Lie-Markov DNA models are superior to time-reversible models when evolution is heterogeneous

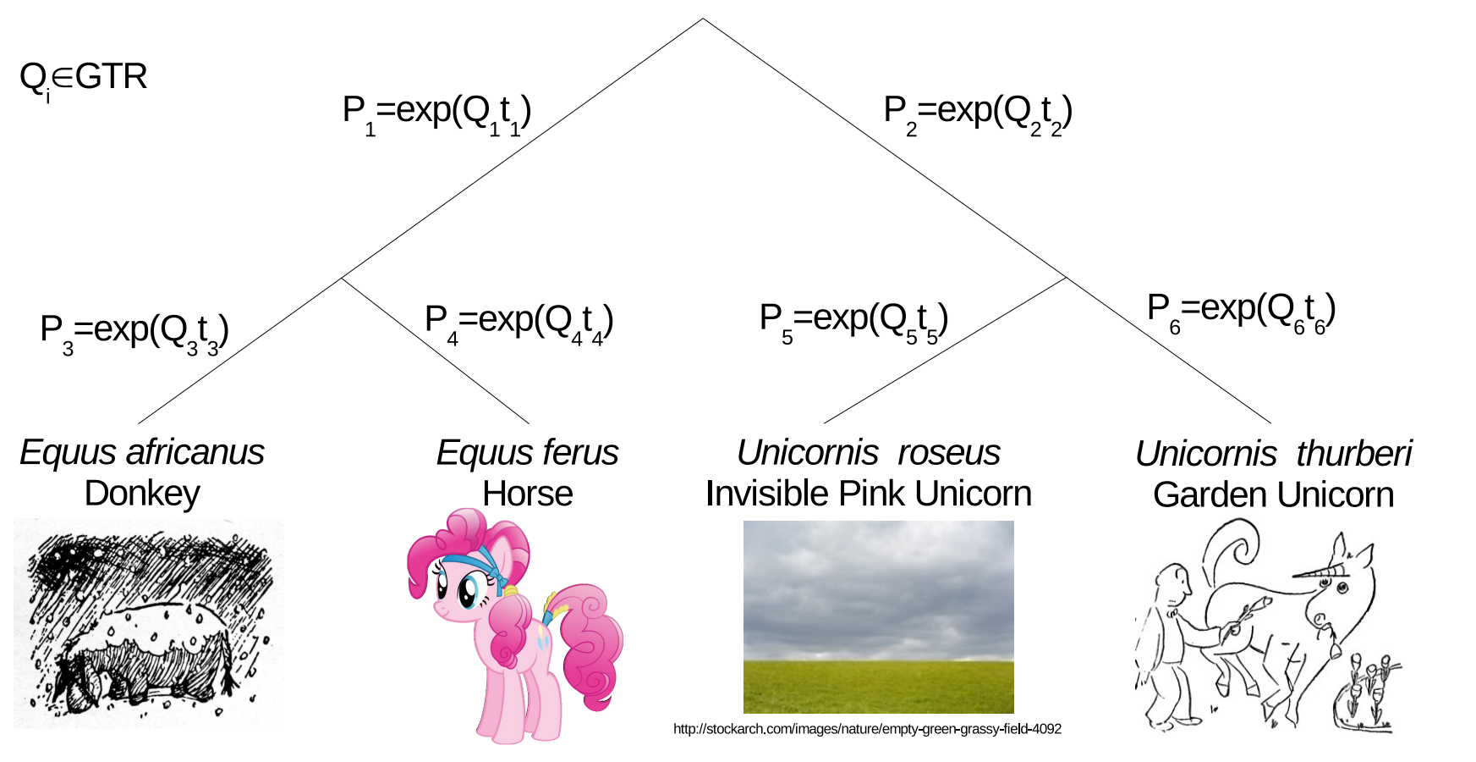
UNIVERSITY of TASMANIA

Michael Woodhams, Barbara Holland, Michael Charleston, Jeremy Sumner

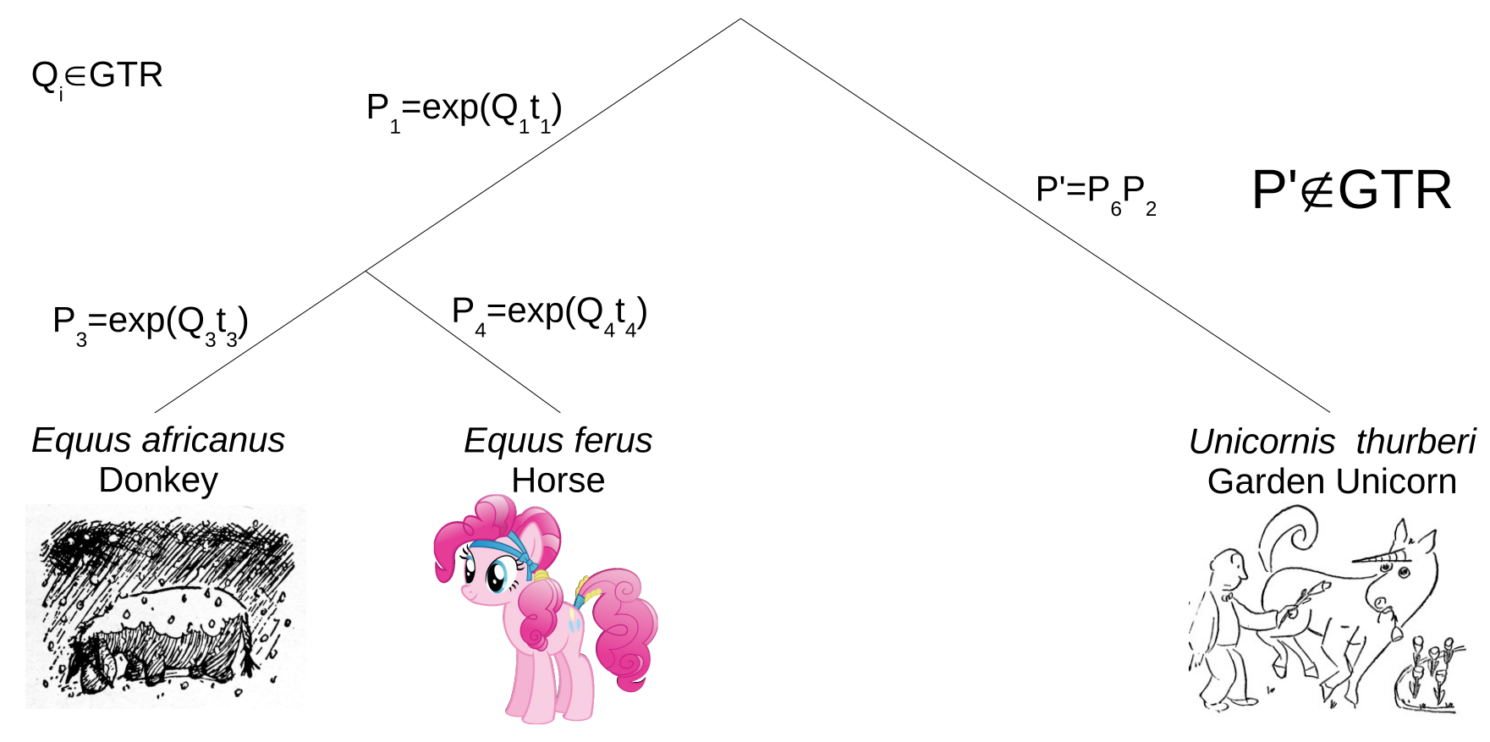
Introduction

It is normal to model DNA evolution as a stationary, reversible, homogeneous process. These assumptions simplify likelihood calculations but are often biologically unrealistic, e.g. in the presence of base composition heterogeneity.

