Evaluating the performance of phylogenetic inference methods on heterogeneous data

Mark Holder
Department of Ecology and Evolutionary Biology
University of Kansas
Lawrence, Kansas, USA

Thanks:

- Dr. Ziheng Yang, Dr. Nick Goldman and the Royal Society
- Dr. Derrick Zwickl
- Dr. Gavin Naylor, Dr. Junhyong Kim, Dr. David Hillis, Dr. Paul Lewis, Dr. Dave Swofford, Tracy Heath, Dr. A. Town Peterson
- National Science Foundation

Weaknesses of early models of sequence evolution

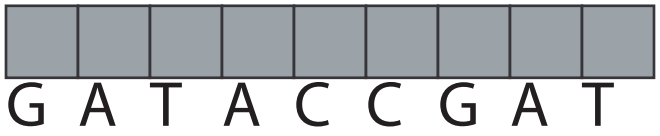
They assumed that

- the process of evolution is identical across all sites,
- the process of evolution is fixed over time,
- sites evolve independently of one another.

Assuming homogeneity across sites

reality:

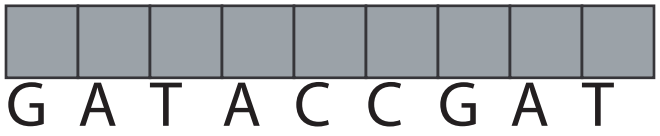
Simplest *iid* model's view:



Among-site rate heterogeneity

reality:

no rate heterogeneity



with rate heterogeneity:

Among-site heterogeneity in substitution patterns

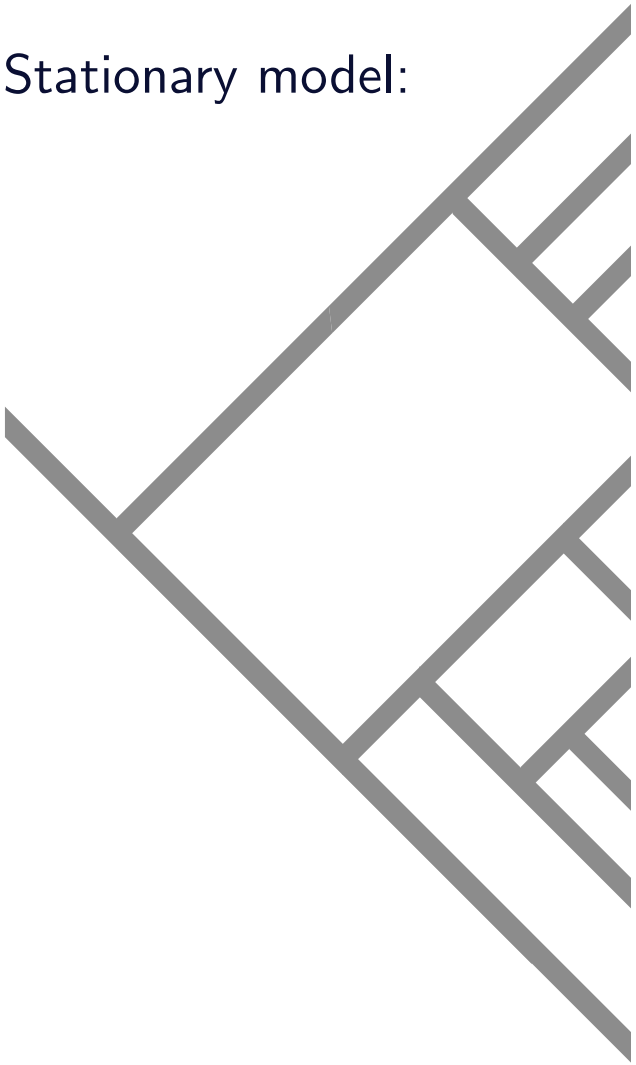
Reality:

Partitioned or mixure model:

Stationarity assumption

Reality (perhaps):

Stationary model:

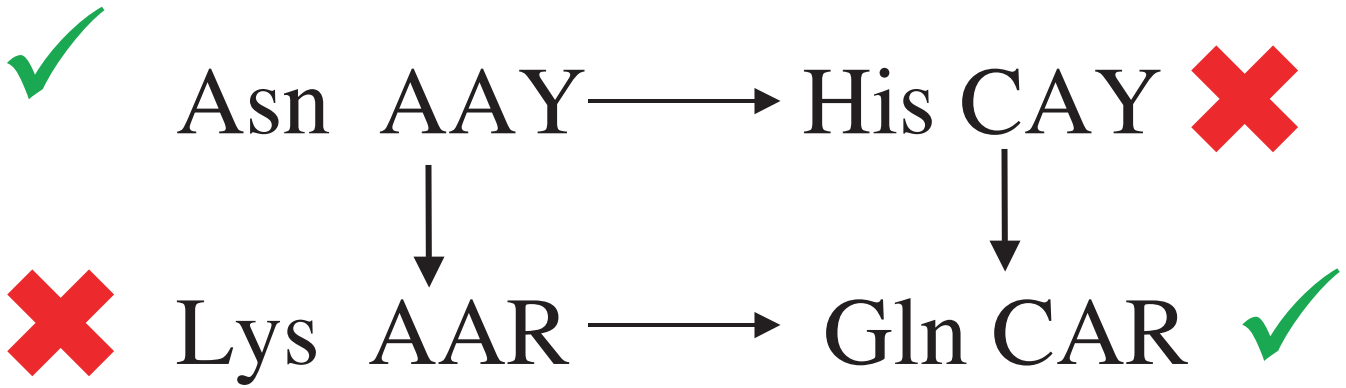


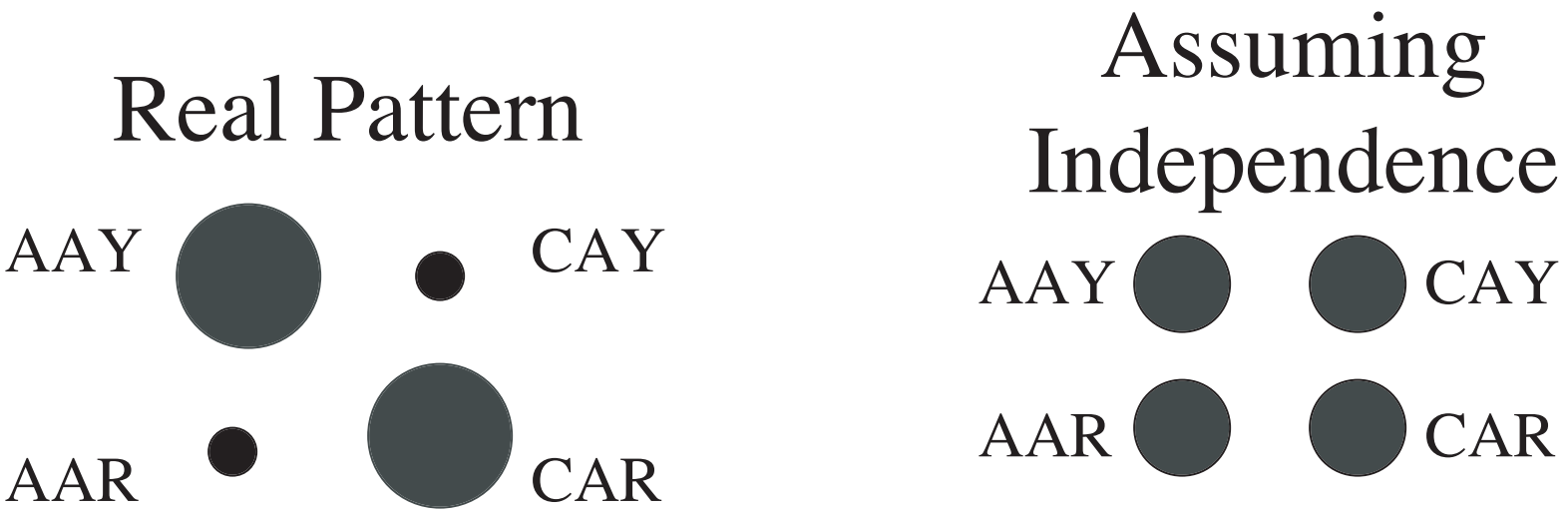
Changes in evolutionary process over time

- Covarion models
- State frequency changes
- models that allow for changes to ω on a specific branch
- relaxed and local clock methods

...

Independence assumption





Interactions among sites and neighborhood effects

- Codon models
- Doublet models for RNA
- Context-dependent models

...

Amazing progress in the development of models for phylogenetic inference, but . . .

