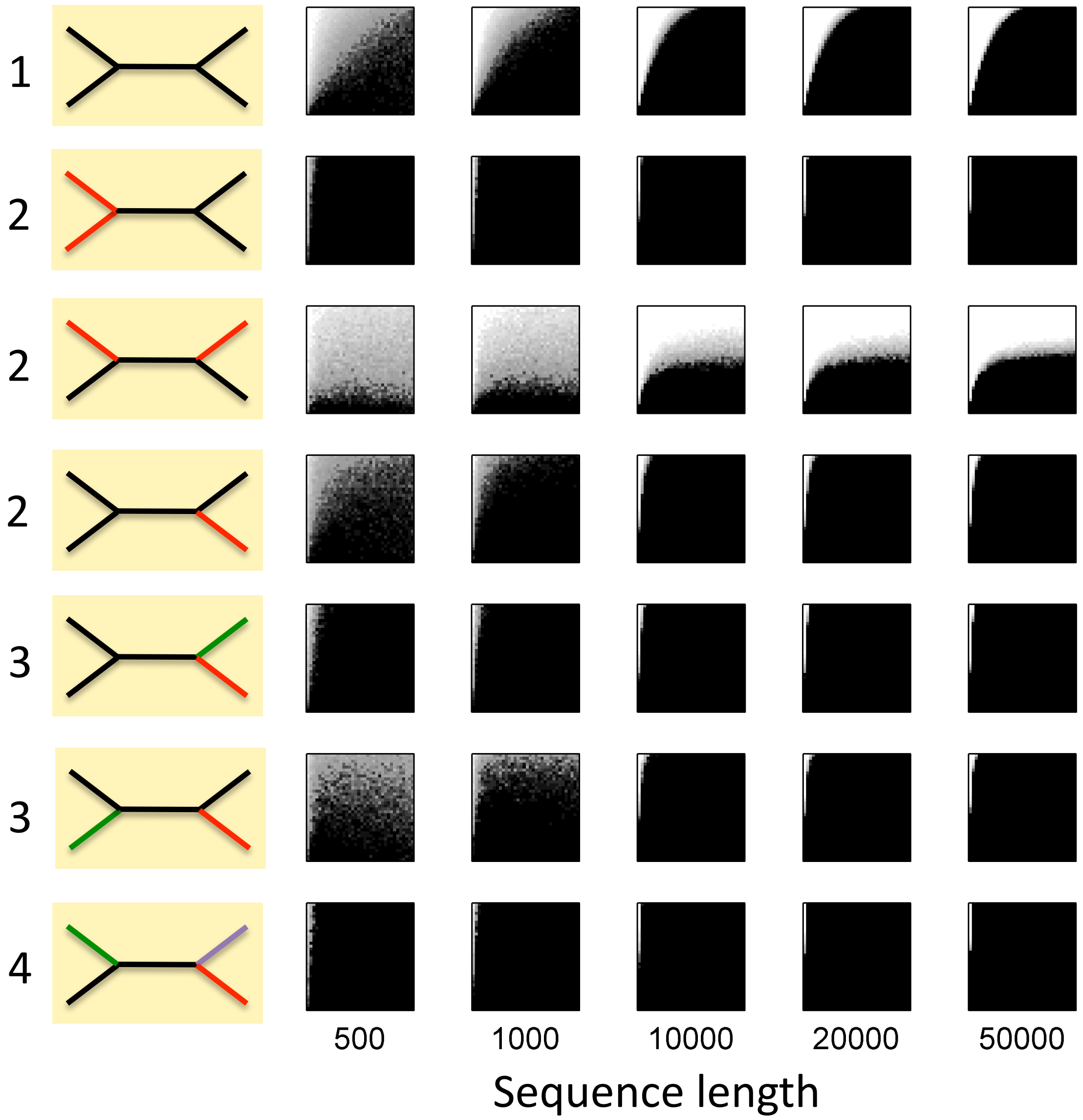
All models: JC + G, alpha = 0.1, Reconstruction: **ML**



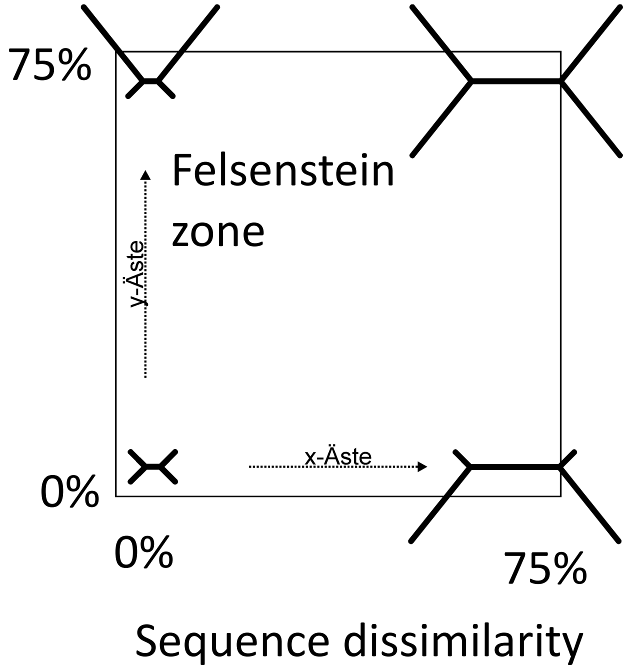
Tree shapes:



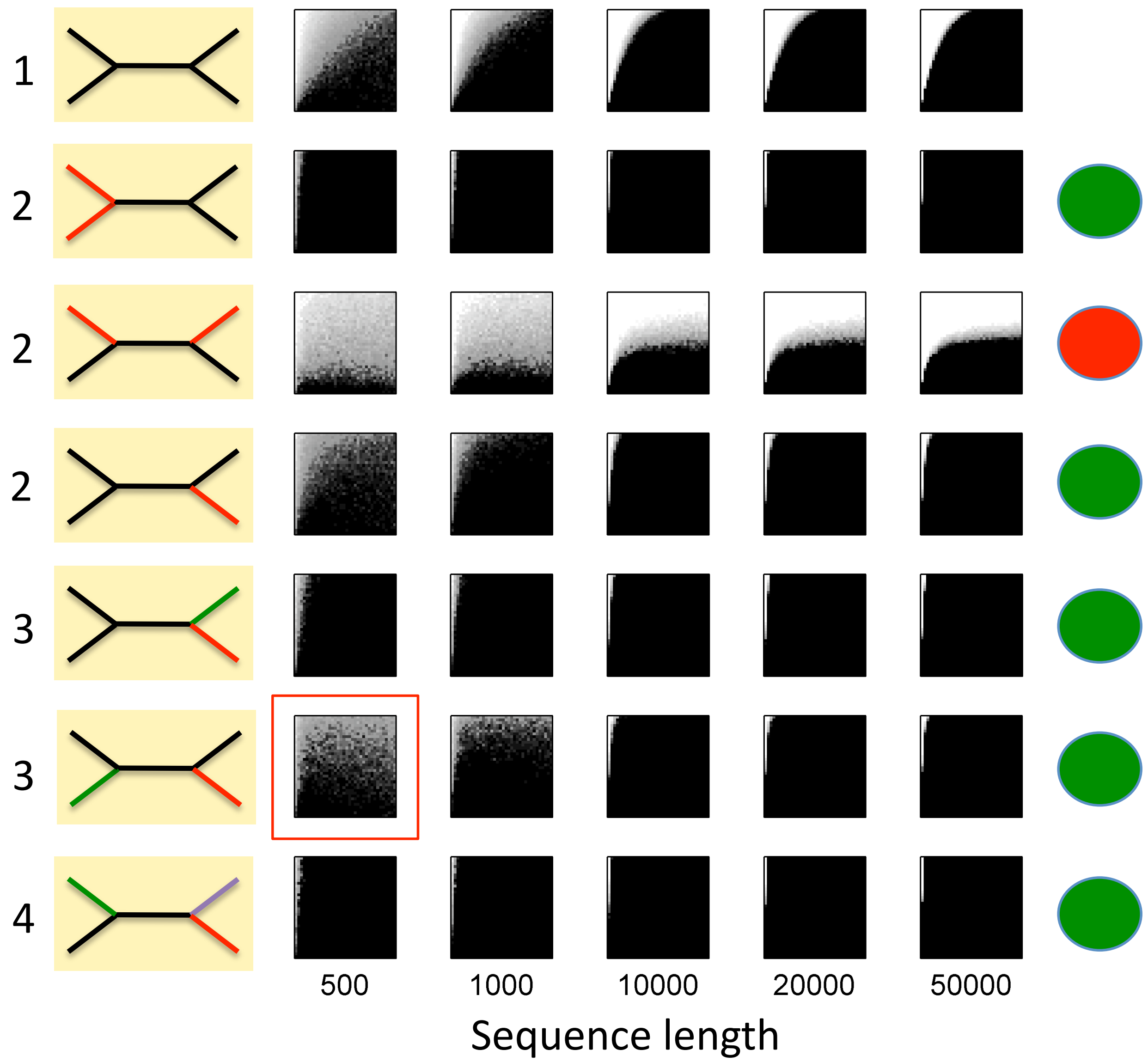
All models: JC + G, alpha = 0.1, Reconstruction: **MP**



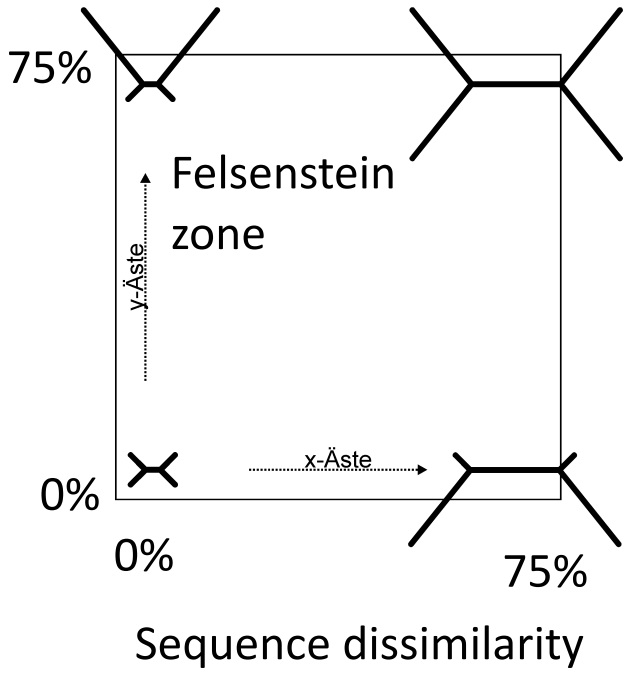
Tree shapes:



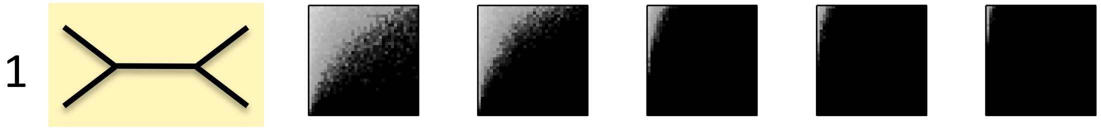
All models: JC + G, alpha = 0.1, Reconstruction: **MP**



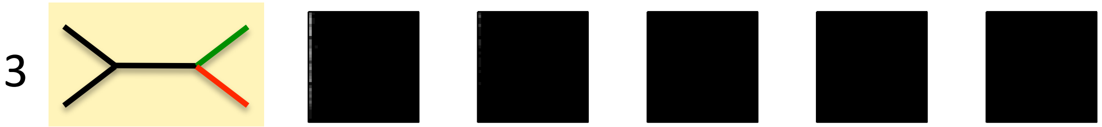
Tree shapes:



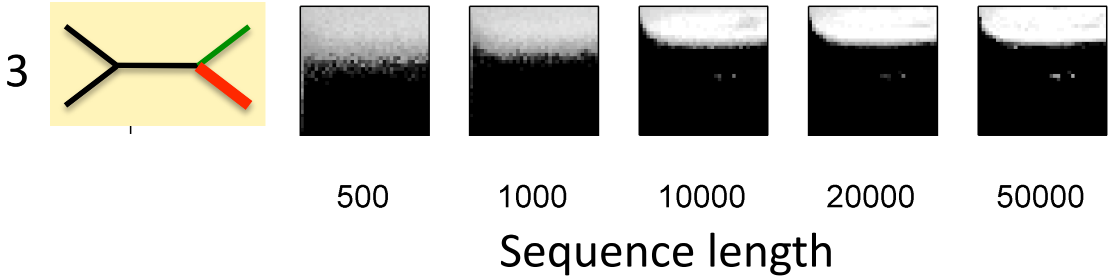
All models: JC + G, alpha = 0.1, Reconstruction: **ML**



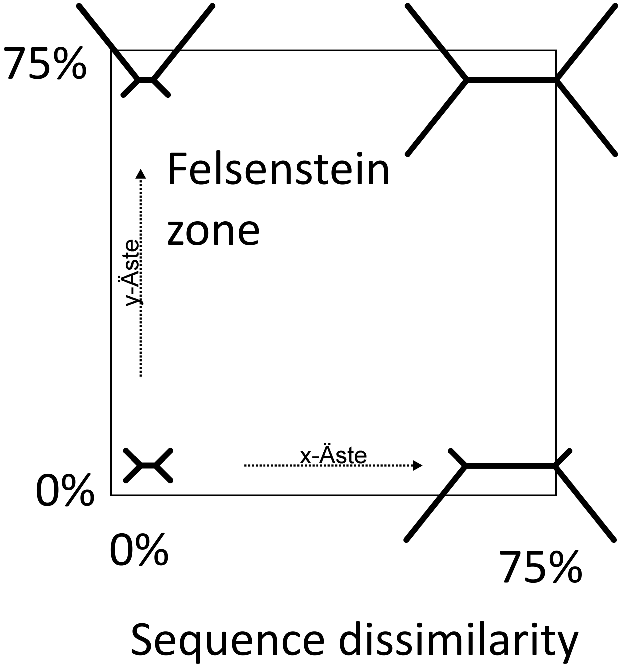
Third model has alpha = 0.1



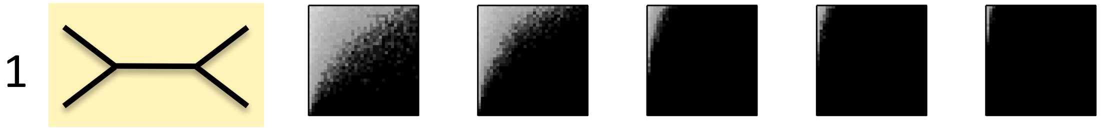
Third model has equal rates



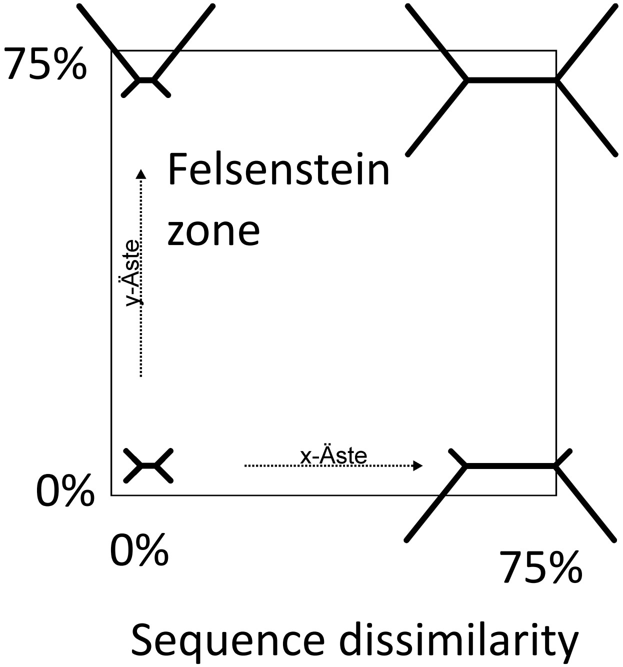
Tree shapes:



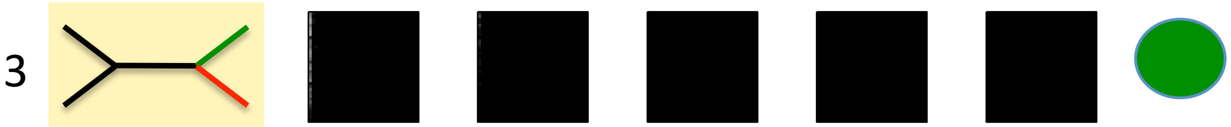
All models: JC + G, alpha = 0.1, Reconstruction: **ML**



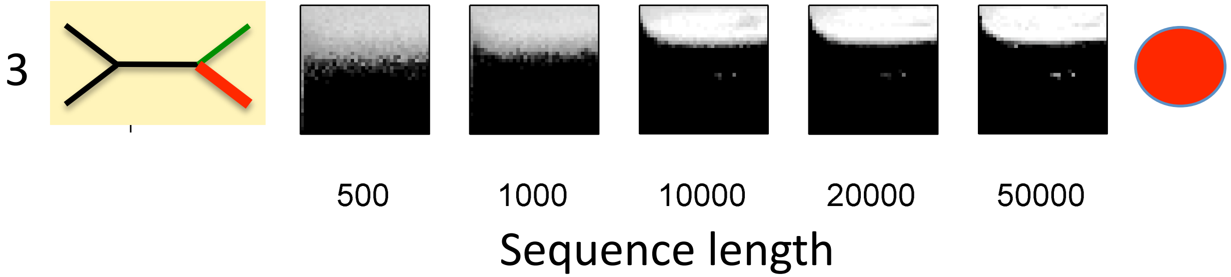
Tree shapes:



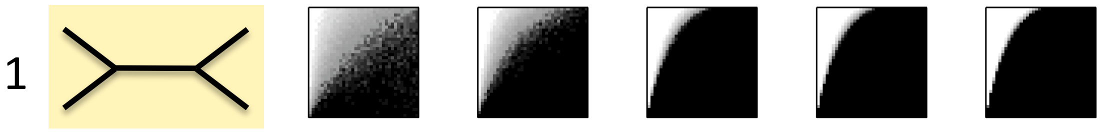
Third model has alpha = 0.1



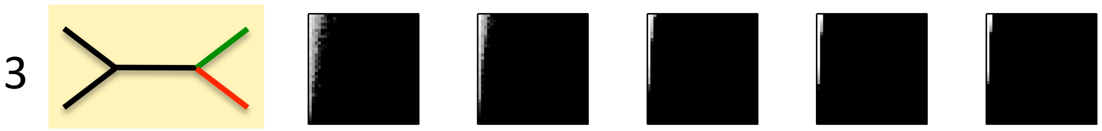
Third model has equal rates



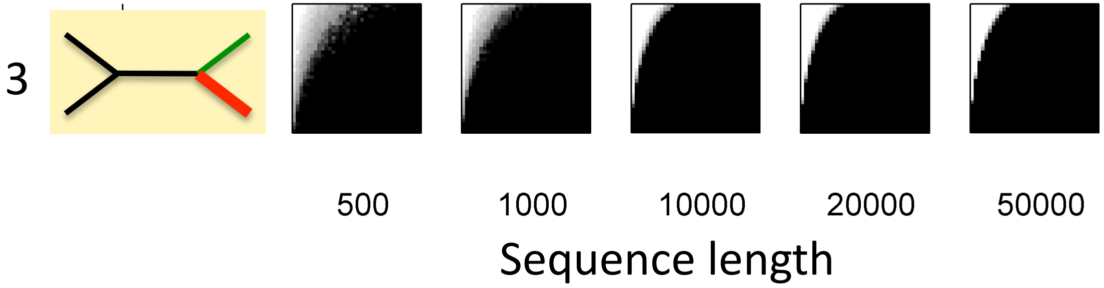
All models: JC + G, alpha = 0.1, Reconstruction: **MP**



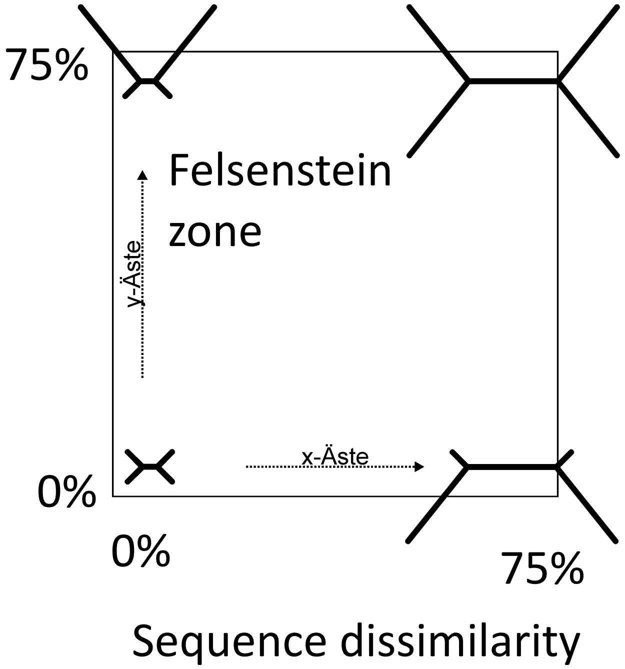
Third model has alpha = 0.1



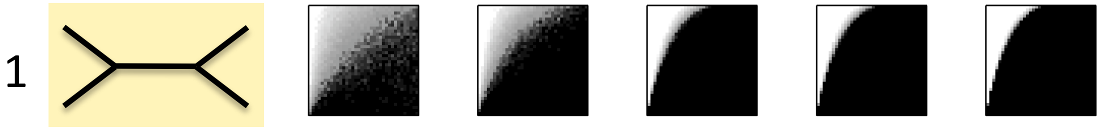
Third model has equal rates



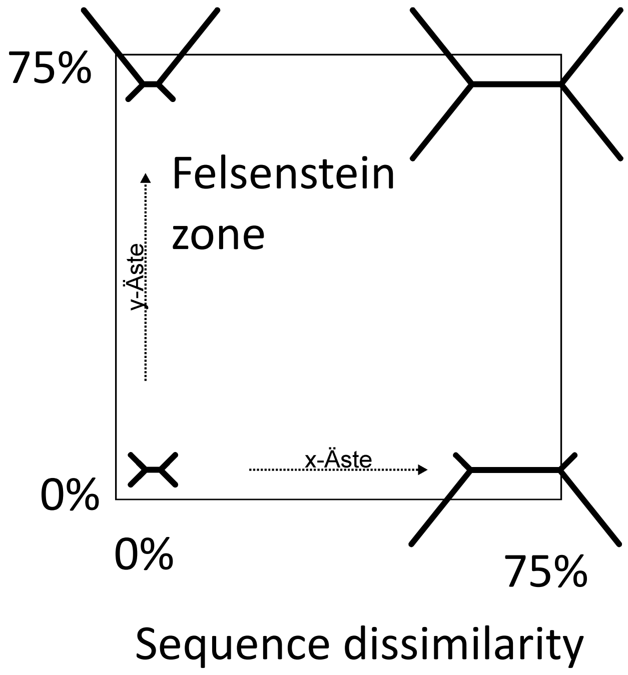
Tree shapes:



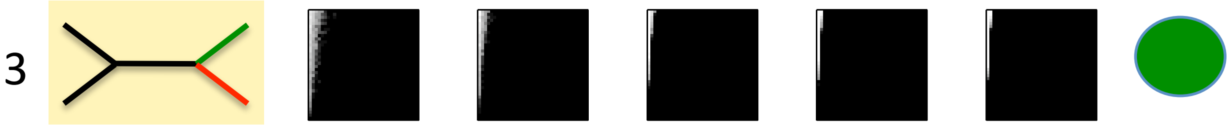
All models: JC + G, alpha = 0.1, Reconstruction: **MP**



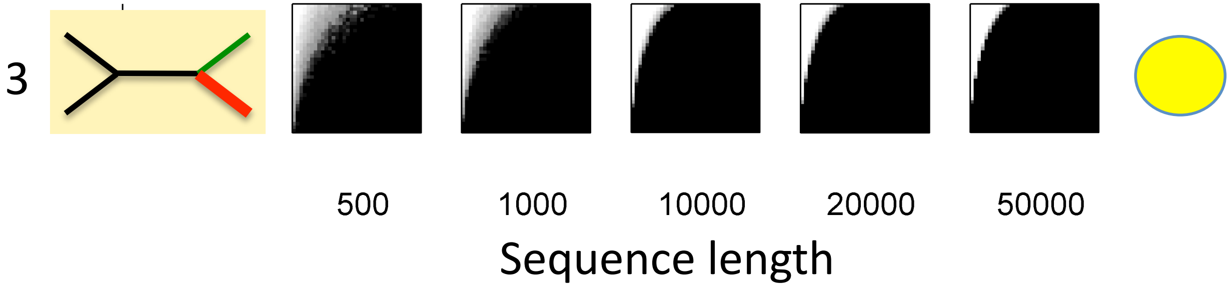
Tree shapes:



Third model has alpha = 0.1

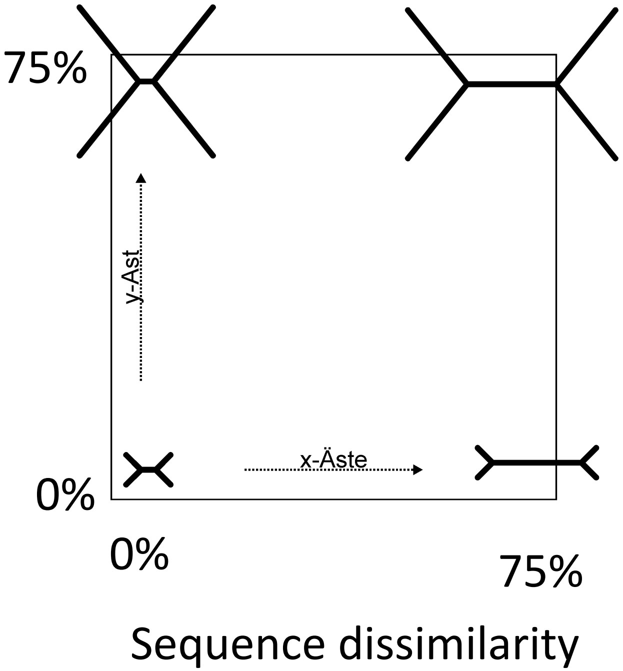


Third model has equal rates

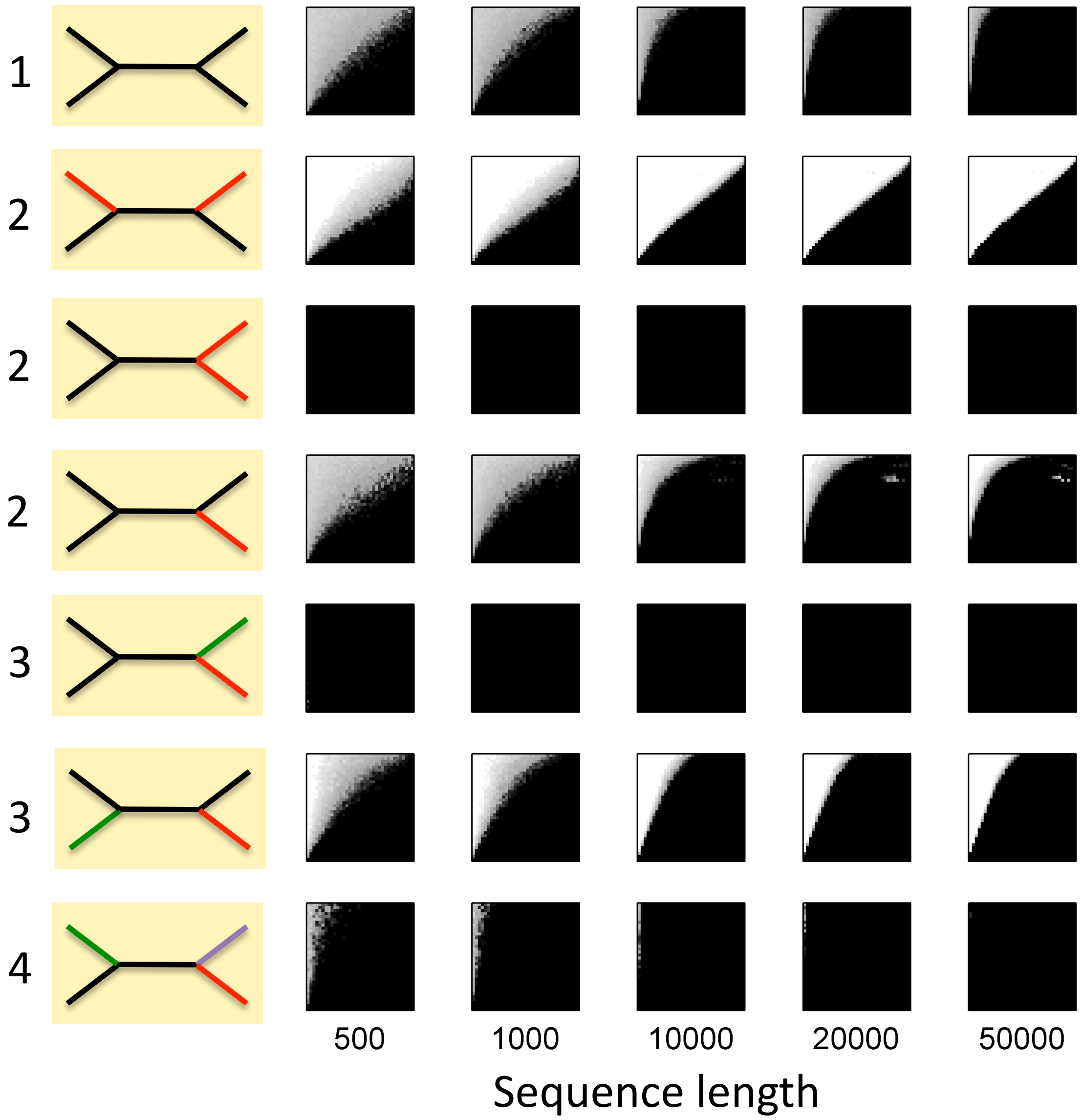


All models: JC + G, alpha = 0.1

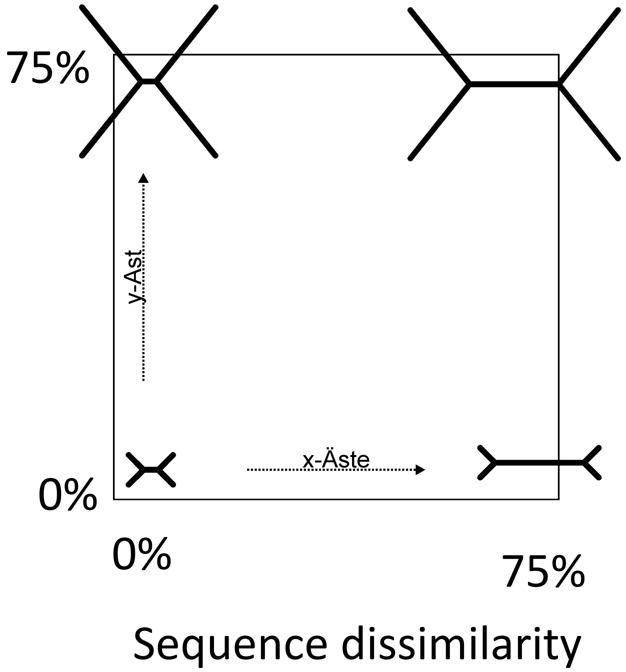
Tree shapes:



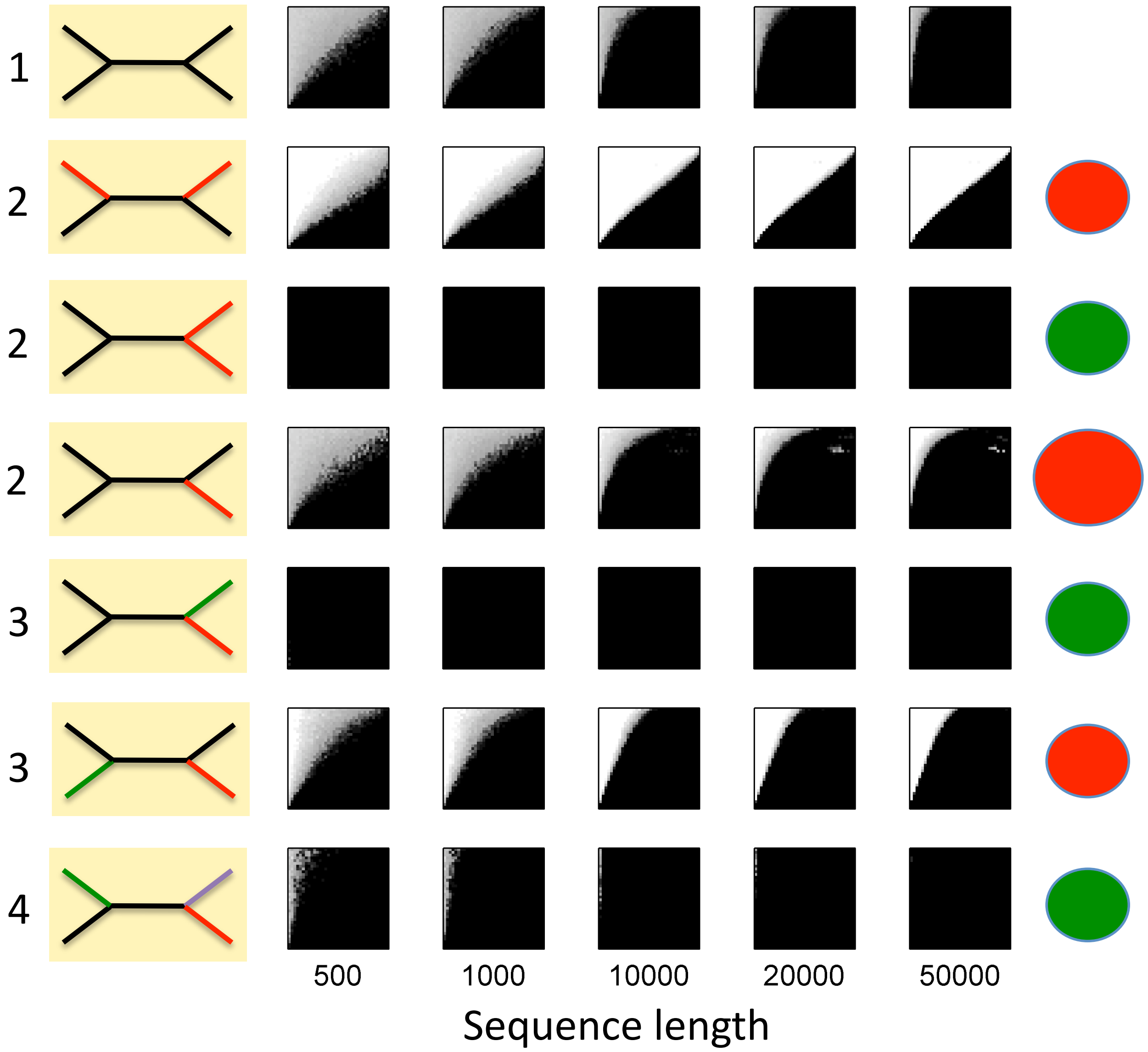
All models: JC + G, alpha = 0.1, Reconstruction: **ML**