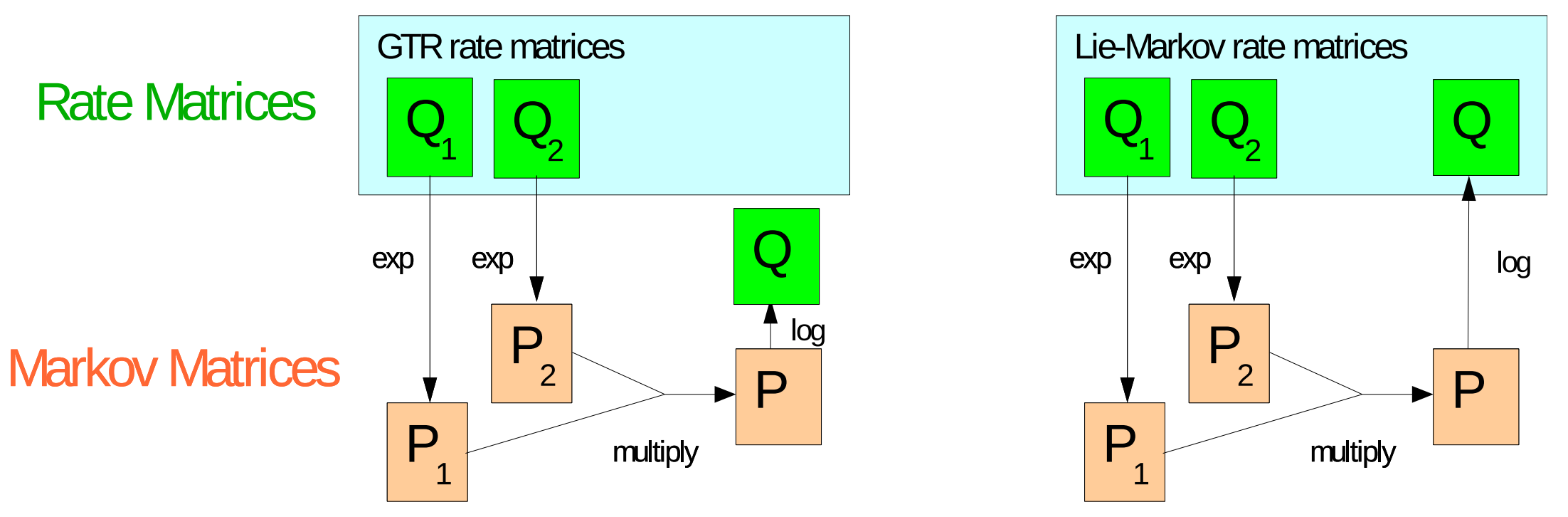
Discarding homogeneity, we potentially have different rate matrices on each branch:



Supposing *U. roseus* was not sampled, this becomes:



Despite Q_2 and Q_6 being GTR, the edge from root to *U. thurberi* is inconsistent with the GTR model.



Closure

When multiplying two Markov matrices from a 'closed' model, the result is also in the model.

GTR and many other time-reversible models lack this property.

Lie-Markov models

The defining property of the Lie-Markov models is that they are closed.