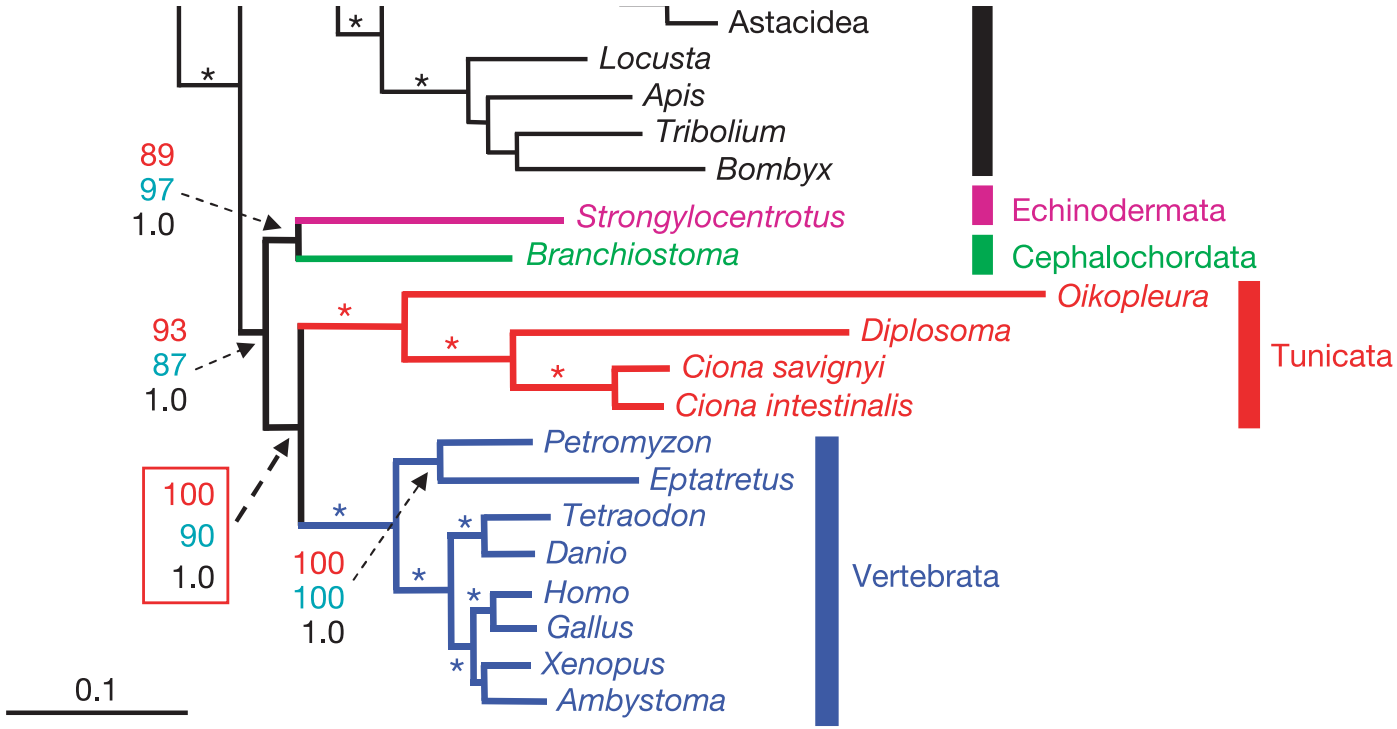
- Are current models sufficient?
- How should we decide between alternative models?

- Are current models sufficient?

  Not for difficult "deep phylogeny" questions. For example, it is fairly common to see strong support for a grouping, but sensitivity to taxon sampling.

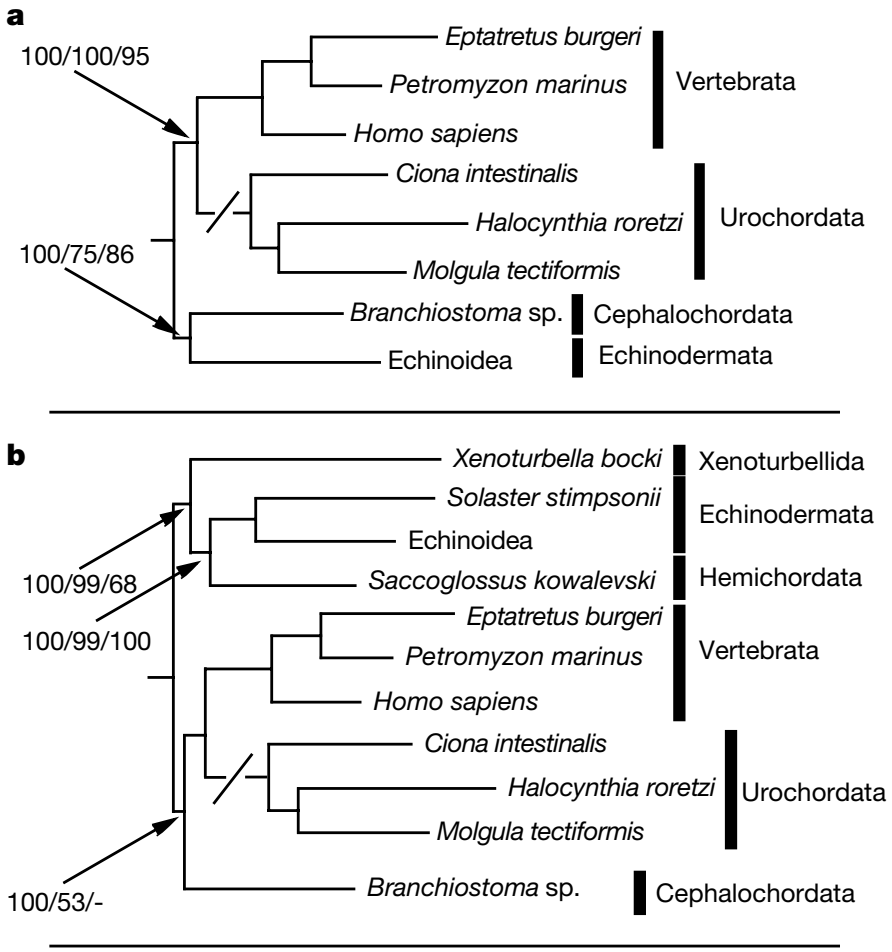
- How should we decide between alternative models?

Delsuc et al. (2006):



Thanks to Gavin Naylor for the slide

Later that year, Bourlat et al. (2006) added taxa including *Xenoturbella*:



Thanks to Gavin Naylor for the slide

- Are current models sufficient?

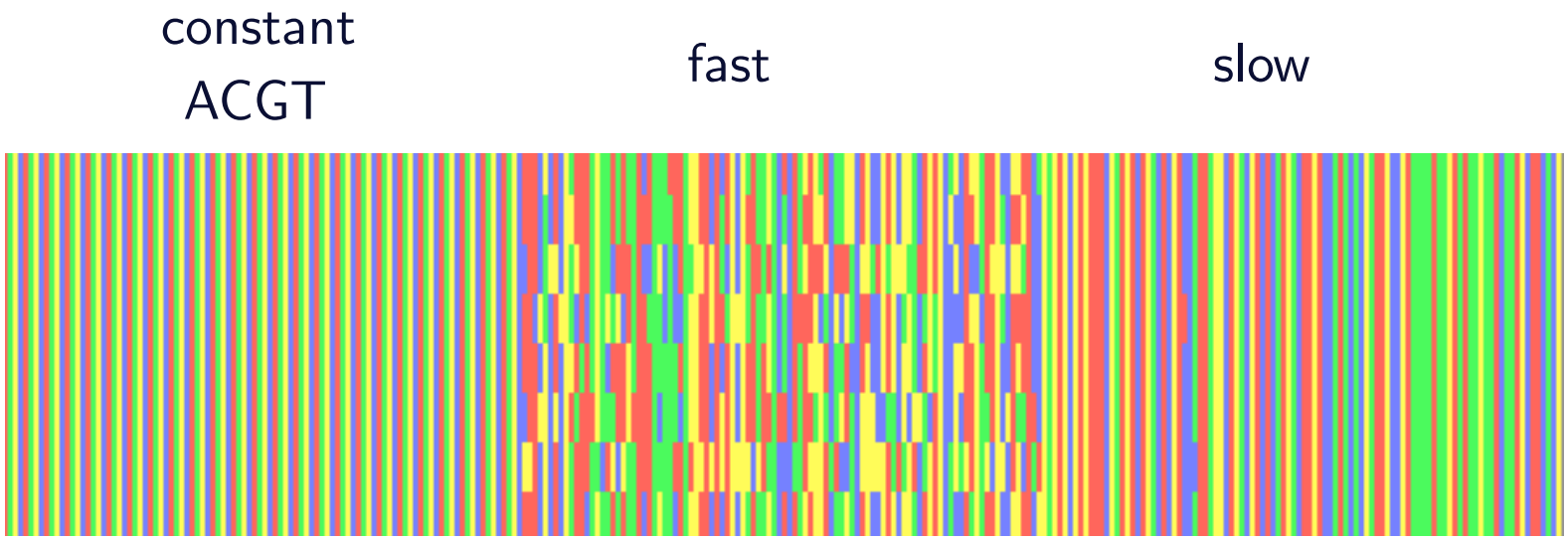
  Not for difficult "deep phylogeny" questions.

- How should we decide between alternative models?

  "Black-box" model choice methods (LRT, AIC, BIC, Bayes Factors, ...) are very helpful at assessing the fit.

  However, they may **not** do a good job of identifying the model that is the best estimator of the parameters that we care about (e.g. the tree).

  This particularly a concern when none of the models considered is close to being "true."



M_1: one rate class – Rejected
M_2: constant-$ACGT$-sites + 1 variable rate class
M_3: 3 rate classes

Either M_2 or M_3 may be chose by an automated model selection procedure, but M_3 is more likely to return the correct tree.

From a paper a *Systematic Biology* by Marshall et al. (2006) comparing models (all of which had Γ-distributed rates).