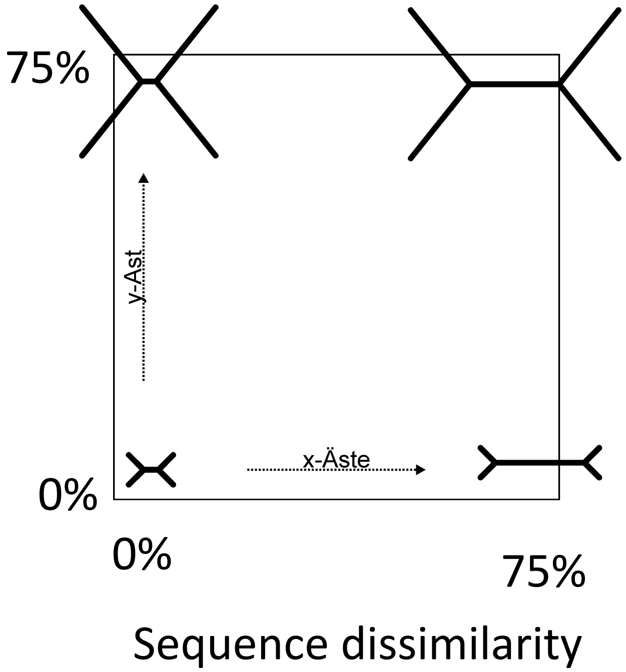
Tree shapes:



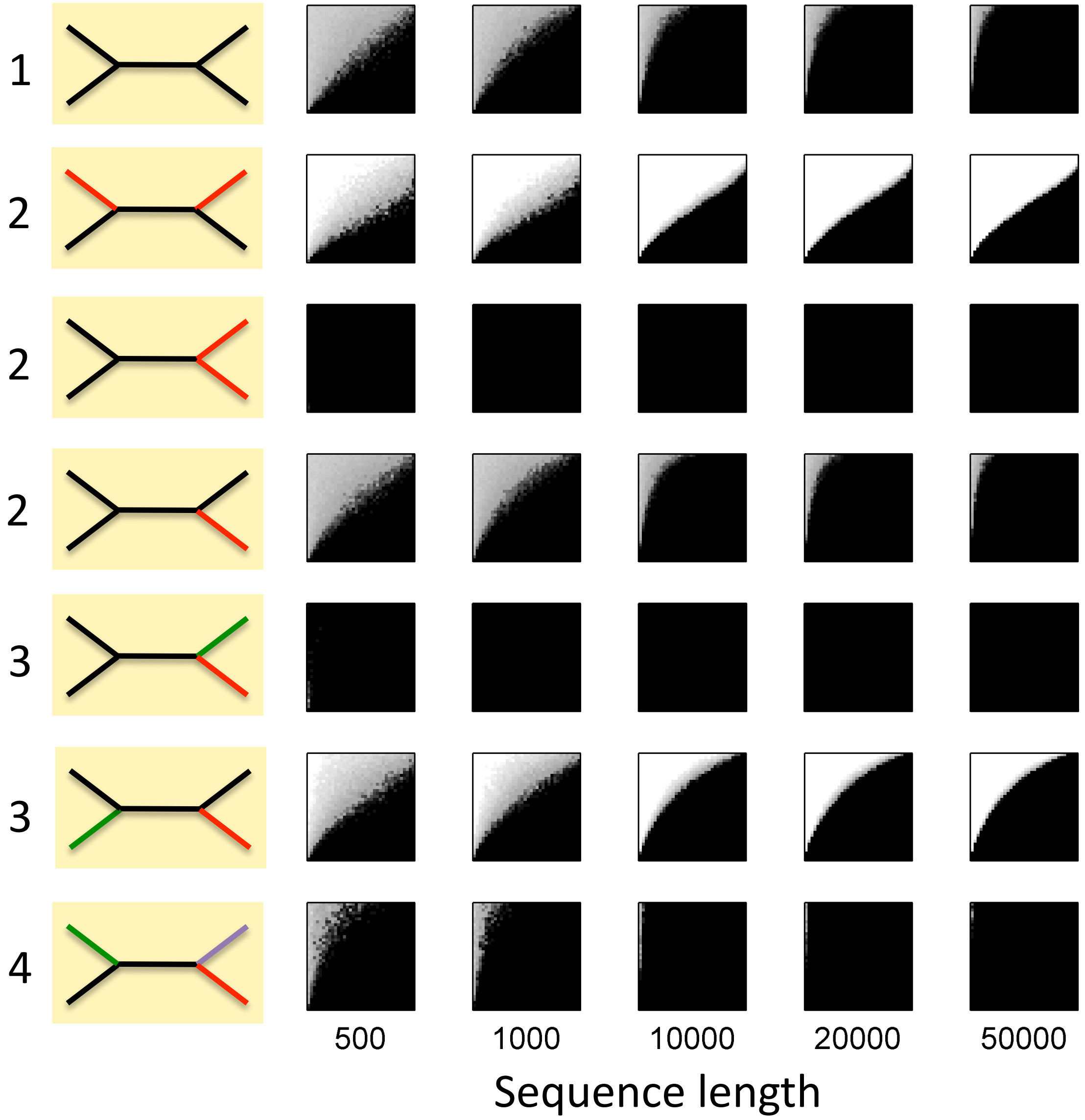
All models: JC + G, alpha = 0.1, Reconstruction: **ML**



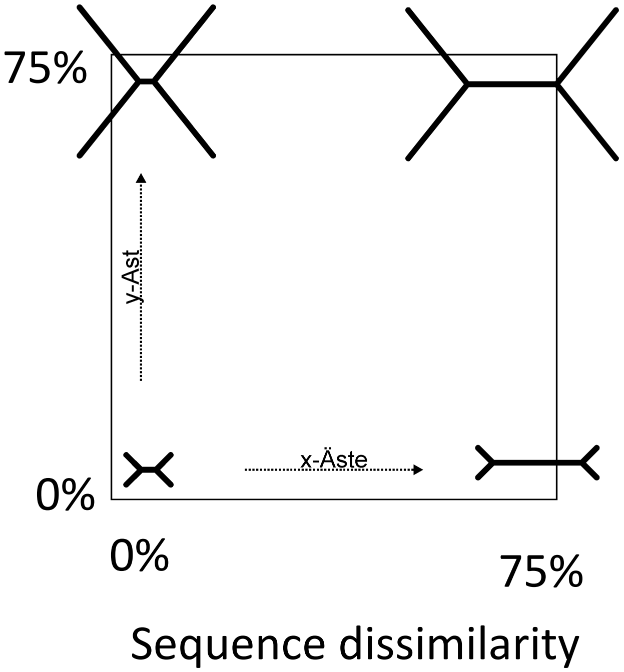
Tree shapes:



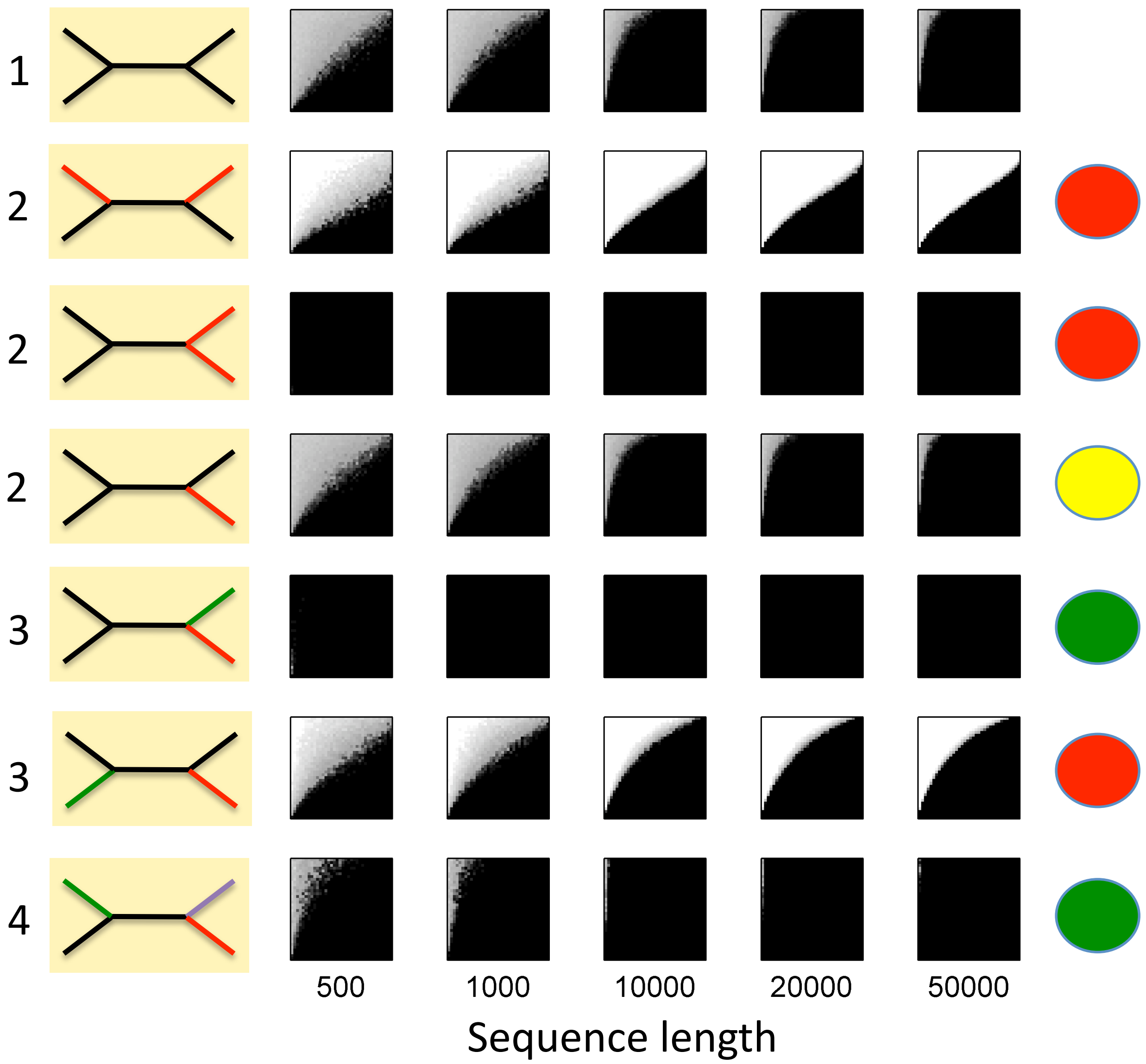
All models: JC + G, alpha = 0.1, Reconstruction: **MP**