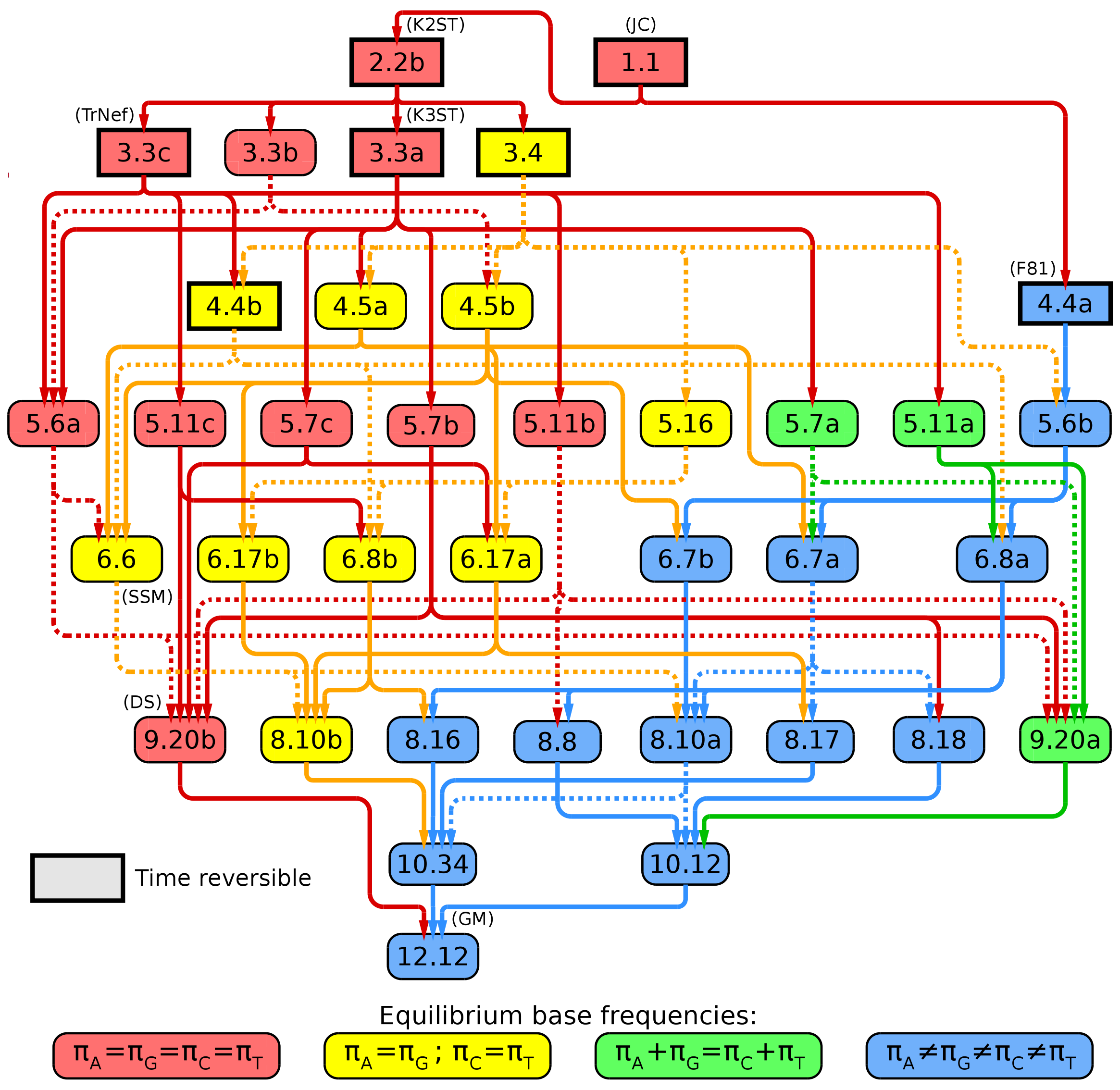
The rate matrices of time-reversible models generally involve *products* of parameters, e.g.

$$Q_{HKY} = \begin{pmatrix} * & \alpha\pi_G & \beta\pi_C & \beta\pi_T \\ \alpha\pi_A & * & \beta\pi_C & \beta\pi_T \\ \beta\pi_A & \beta\pi_G & * & \alpha\pi_T \\ \beta\pi_A & \beta\pi_G & \alpha\pi_C & * \end{pmatrix}$$

Lie-Markov models are always *linear* in their parameters, e.g.

$$Q_{5.6b} = \begin{pmatrix} * & \alpha+\rho_G & \beta+\rho_C & \beta+\rho_T \\ \alpha+\rho_A & * & \beta+\rho_C & \beta+\rho_T \\ \beta+\rho_A & \beta+\rho_G & * & \alpha+\rho_T \\ \beta+\rho_A & \beta+\rho_G & \alpha+\rho_C & * \end{pmatrix}$$

Our complete hierarchy of 37 Lie-Markov models distinguishing transitions and transversions



Base frequency degrees of freedom (BDF)

i. Time-reversible models have BDF=0 (e.g., K2ST), BDF=1 (T92), or BDF=3 (HKY85, GTR).

ii. Lie-Markov models allow all cases, including BDF=2 (e.g., 5.7a above).