Model	Tree length	ln L
partitioned by codon, one mean rate	2.00	-9942.24
unpartitioned	1.13	-10414.96

From a paper a *Systematic Biology* by Marshall et al. (2006) comparing models (all of which had Γ-distributed rates).

Model	Tree length	ln L
partitioned by codon, one mean rate	2.00	-9942.24
unpartitioned	1.13	-10414.96
partitioned by codon, SSB	1.13	-9866.67
partitioned by codon, SSR	1.15	-9842.41

See also Buckley et al. (2001) for discussion of performance of subset-specific rate models without Γ-distributed rate heterogeneity.

How should we decide between alternative models?

- Standard model selection tools are certainly *helpful*,
- Performance-based model selection (e.g. Minin et al., 2003),
- Performance of a model on previous real data analyses,
- **Performance of inference methods on data derived from computer simulations.**

Using computer simulations to evaluate models and methods:

- flexible, fast, and transparent

but. . .

- simulations often generate data that is too "clean" (but see work of Junhyong Kim and his collaborators).

- simulated data may bear little resemblance to the real world.

Evaluating performance using computer simulation

1. Choose a simulation model

2. Choose a tree shape

3. Simulate many data sets

4. Infer trees using a variety of methods

5. Compare inferred tree to true (model) tree.

In the following slides we'll deal with a *very* complex, parameter-rich simulation model – Halpern and Bruno (1998) model of site-specific residue frequencies.

Halpern and Bruno (1998) model of coding sequence evolution

- All nucleotides share a set of parameters for a mutational model,
- Each amino acid residue has a set of equilibrium frequencies

Halpern and Bruno (1998) model properties

1. Extreme heterogeneity of process

2. Sites (within the same codon) are not independent

3. Among-site rate heterogeneity, but not following a simple distribution (such as the Γ-distribution).

I wrote software to:

1. find MLEs of parameters of the Halpern and Bruno (1998) model, and

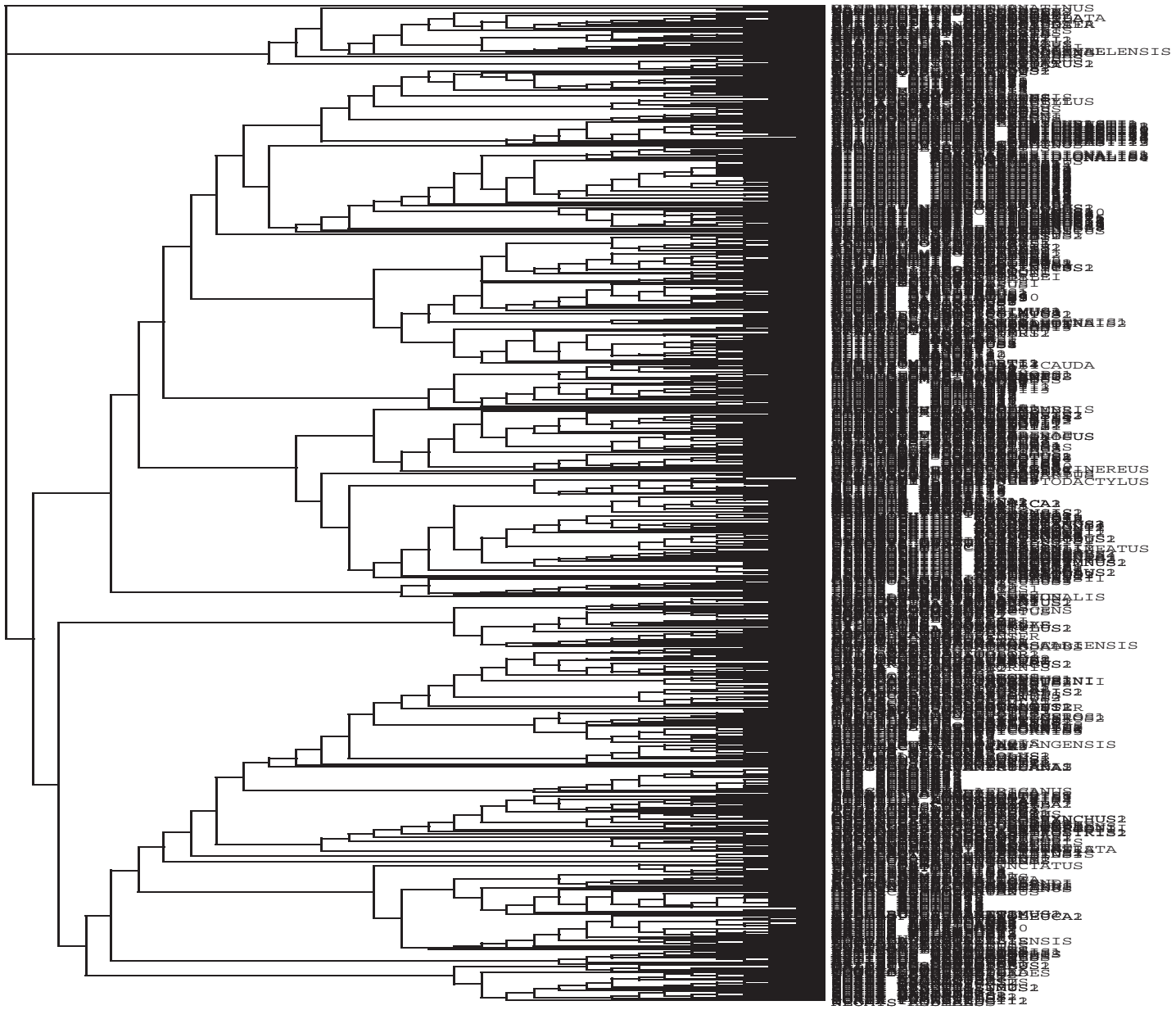
2. simulate under the Halpern-Bruno model (on a user-defined tree).

Estimated parameters on a dataset of 1610 mammalian cytochrome-b sequences (376 amino acid residues, $\approx 7,150$ substitution parameters)

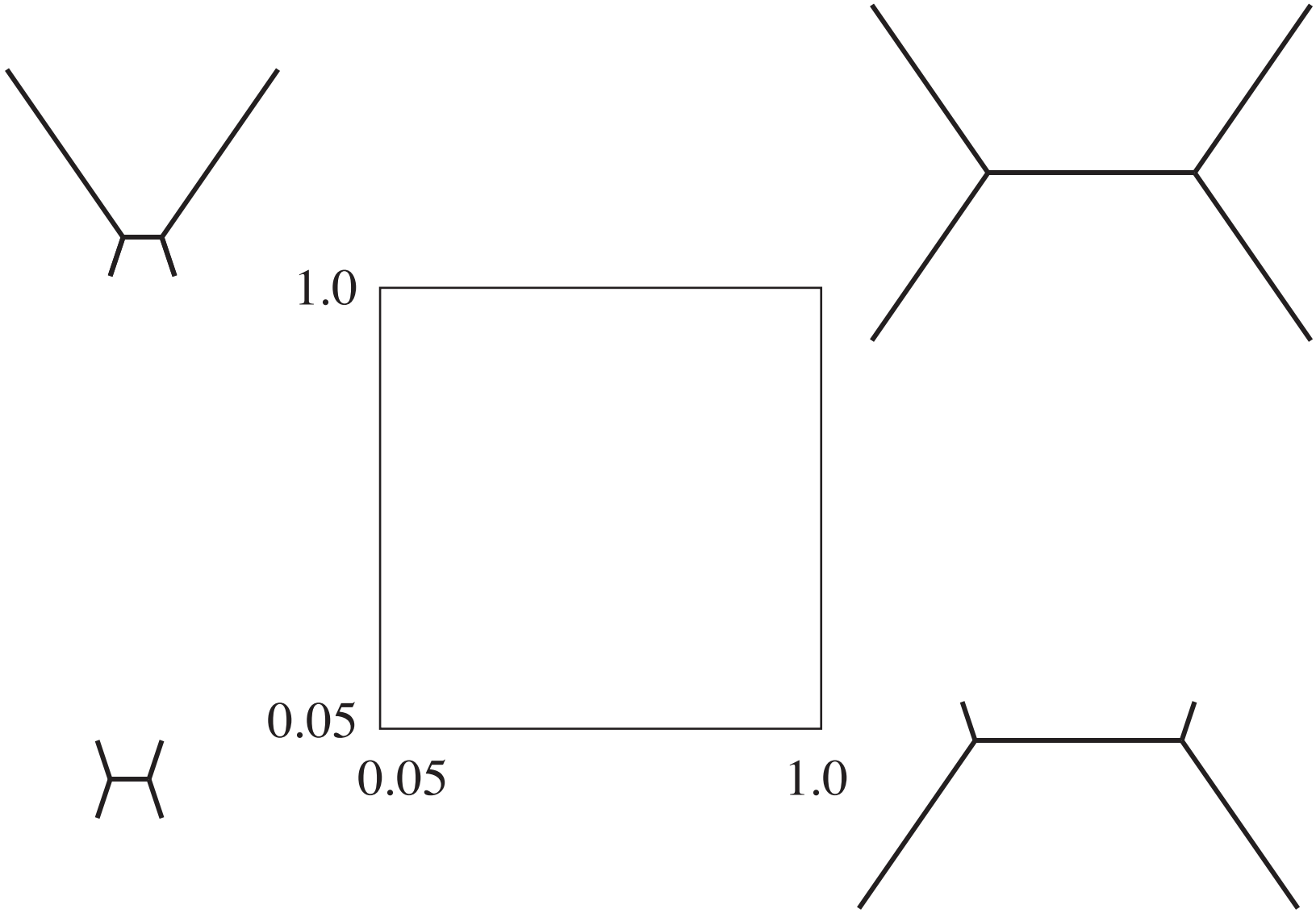
`http://nladr-cvs.sdsc.edu/svn/CIPRES/cipresdev/trunk/python-example/org.cipres.bull`