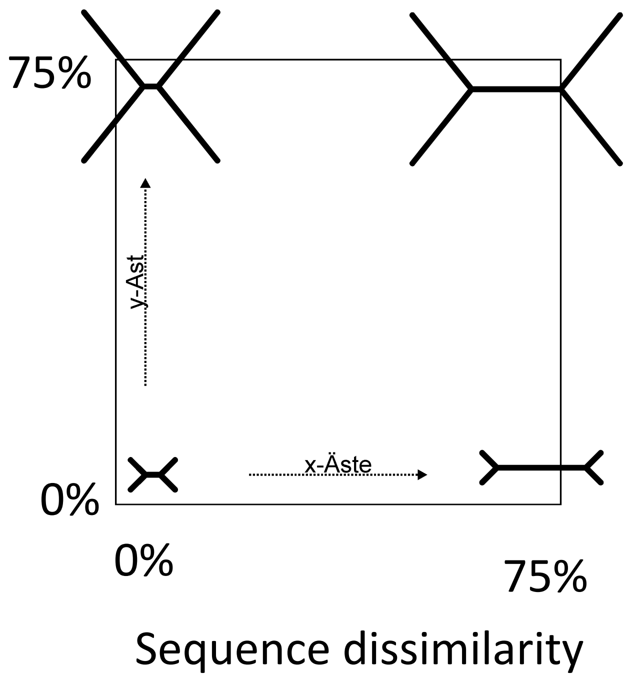
Tree shapes:



All models: JC + G, alpha = 0.1, Reconstruction: **MP**

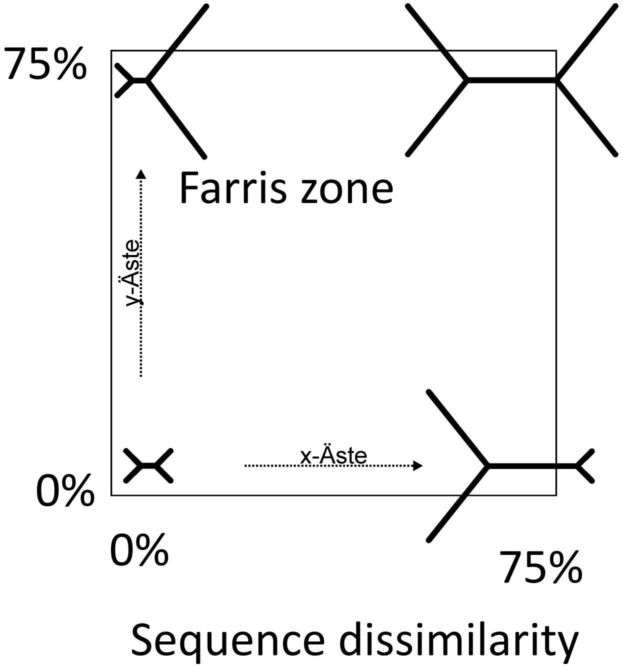


Tree shapes:

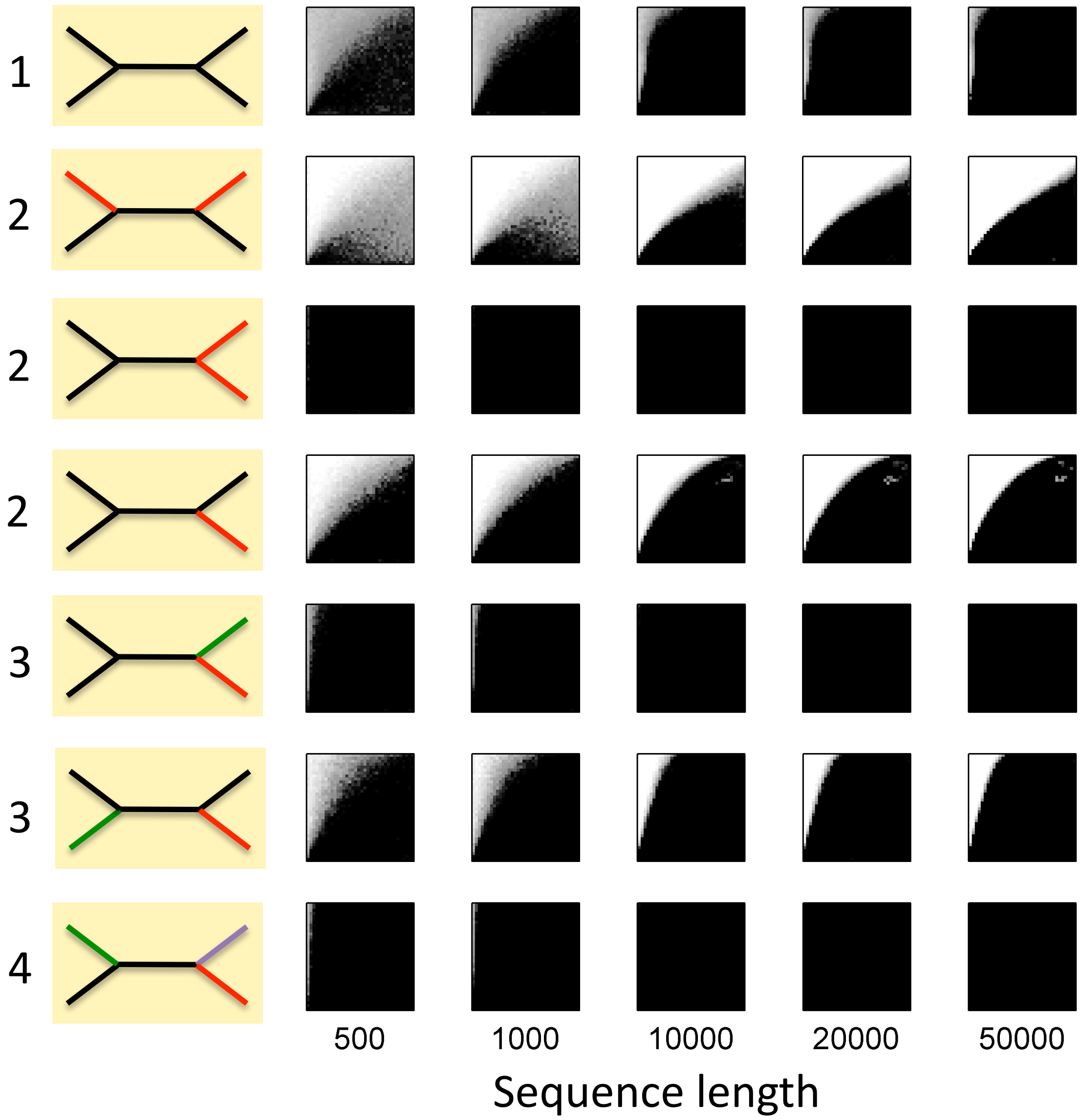


All models: JC + G, alpha = 0.1

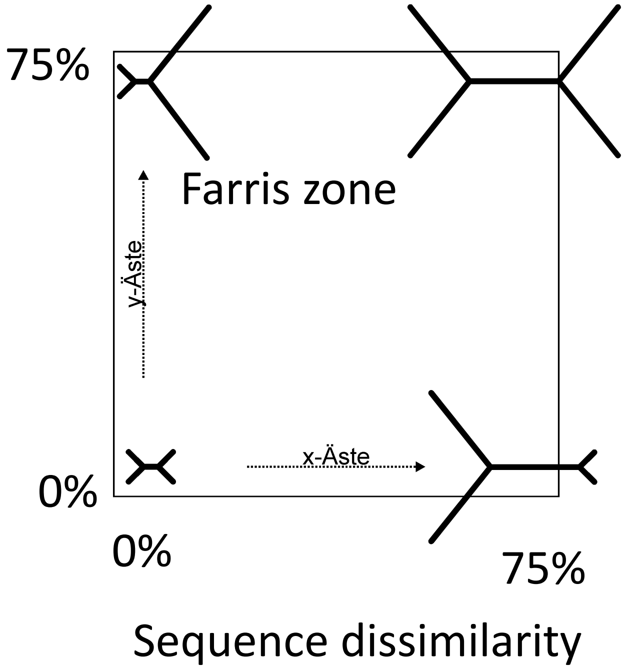
Tree shapes:



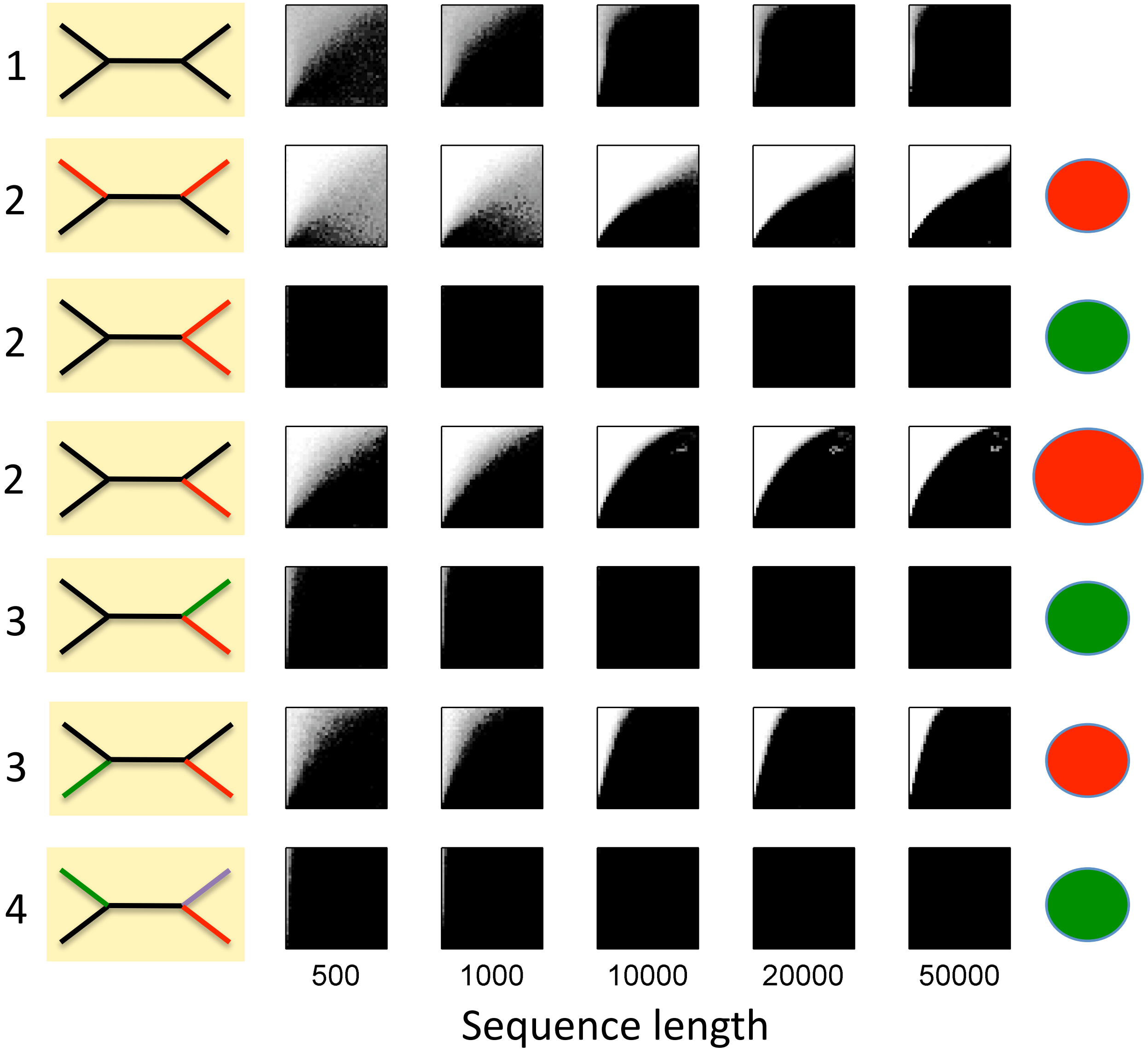
All models: JC + G, alpha = 0.1, Reconstruction: **ML**