In this work we simulate and analyse time-reversible model variants with BDF=0, 1, 2, and 3.

Testing

We simulated length 1000 sequences on a 5 taxon 'giraffe' tree, using three time-reversible rate matrices (all randomly sampled from the same time-reversible model) applied to branches as shown below.

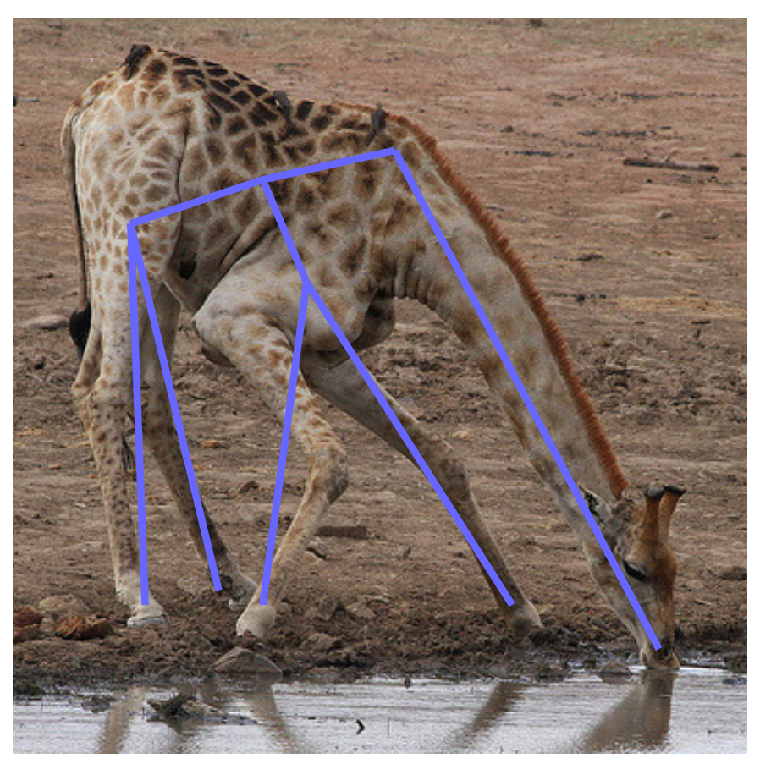
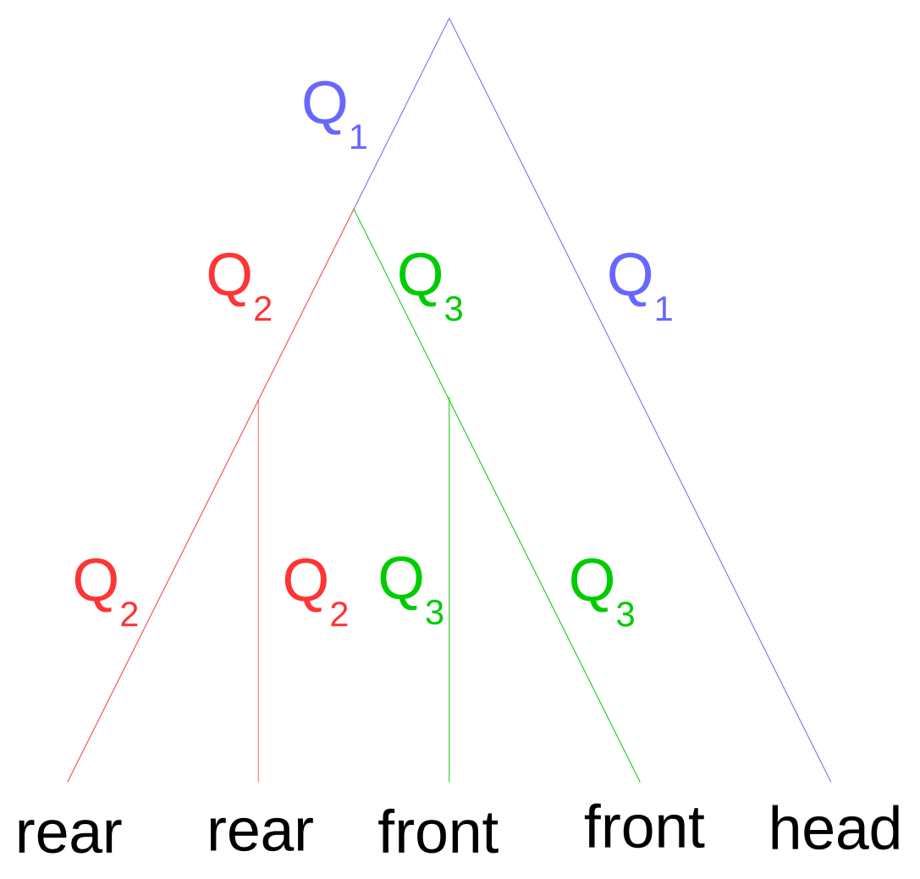