username: guest

password: guest

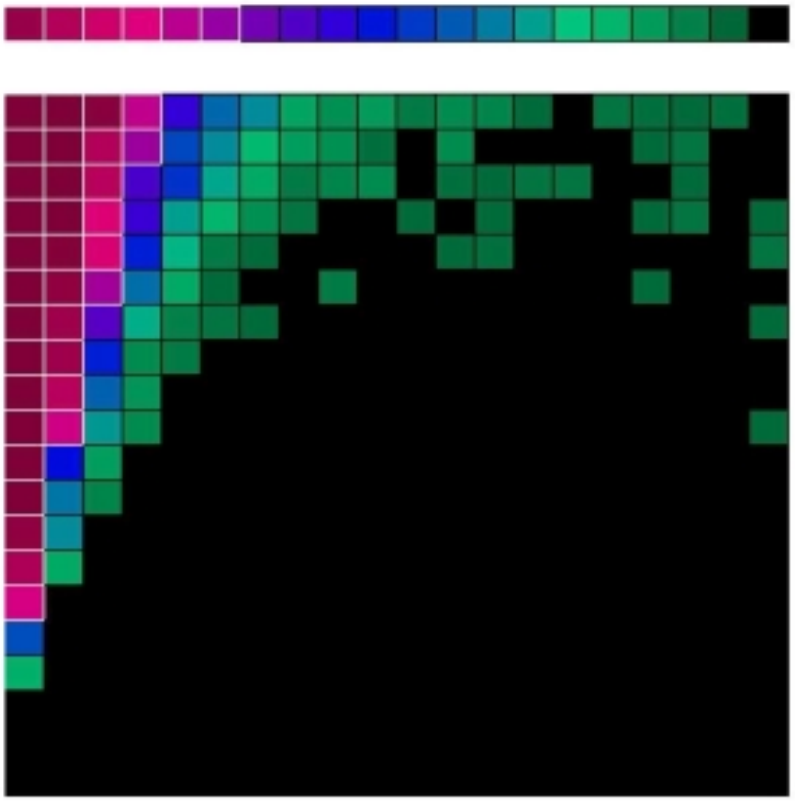


Space of 4-taxon simulation trees

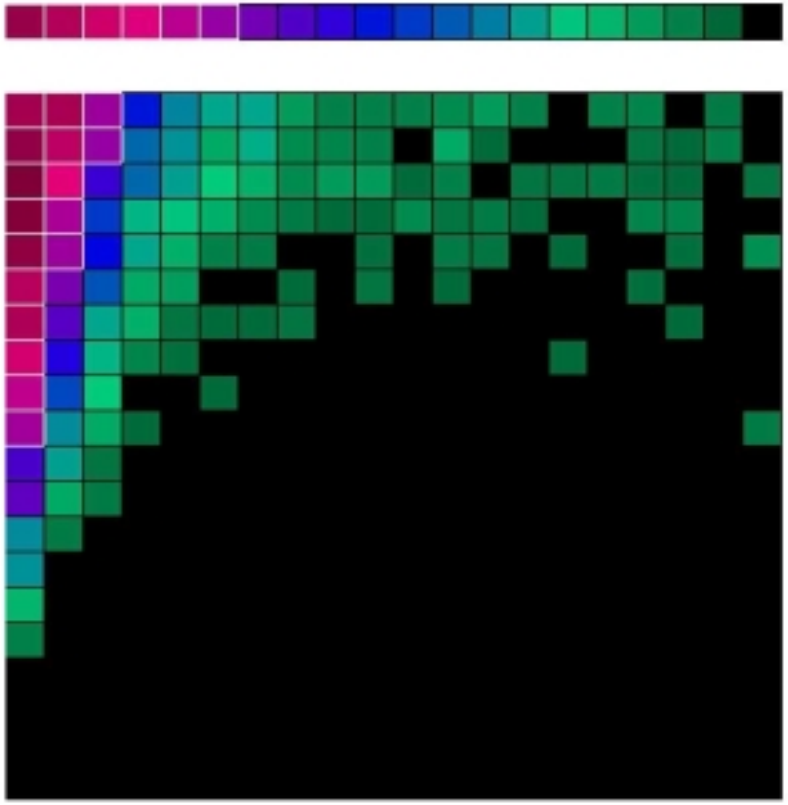


Parsimony and maximum likelihood

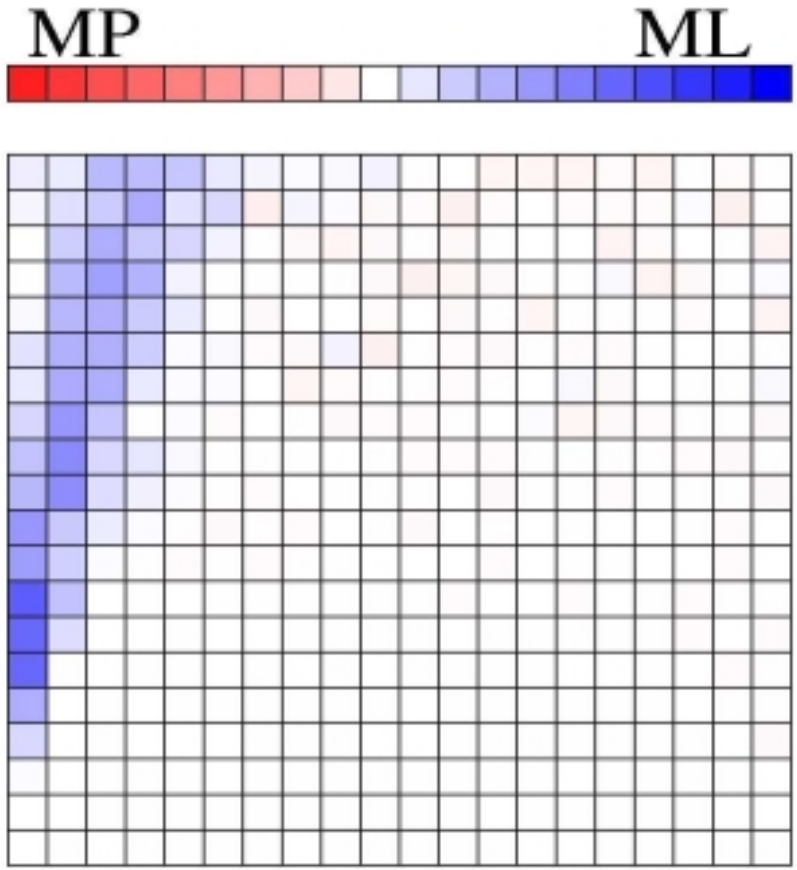
Parsimony with step matrix



Maximum likelihood (GTR + rate het.)