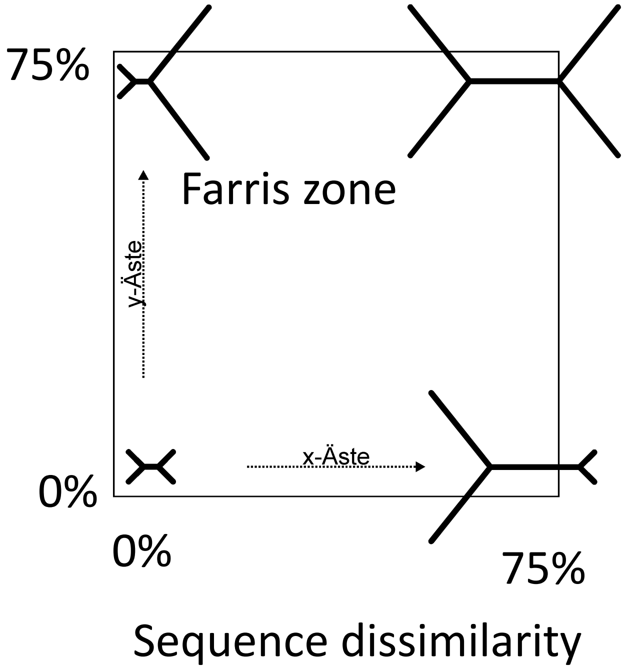
Tree shapes:



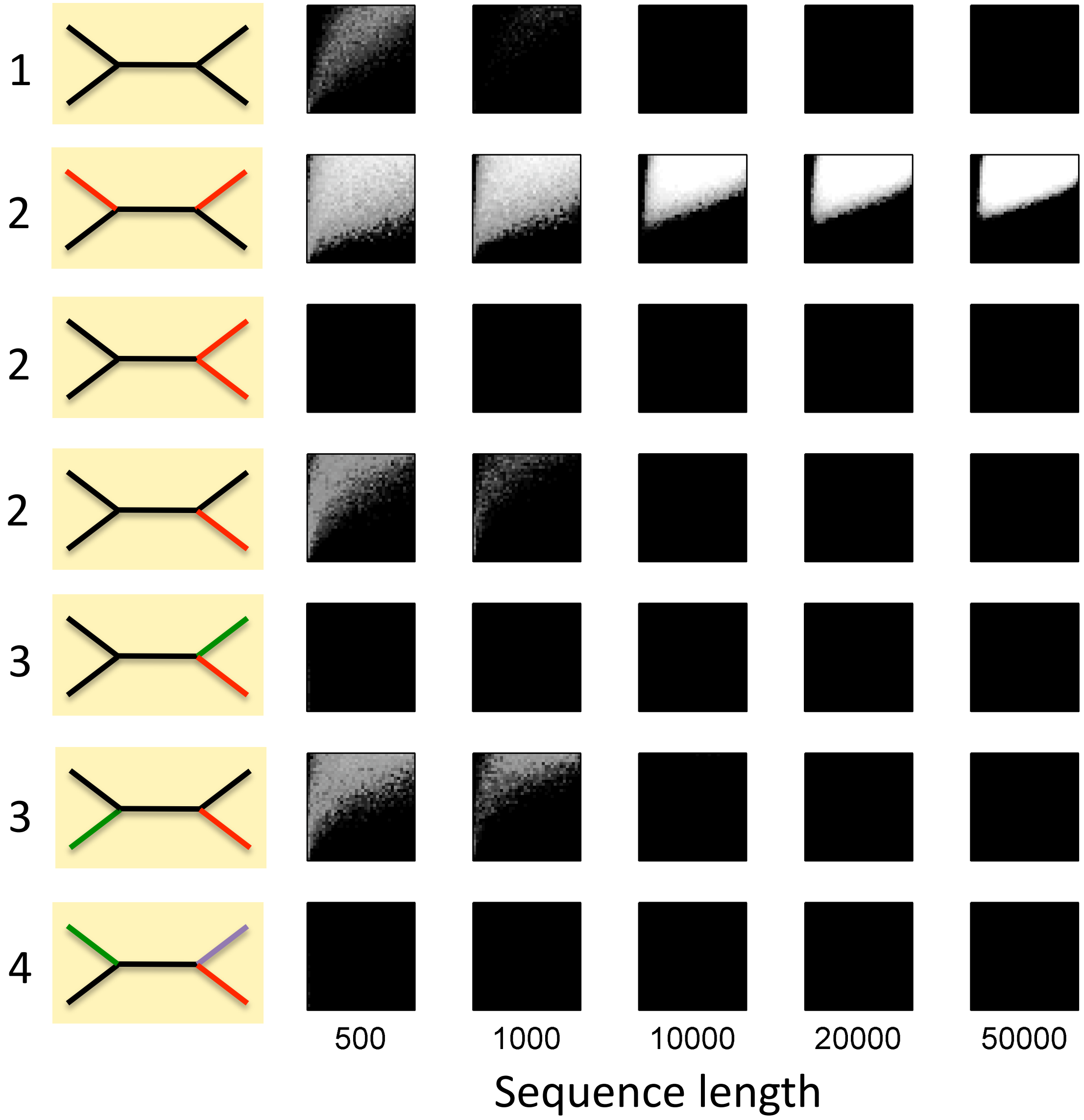
All models: JC + G, alpha = 0.1, Reconstruction: **ML**



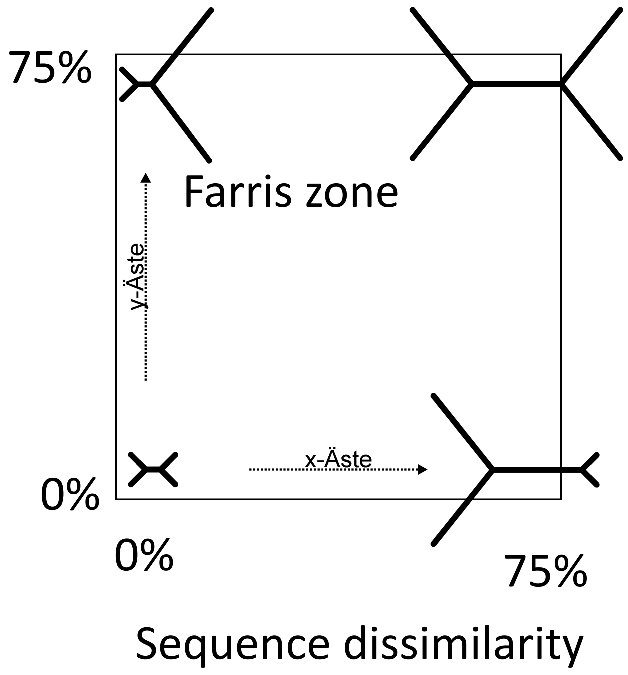
Tree shapes:



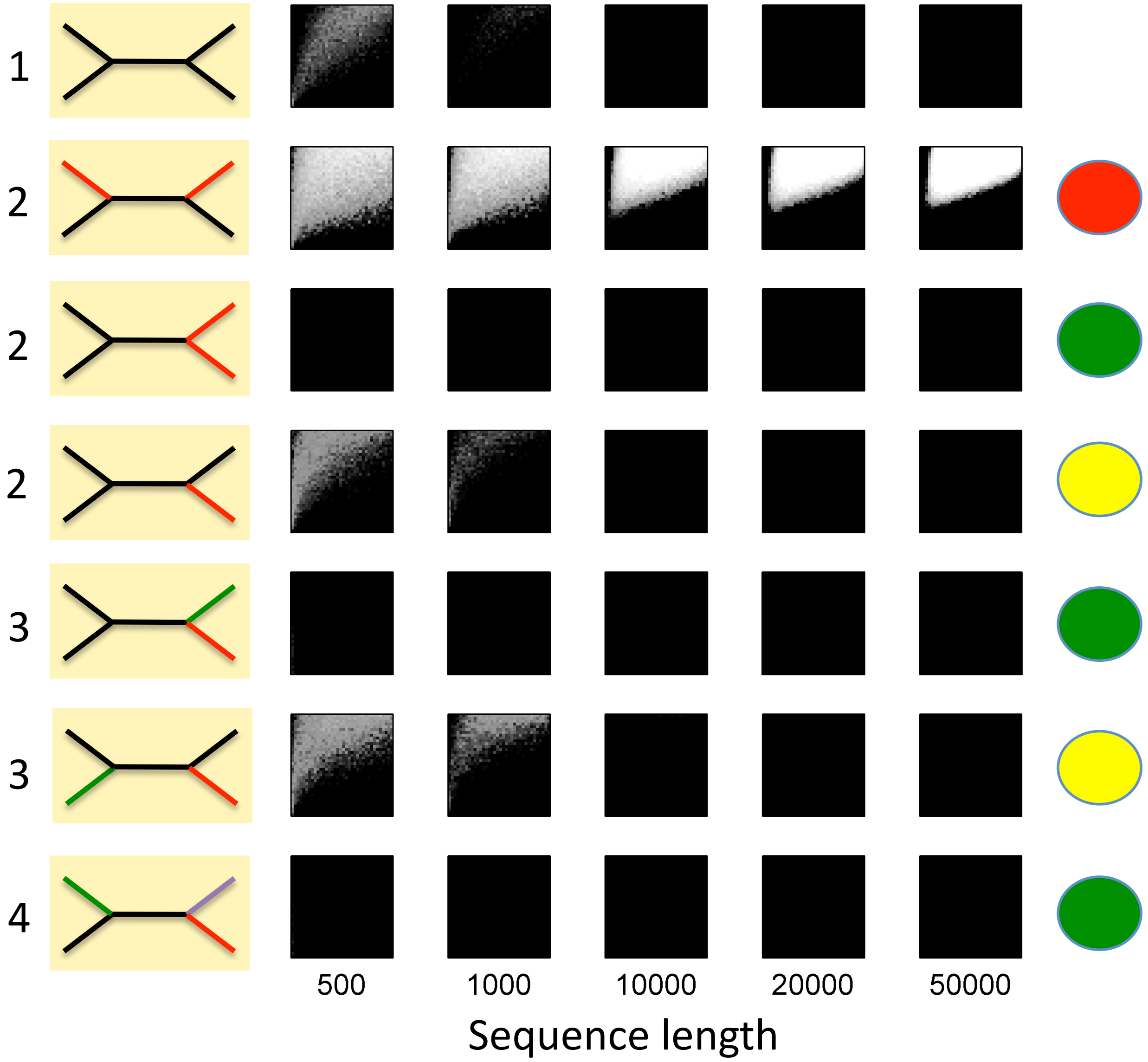
All models: JC + G, alpha = 0.1, Reconstruction: **MP**