Photo credit: Derek Keats https://www.flickr.com/photos/dkeats/10383834983/ CC BY 2.0 license



The ISO standard heterogeneous giraffe

Our experimental grid is (6 simulation models) × (1000 replicates) × (29 Lie-Markov + 22 time-reversible analysis models) × (15 tree topologies).

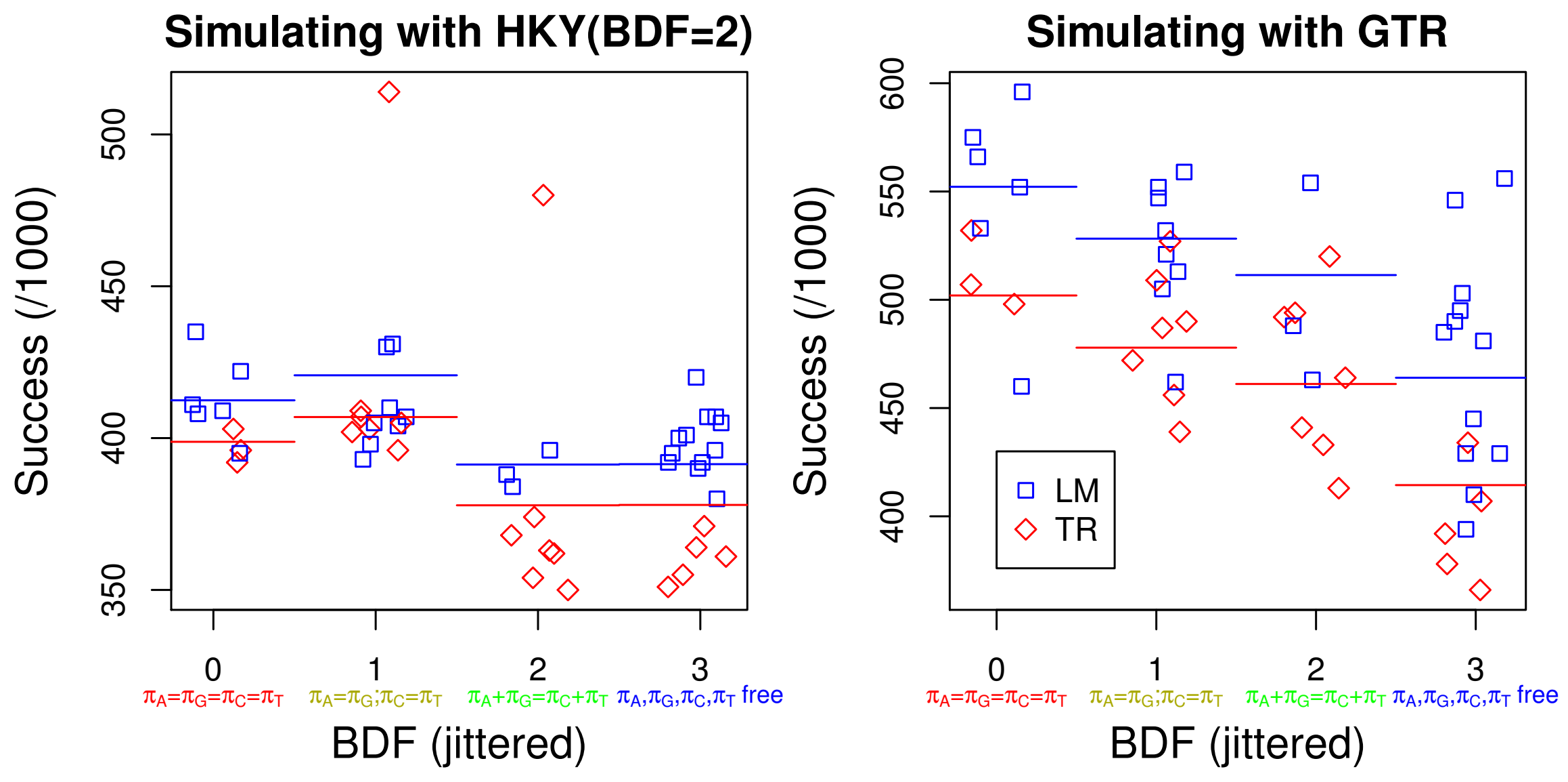
Simulation: We used time-reversible models to avoid unfairly favouring the Lie-Markov models — HKY and GTR, with BDF = 1, 2, and 3 variants. For each replicate, we randomly sampled three rate matrices within the model, and simulated an alignment of length 1000 on the illustrated 'giraffe' tree.