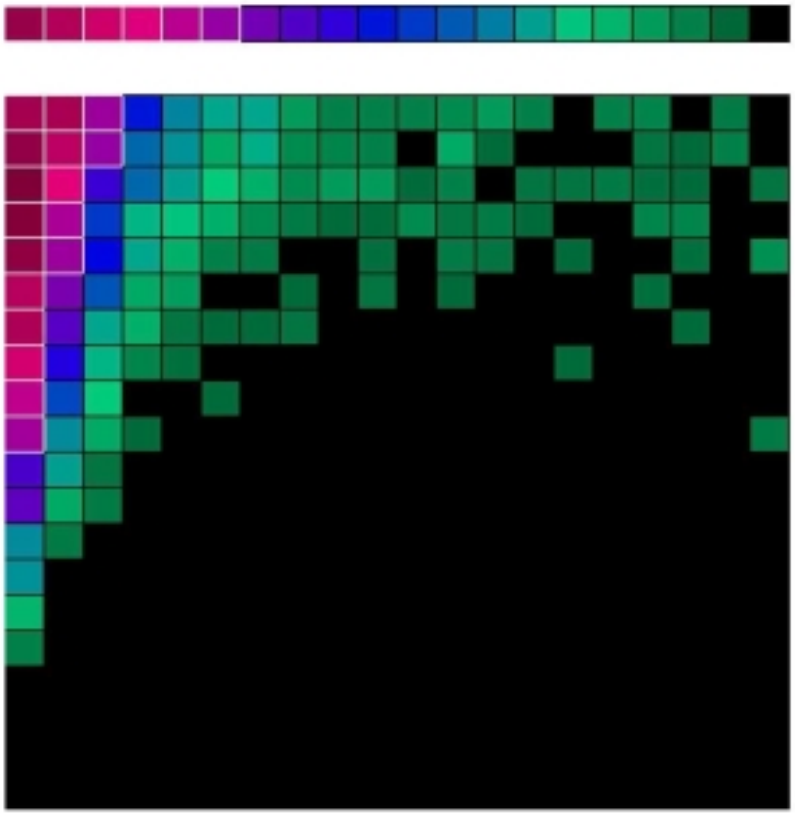


Contrasting parsimony and maximum likelihood

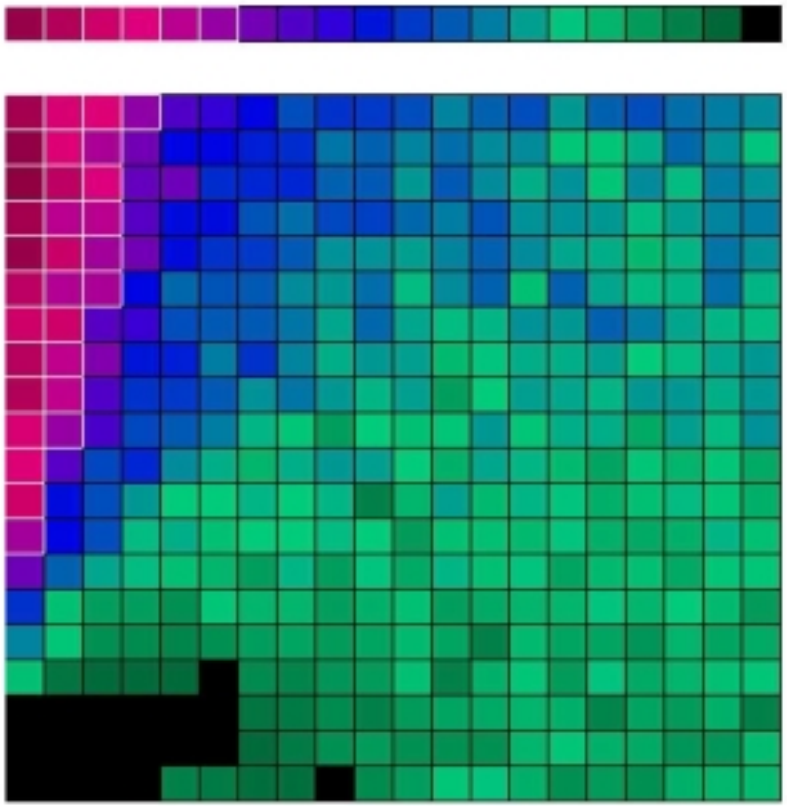


Maximum likelihood and distance (GTR + rate het.)