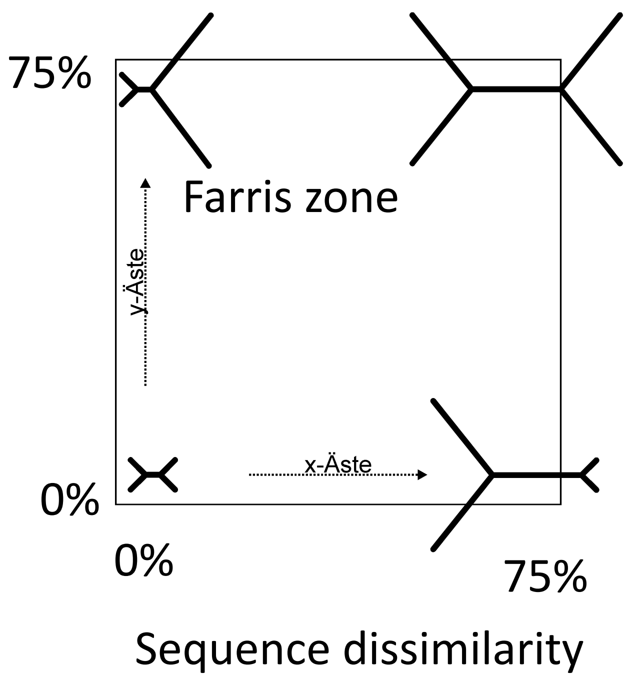
Tree shapes:



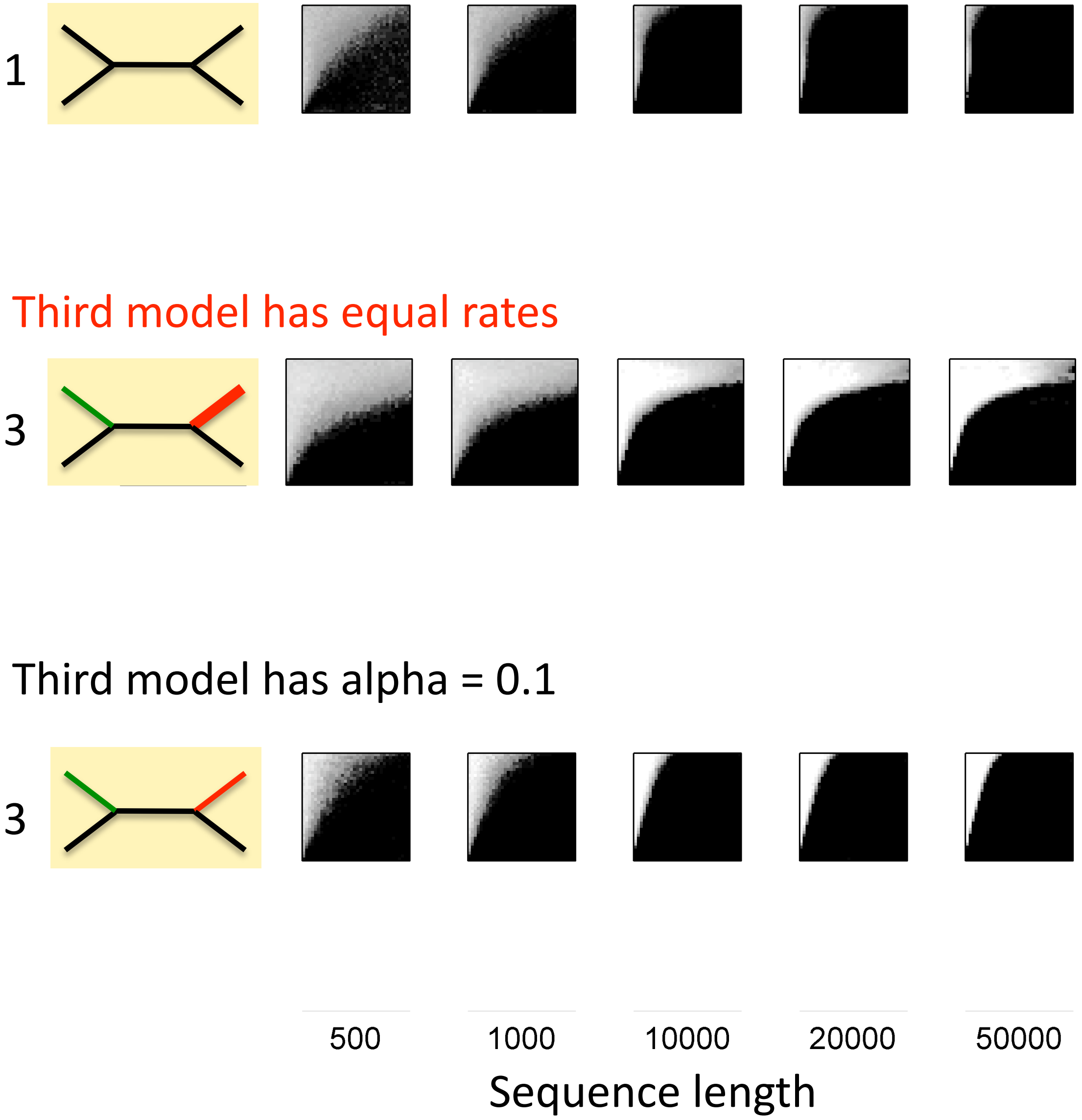
All models: JC + G, alpha = 0.1, Reconstruction: **MP**



Tree shapes:



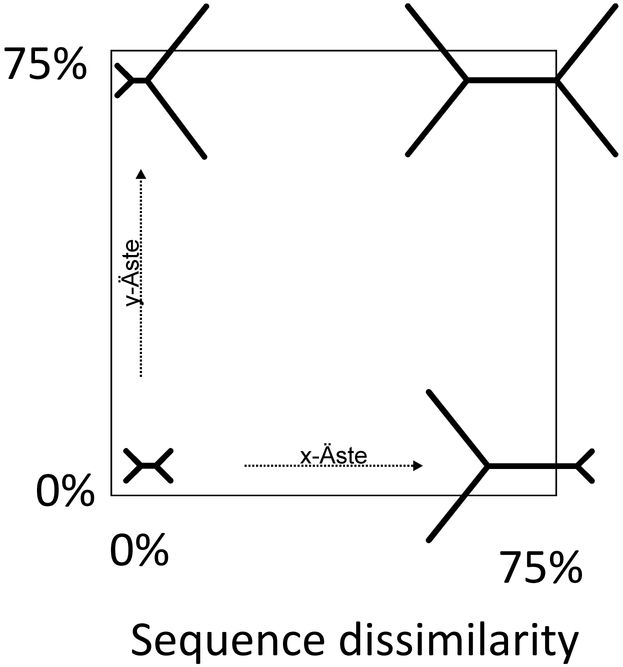
All models: JC + G, alpha = 0.1, Reconstruction: **ML**



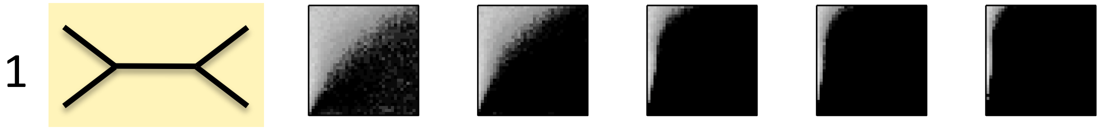
Third model has equal rates

Third model has alpha = 0.1

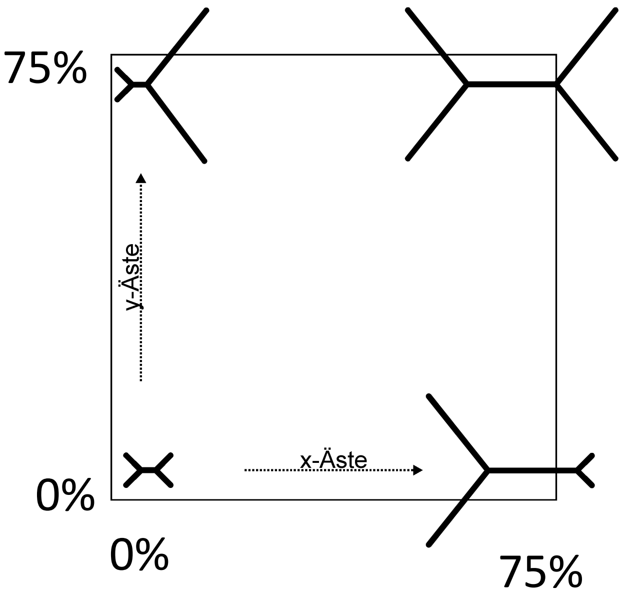
Tree shapes:



All models: JC + G, alpha = 0.1, Reconstruction: **ML**

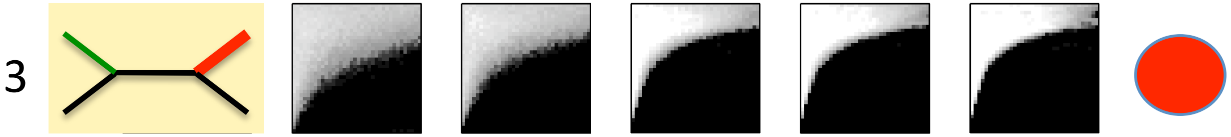


Tree shapes:

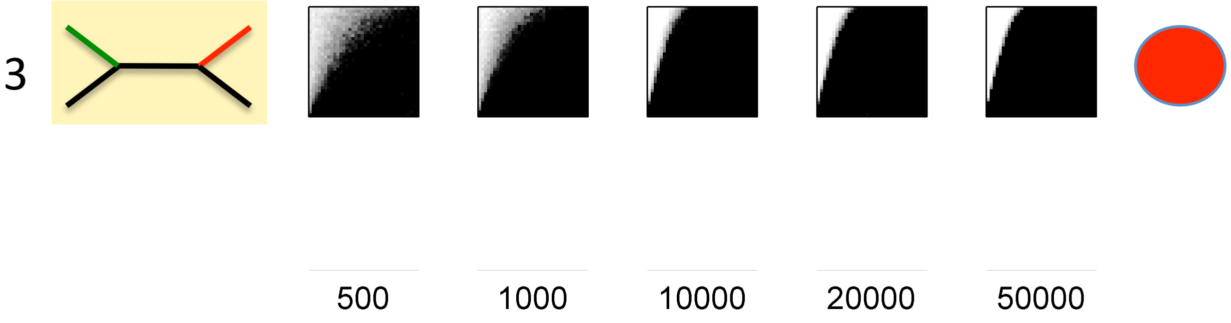


Sequence dissimilarity

Third model has equal rates

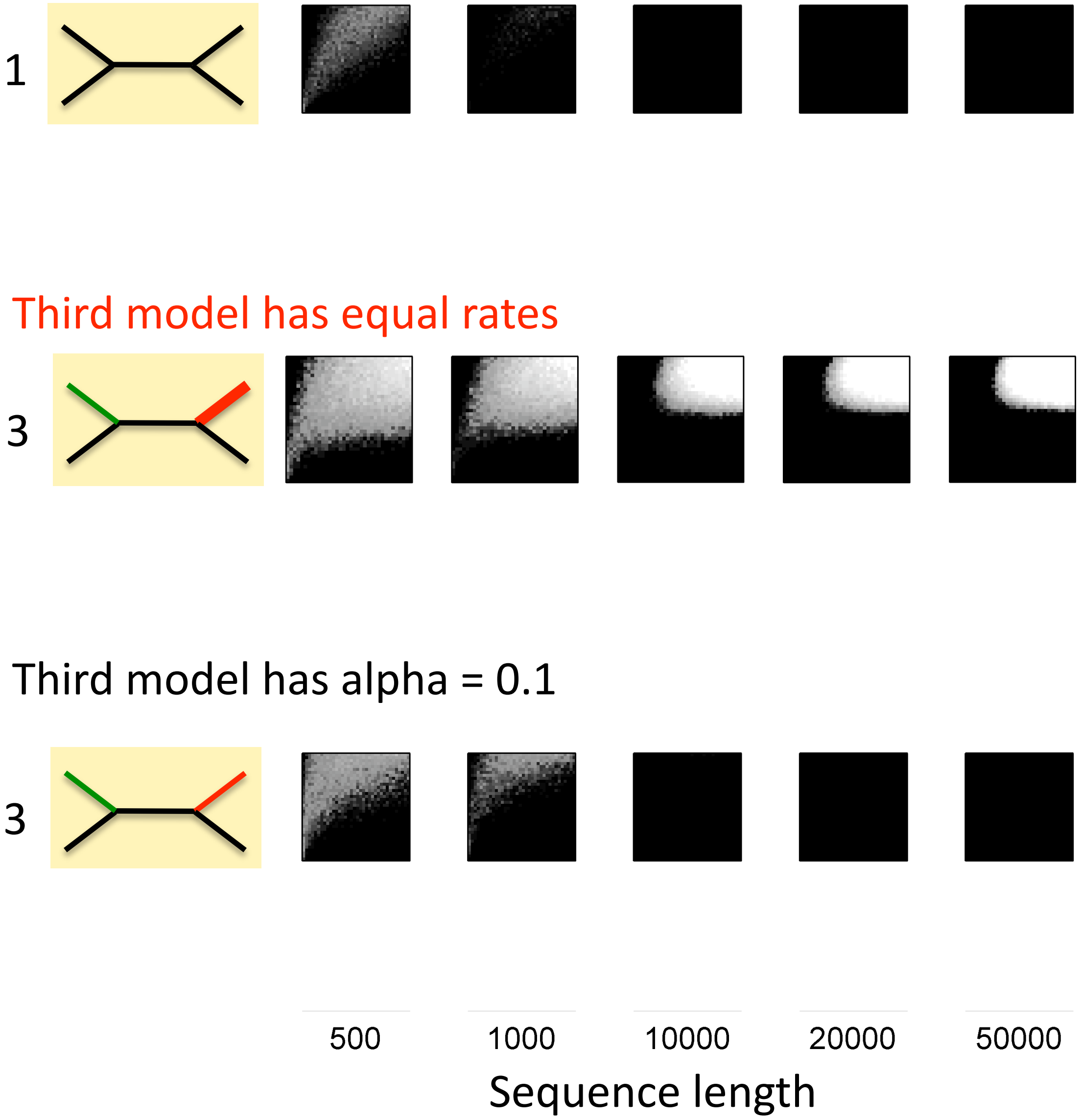


Third model has alpha = 0.1

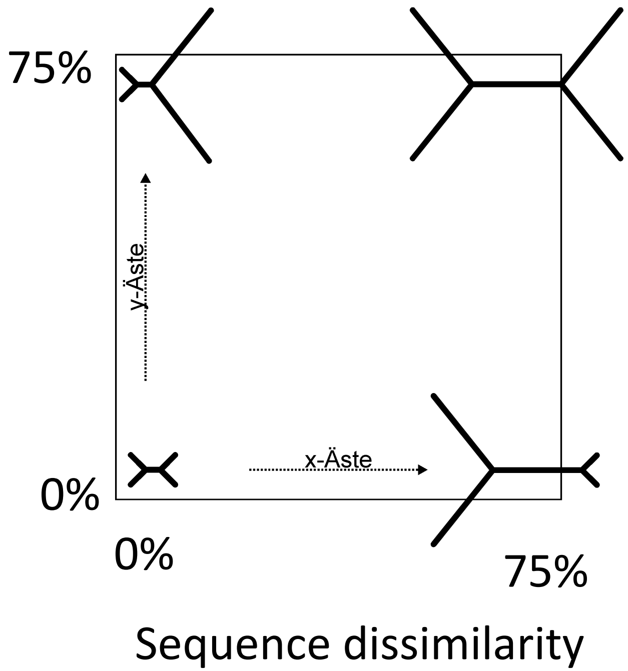


Sequence length

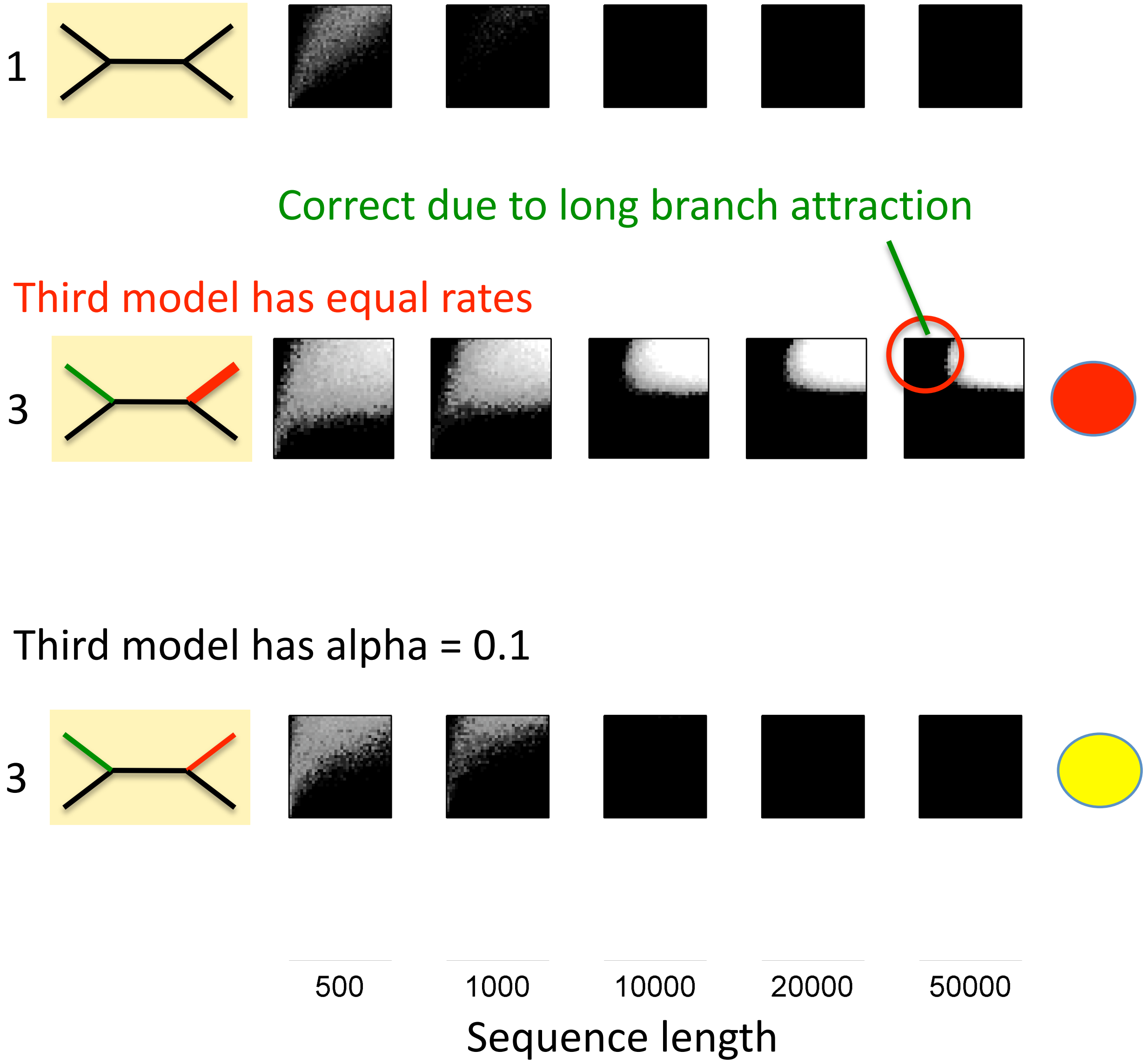
All models: JC + G, alpha = 0.1, Reconstruction: **MP**



Tree shapes:



All models: JC + G, alpha = 0.1, Reconstruction: **MP**

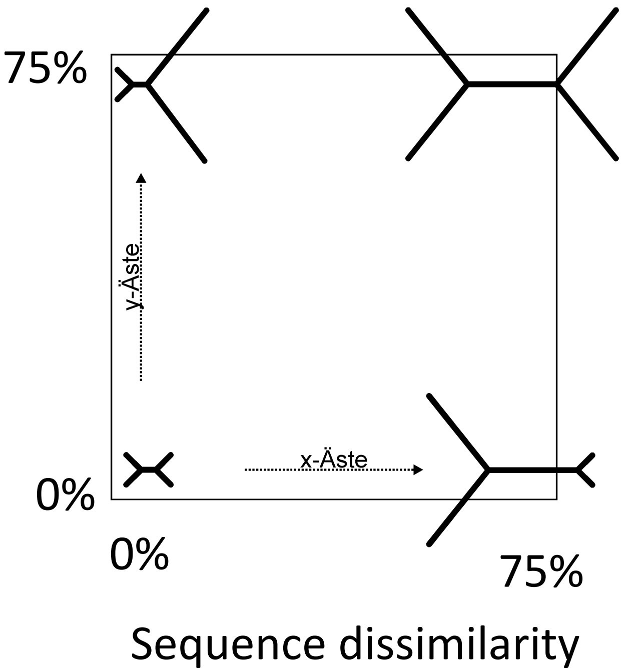


Correct due to long branch attraction

Third model has equal rates

Third model has alpha = 0.1

Tree shapes:



Conclusions

- Heterotachy can strongly decrease and increase phylogenetic accuracy.

- It is worrying that a different model on a single branch decreases the accuracy of ML considerably.

- Likelihood gets strongly affected if heterogeneity differs in different lineages.

Selected References

- N. Galtier. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol*. 18:866–873.
- J.P. Huelsenbeck. Performance of Phylogenetic Methods in Simulation. *Syst. Biol.,* 44(1):17-48, 1995.
- B. Kolaczkowski and J.W. Thornton. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*, 431:980984, 2004.
- B. Kolaczkowski and J.W. Thornton. A mixed branch length model of heterotachy improves phylogenetic accuracy. *Molecular Biology and Evolution*, page (advance access), 2008.
- P. Lopez, D. Casane, and H. Philippe. Heterotachy, an important process in protein evolution. *Molecular Biology and Evolution*, 19(1):1–7, 2002.
- F.A. Matsen and M. Steel. Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Sys. Bio*., 56:767775, 2007.
- D. Penny, B.J. McComish, M.A. Charleston, and M.D. Hendy. Mathematical elegance with biochemical realism: The covarion model of molecular evolution. *Journal of Molecular Evo- lution*, 53:711723, 2001.
- H. Philippe, Y. Zhou, H. Brinkmann, N. Rodrigue, and F. Delsuc. Heterotachy and long- branch attraction in phylogenetics. *BMC Evolutionary Biology*, 5(50), 2005.
- M. Spencer, E. Susko, and A.J. Roger. Likelihood, parsimony and heterogeneous evolution. *Mol. Biol. Evol*., 22:1161–1164, 2005.
- C. Tuffley and M. Steel. 1998. Modeling the covarion hypothesis of nucleotide substitution. *Math. Biosci*. 147:63– 91.