Analysis: We used Maximum likelihood to test time-reversible *vs* Lie-Markov models. The time-reversible models were F81, HKY, K81uf, TrN, TIM, TVM, and GTR (all BDF = 3), together with any BDF = 0, 1, or 2 variants. Some models are simultaneously time-reversible and Lie-Markov, and these were omitted, leaving 29 Lie-Markov models and 22 time-reversible models.

We analysed the alignments by maximum likelihood, using a single rate matrix, i.e., assuming homogeneity — a model mis-specification. We exhaustively searched tree space (15 topologies) and counted a 'success' when the maximum likelihood topology was the correct one.

Results

Analysis models vary by number of model parameters and BDF, which could affect their success scores. The best fit generalised linear model (optimal AIC) treated Lie-Markov *vs* time-reversible and BDF as factors, and was independent of the number of parameters in the model.



In all six test cases (simulation models), LM models outperform TR models on average.

Simulation		HKY	
model	BDF=1	BDF=2	BDF=3
p value	3×10^{-5}	3×10^{-3}	10^{-10}
Effect Size	4%	3%	6%
Simulation		GTR	
model	BDF=1	BDF=2	BDF=3
p value	5×10^{-15}	$<2\times10^{-16}$	$<2\times10^{-16}$
Effect Size	8%	10%	11%

Conclusion

When we have heterogeneous DNA mutation processes, but analyse them as if homogeneous, the Lie-Markov models are superior to time-reversible models in reconstructing the phylogeny.

Software: Beast and IQ-TREE

`github.com/MichaelWoodhams/BeastLieMarkov`

Coming soon: www.cibiv.at/software/iqtree

References

[1] J. Fernández-Sánchez, J. G. Sumner, P. D. Jarvis, and M. D. Woodhams. Lie Markov models with purine/pyrimidine symmetry. *J Math Biol*, pages 1–37, 2014.

[2] J. G. Sumner, J. Fernández-Sánchez, and P. D. Jarvis. Lie Markov models. *J Theor Biol*, 298:16–31, 2012.

[3] J. G. Sumner, P. D. Jarvis, J. Fernández-Sánchez, B. T. Kaine, M. D. Woodhams, and B. R. Holland. Is the general time-reversible model bad for phylogenetics? *Syst Biol*, 61(6):1069–1074, 2012.

[4] M. D. Woodhams, J. Fernández-Sánchez, and J. G. Sumner. A new hierarchy of phylogenetic models consistent with heterogeneous substitution rates. *Syst Biol*, 64(4):638–650, 2015.

Australian Government
Australian Research Council

This work was supported by Discovery Grant DP150100088 to JS and BH.