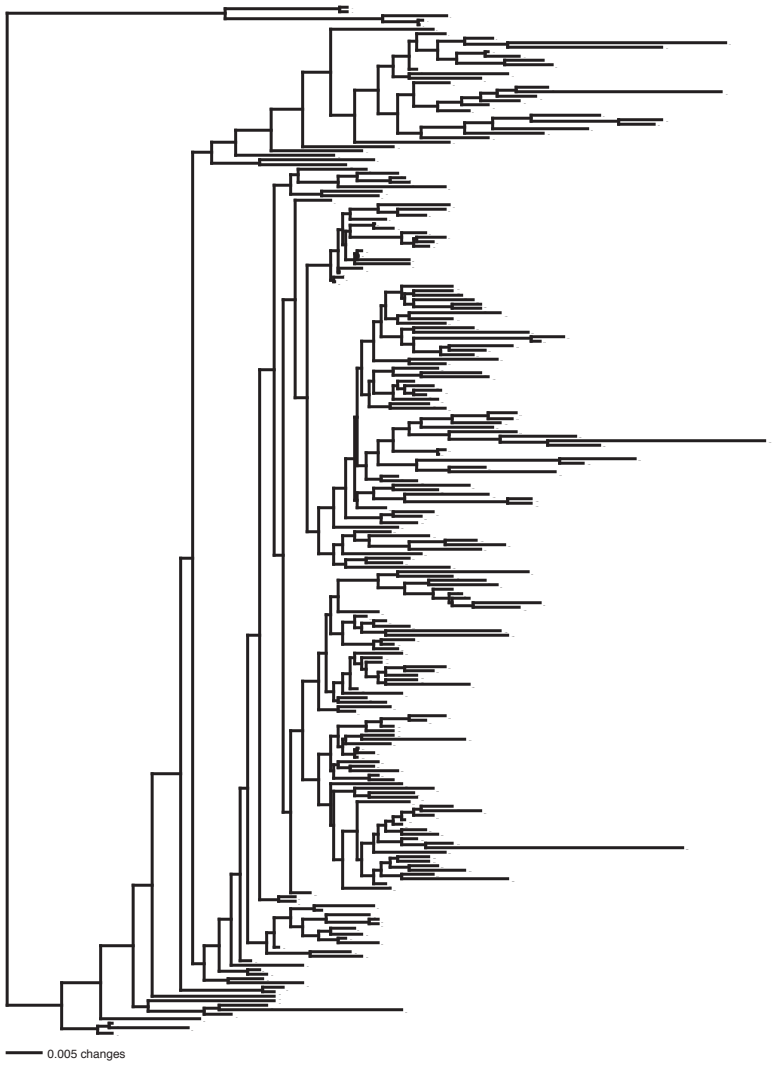
Maximum likelihood



Minimum evolution

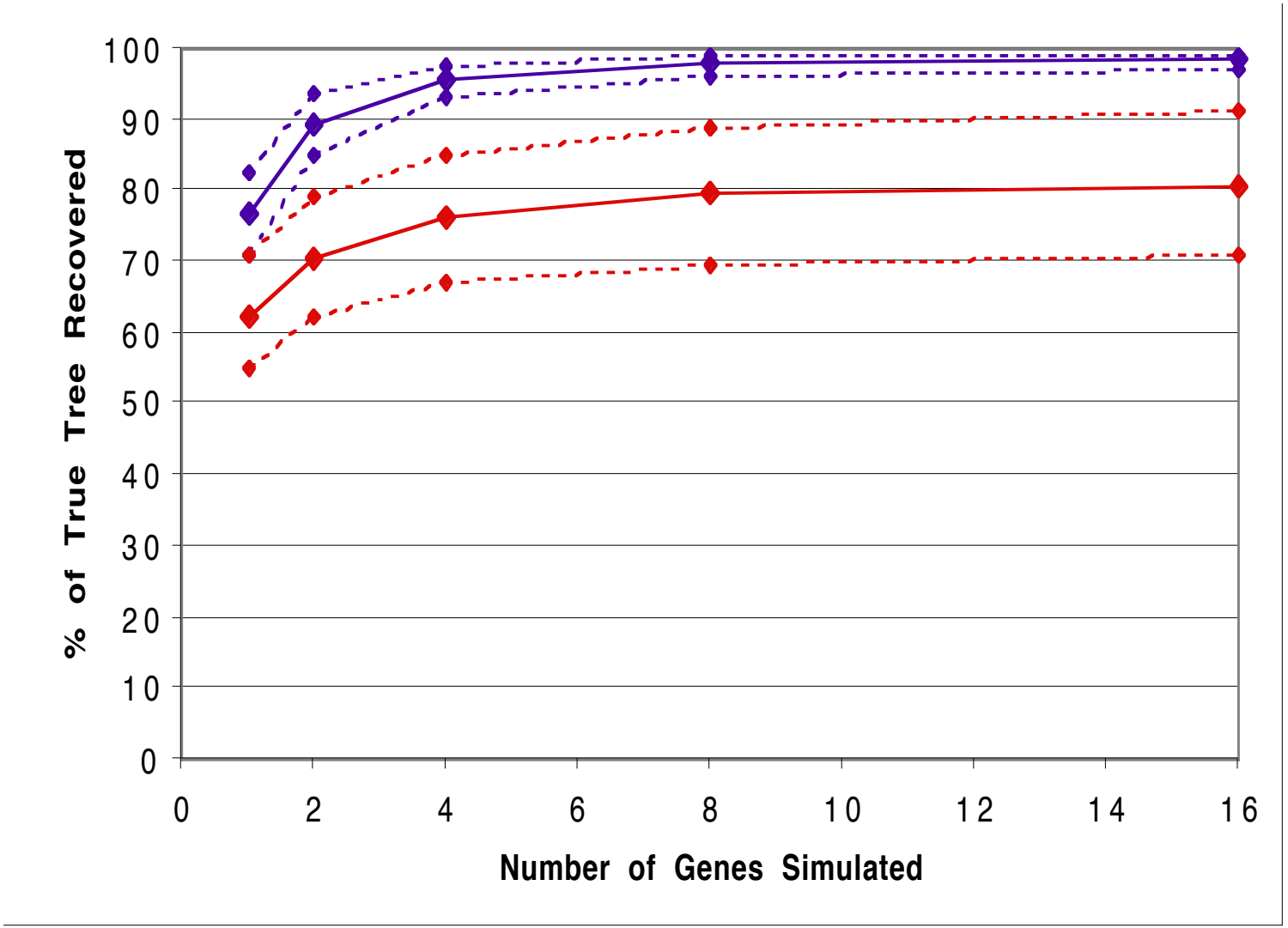


228-taxon model tree

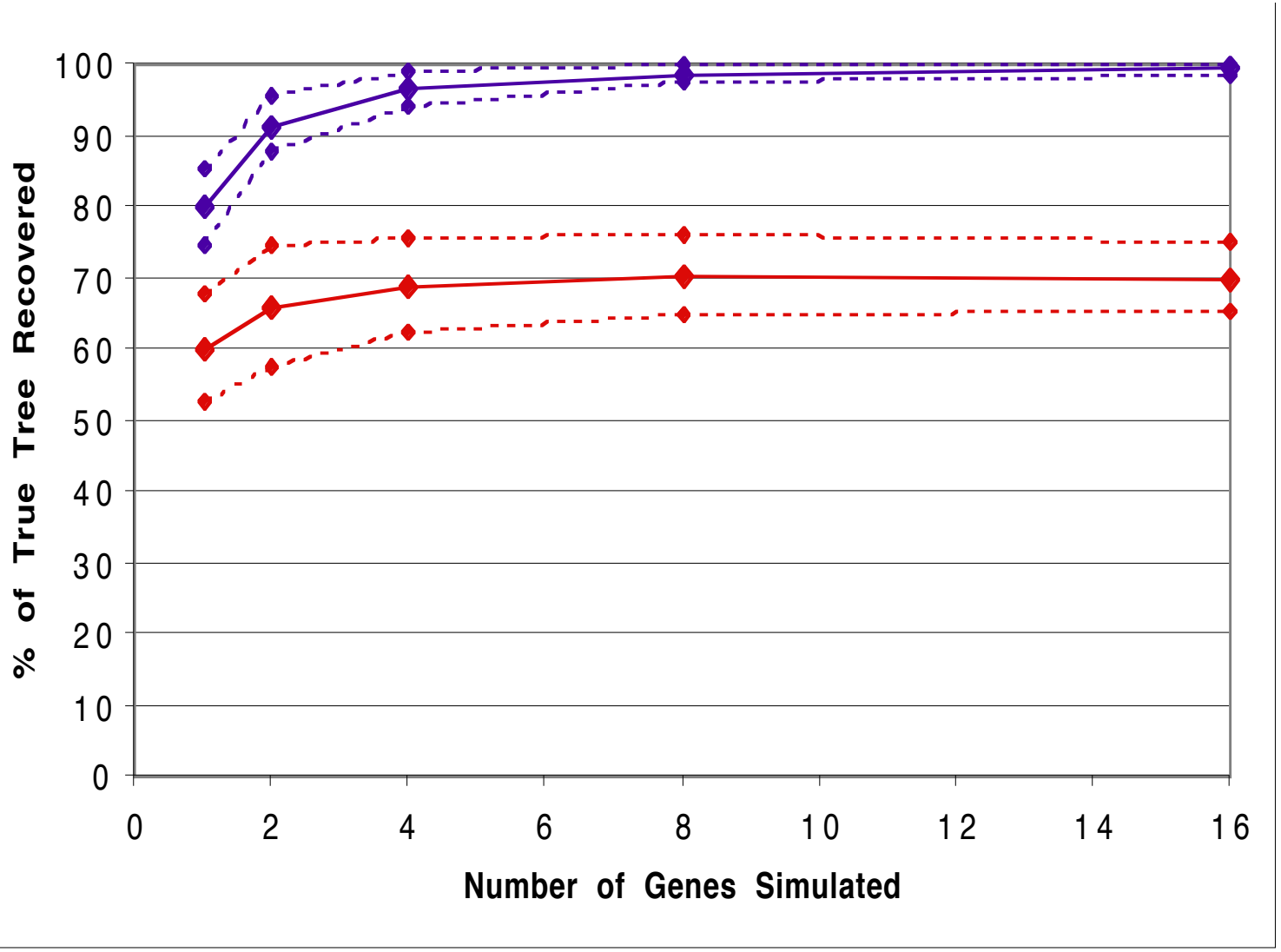


- Parsimony tree for angiosperm tree
- Model tree for Hillis (1996, 1998)
  (blue on the next slides)
- Inferred branch lengths and 10×
  (red on the next slides)

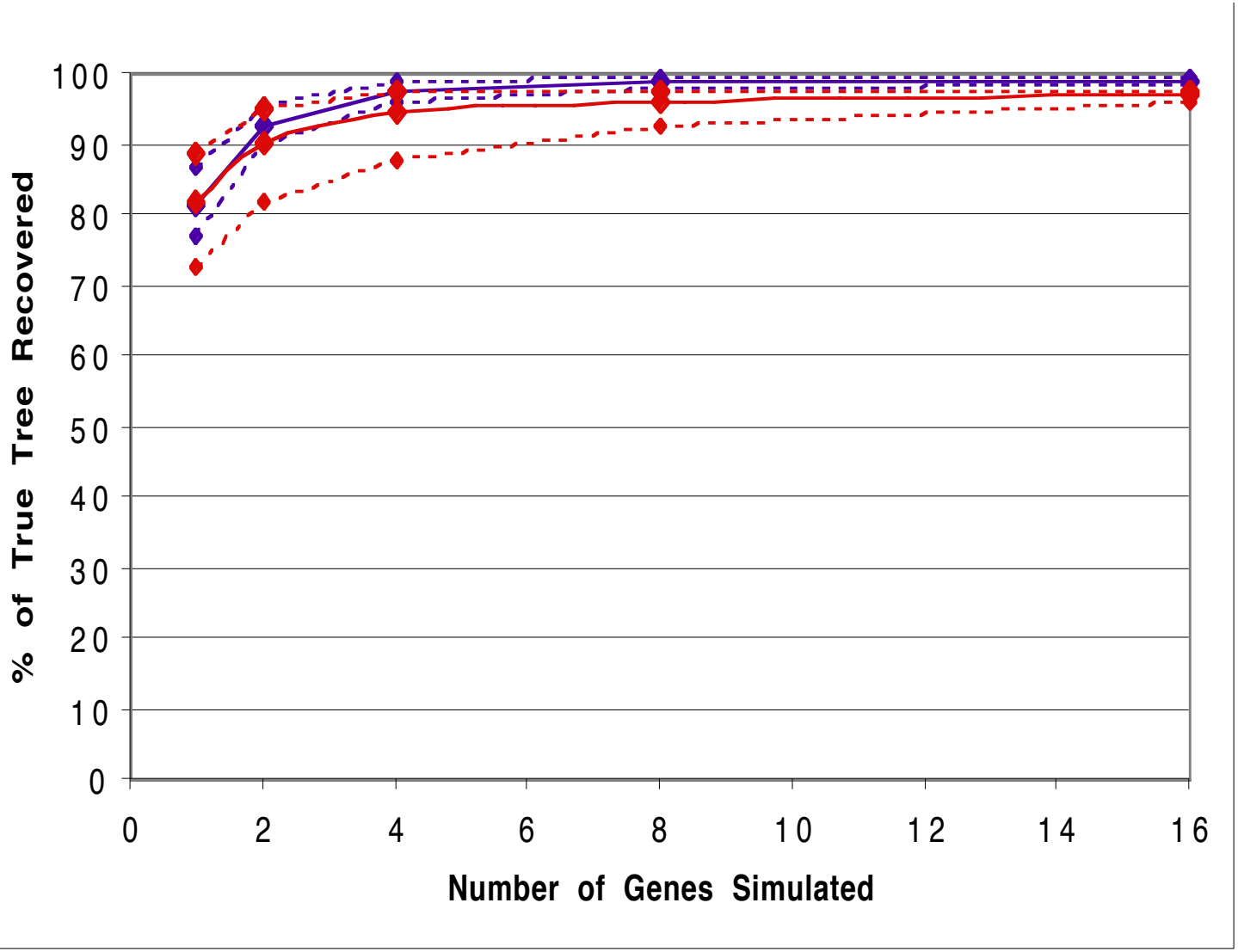
228-taxon tree: Parsimony (stepwise addition)



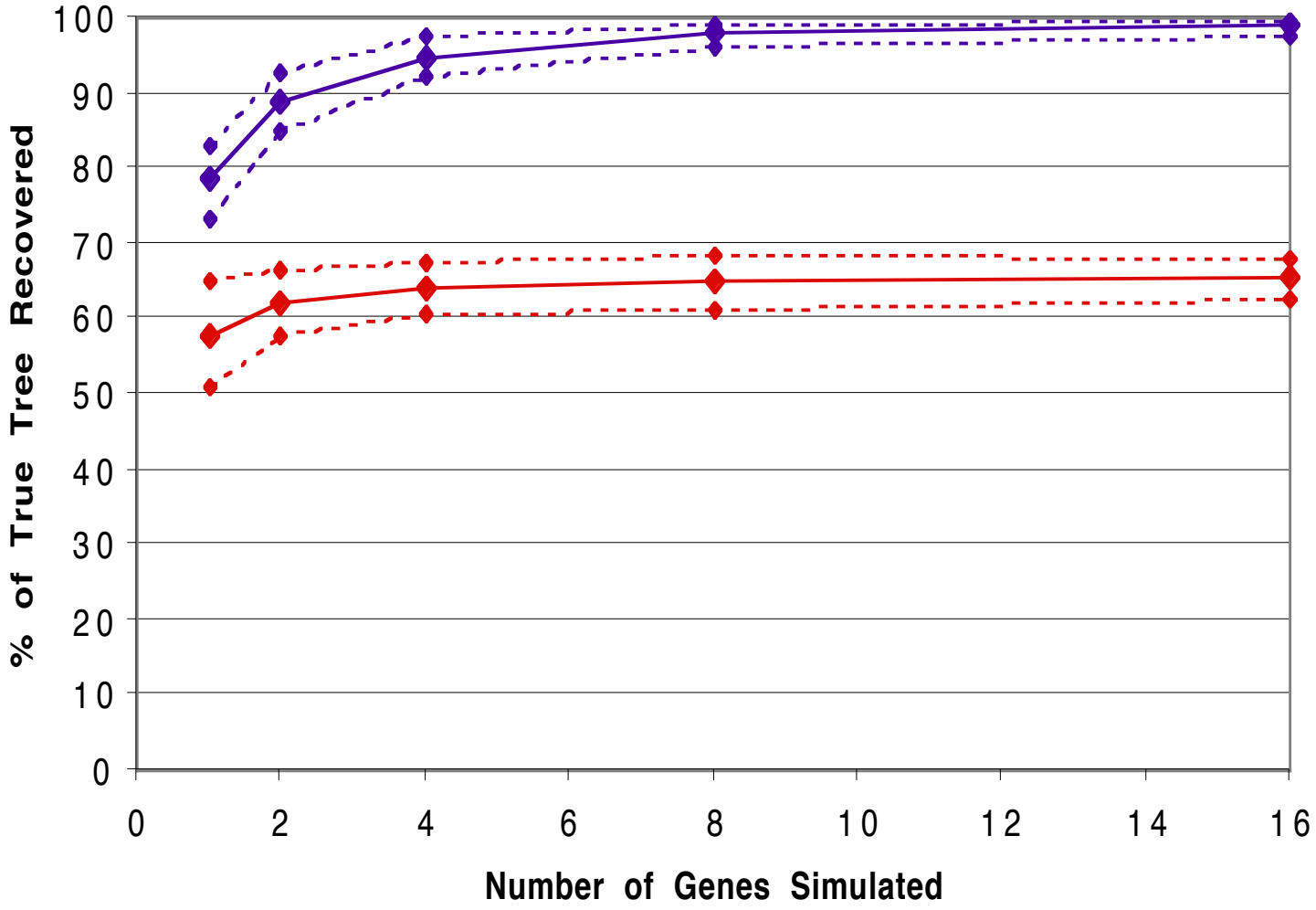
228-taxon tree: Neighbor-joining



228-taxon tree: Parsimony (searching using SPR)



228-taxon tree: More thorough distance searches (minimum evolution)

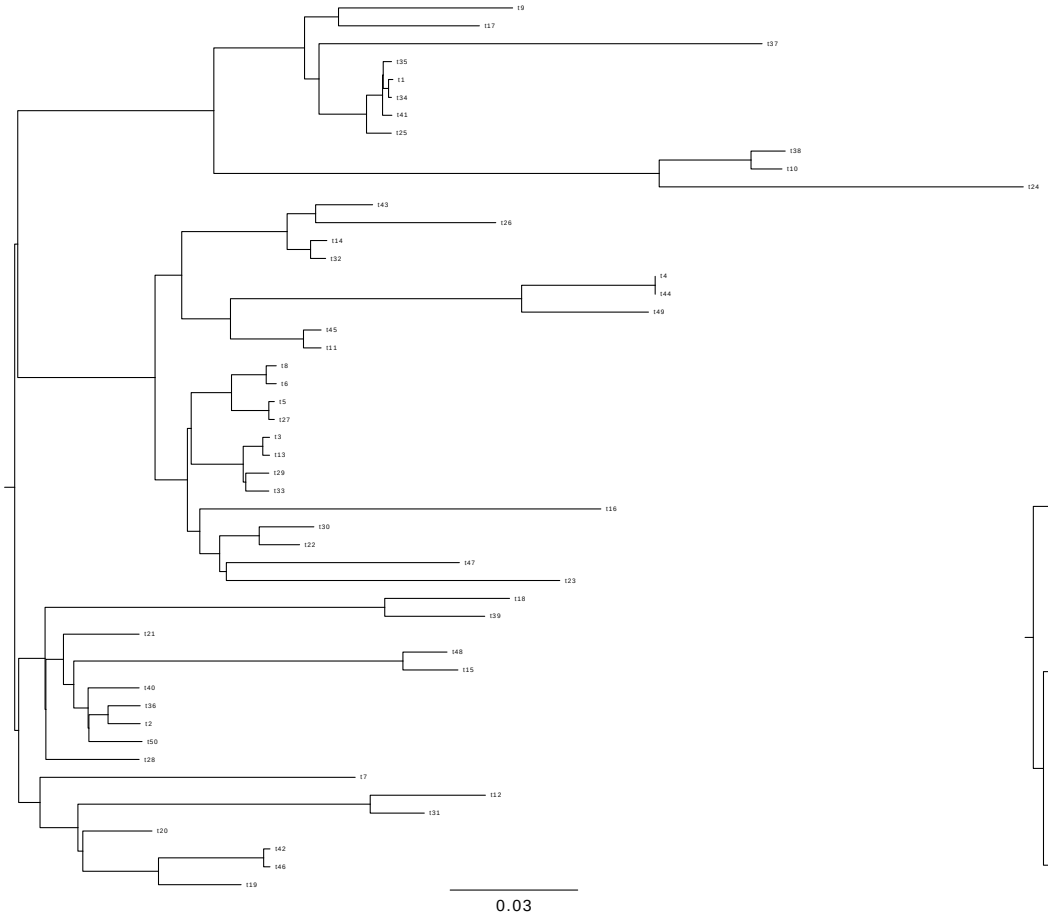


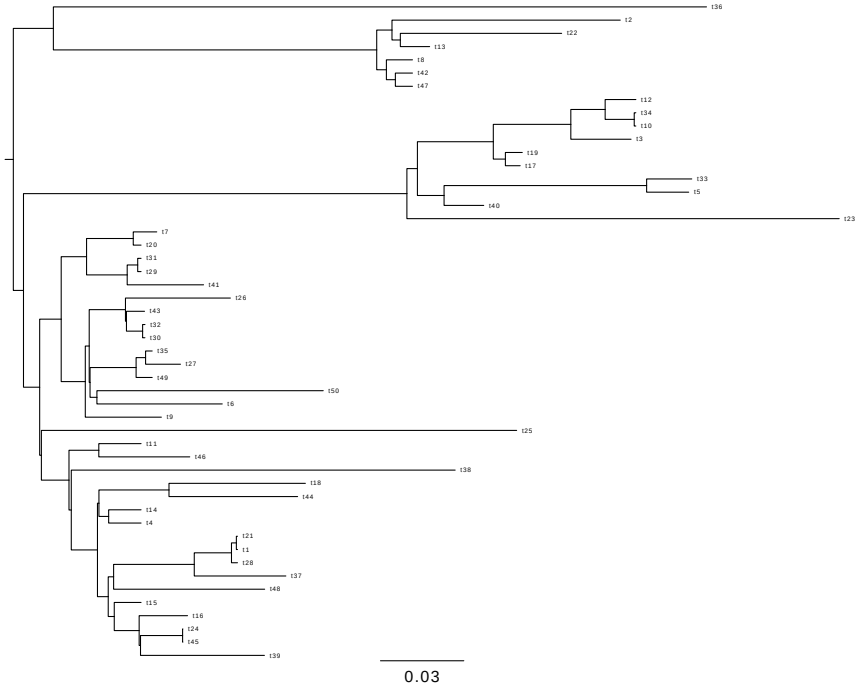
Randomly generated trees

Model trees for 50 leaves generated:

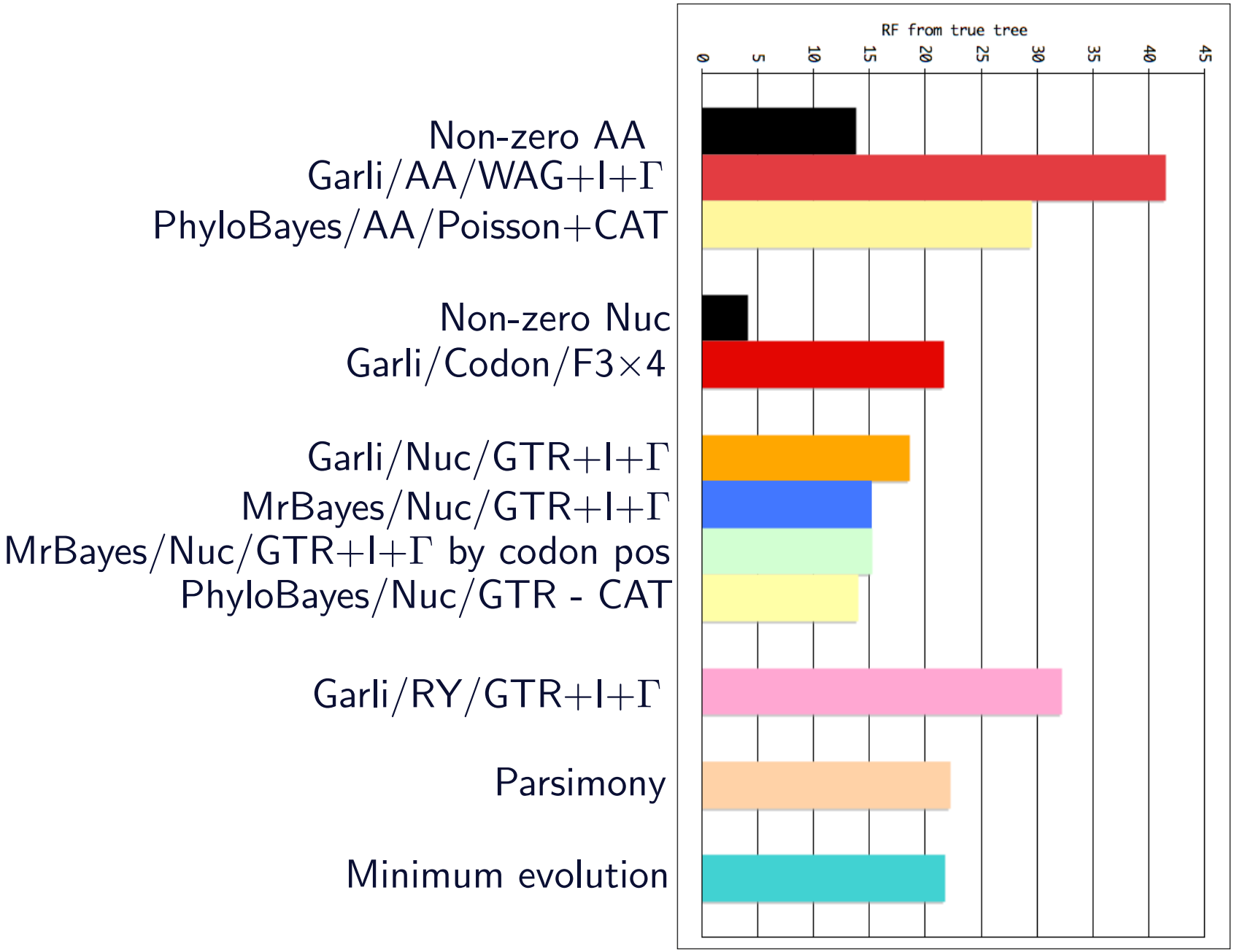
- by a pure birth process, followed by selection of 50% of the taxa.
- rate of evolution then allowed to evolve along the tree Kishino et al. (2001)
- mean branch length has an expected # changes per site ≈ 0.02

The following results are **very** preliminary.

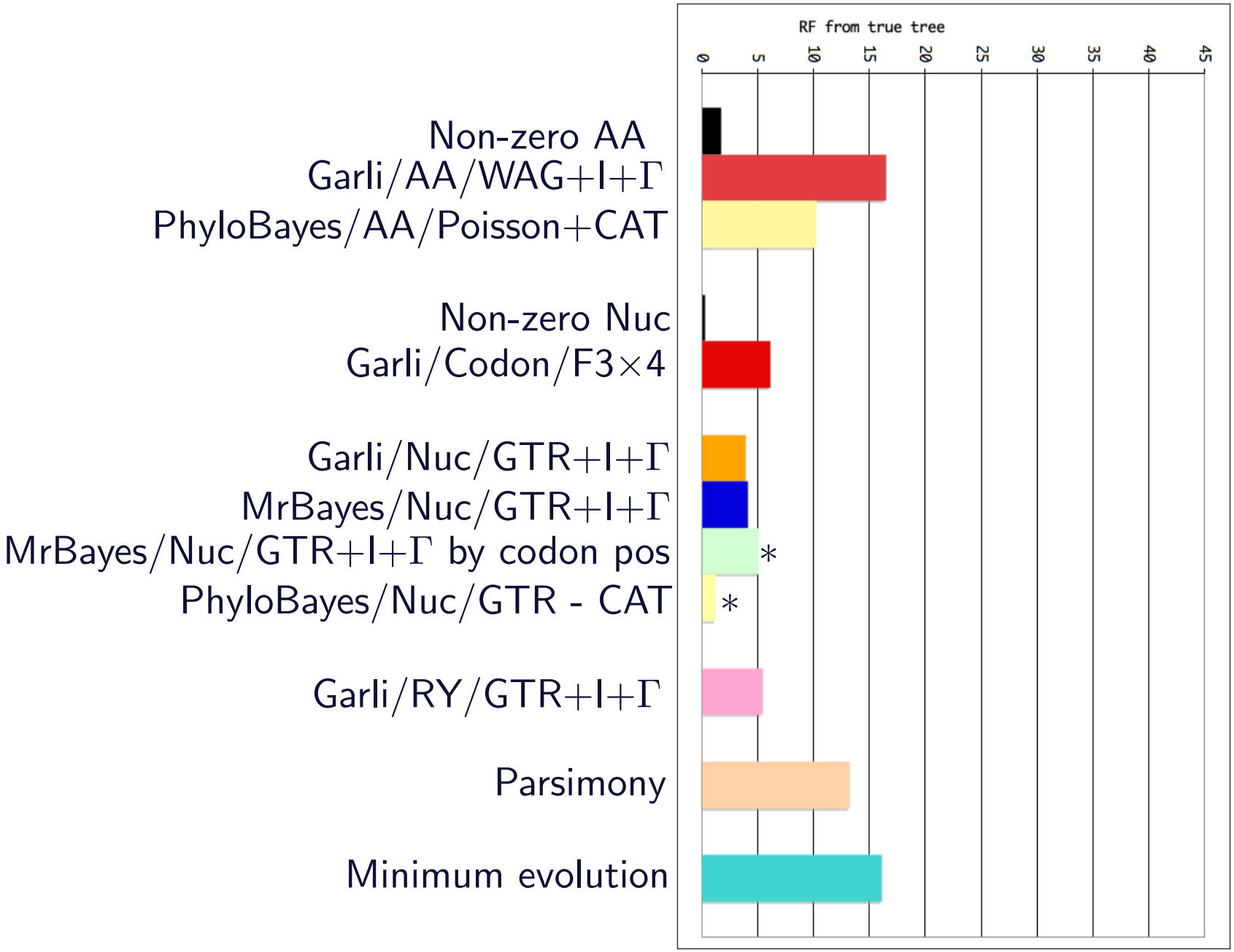

t9
t17
t37
t35
t1
t34
t41
t25
t38
t10
t24
t43
t26
t14
t32
t4
t44
t49
t45
t11
t8
t6
t5
t27
t3
t13
t29
t33
t16
t30
t22
t47
t23
t18
t39
t21
t48
t15
t40
t36
t2
t50
t28
t7
t12
t31
t20
t42
t46
t19
0.03


t36
t2
t22
t13
t8
t42
t47
t12
t34
t10
t3
t19
t17
t33
t5
t40
t23
t7
t20
t31
t29
t41
t26
t43
t32
t30
t35
t27
t49
t50
t6
t9
t25
t11
t46
t38
t18
t44
t14
t4
t21
t1
t28
t37
t48
t15
t16
t24
t45
t39
0.03

Performance on a simulation of 1 copy of cyt. b



Performance on a simulation of 10 copies of cyt. b



Conclusions from simulations under Halpern-Bruno model fit to cytochrome b data

- in general, likelihood-based inference on the nucleotide level appears robust,
- mixture/partitioned approaches performing well,
- parsimony's performance was quite variable (depending on model tree shape),
- distance methods are much more sensitive to model violation,
- analyses at the amino acid level were substantially less accurate

however...

Caveats about simulations under Halpern-Bruno model fit to cytochrome b data

- On small trees, the simulator generates amino sequences with relatively little variation
  - cyt b is constrained,
  - overfitting of amino acid parameters could be decreasing the variability
- Still too simple:
  - homogeneity of mutational process (over space and time),
  - lack of codon bias

  Could cause the nucleotide-level conclusions to be too optimistic.

These drawbacks can be addressed by fitting parameters from more genes, and using parameter estimates from other clades estimate the rate of change in parameter values over time.

Perspective and Discussion questions

Devising realistic, complex simulators still an open area of research. Such tools could allow us to compare fundamentally different analysis styles (e.g. codon vs. nucleotide vs amino acid), and could provide sound guidance about the appropriateness of a model – evidence to be used in conjuction with standard model choice techniques.

- Is it feasible to develop compelling simulators? Can we really make them complex, and yet realistic without knowing the "true" model?

- Suppose that AIC (or your-favorite-model-selection) framework strongly prefers model X over model Y. Would a simulation study saying that model Y is more robust and reliable sway you?

References

Bourlat, S., Juliusdottir, T., Lowe, C., and Freeman, R. (2006). Deuterostome phylogeny reveals monophyletic chordates and the new phylum xenoturbellida. *Nature*.

Buckley, T. R., Simon, C., and Chambers, G. K. (2001). Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. *Systematic Biology*, 50(1):67–86.

Delsuc, F., Brinkmann, H., Chourrout, D., and Philippe, H. (2006). Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*.

Halpern, A. L. and Bruno, W. J. (1998). Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular Biology and Evolution*, 15(7):910–917.

Hillis, D. M. (1996). Inferring complex phylogenies. *Nature*, 383(6596):130–131.

Hillis, D. M. (1998). Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Systematic Biology*, 47(1):3–8.

Kishino, H., Thorne, J. L., and Bruno, W. J. (2001). Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Molecular Biology and Evolution*, 18:352–361.

Marshall, D., Simon, C., and Buckley, T. (2006). Accurate branch length estimation in partitioned bayesian analyses requires accommodation of among-partition rate variation and attention to branch length priors. *Systematic Biology*, 55(6):993–1003.

Minin, V., Abdo, Z., Joyce, P., and Sullivan, J. (2003). Performance-based selection of likelihood models for phylogeny estimation. *Systematic Biology*, 52(5):674–683.