Polymorphism-aware phylogenetic models

MIC-Phy 2021

Dominik Schrempf

February 16, 2021

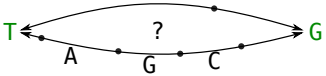
Eötvös Loránd
University

Comparative genomics and phylogenetics

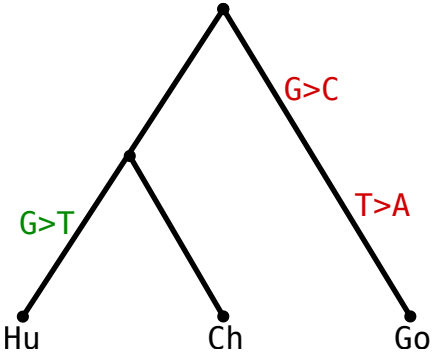
Alignment

```
Human | ACGTACGT
Chimp | ACGGACGT
Goril | ACGGACCA
```

Evolutionary model

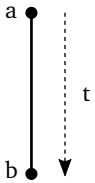


Phylogeny



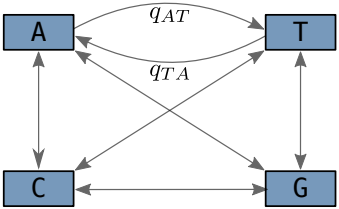
DNA substitution models

Evolution as a series of substitutions



$$a, b \in \{A, C, G, T\}$$

$$\Pr(X_t = b | X_0 = a) = \left(e^{t\boldsymbol{Q}}\right)_{ab}$$



State space $\{A, C, G, T\}$.

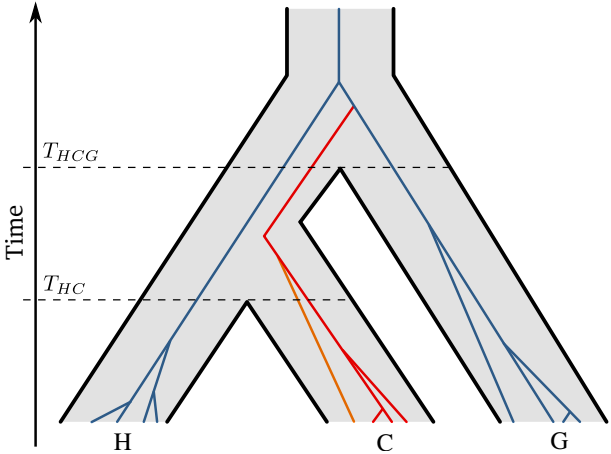
$$\boldsymbol{Q} = \begin{pmatrix} \cdot & q_{AC} & q_{AG} & q_{AT} \\ q_{CA} & \cdot & q_{CG} & q_{CT} \\ q_{GA} & q_{GC} & \cdot & q_{GT} \\ q_{TA} & q_{TC} & q_{TG} & \cdot \end{pmatrix}$$

Transition rate matrix.

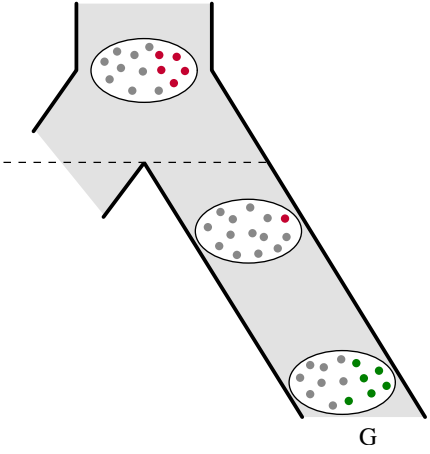
Species are populations and recombination separates histories of genes

Incomplete lineage sorting

Multispecies coalescent models[1]



Polymorphism-aware models[2]



[1] Rannala and Yang (2003).
[2] De Maio et al. (2015) and Schrempf et al. (2016).

Neutral, K-allelic Wright-Fisher[3] model

Discrete-time, discrete-state Markov chain

- N constant haploid population size.
- K alleles.
- $\boldsymbol{z}_\tau$ state of population $(z_\alpha, \ldots, z_\kappa)$ in generation τ; $\|\boldsymbol{z}_\tau\|_1 = N$; the total number of states is $\binom{N+K-1}{K-1}$.
- $\boldsymbol{U}$ $K \times K$ mutation probability matrix; the elements describe the probability to mutate from one state to another.

The distribution of alleles in the next generation $\tau + 1$ is derived by sampling with replacement from the alleles of generation τ

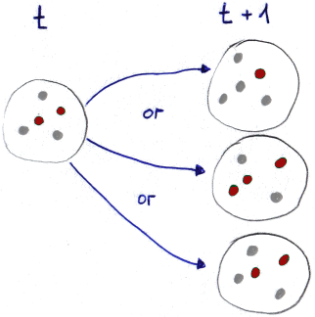
$$\boldsymbol{z}_{\tau+1} | \boldsymbol{z}_\tau \sim \text{Mult}(N, \frac{\boldsymbol{z}_\tau}{N} \boldsymbol{U}).$$

For $K = 4$, and $N = 10$, we have 286 states.

[3] Wright (1931) and Fisher (1930).

Neutral, K-allelic Moran[4] model with mutation

Continuous-time, discrete-state Markov process



Individuals are randomly chosen to reproduce. The offspring is of the same type as the parent and replaces another randomly chosen individual from the population.

For $a, b \in \{\alpha, \ldots, \kappa\}$ and with mutation rates q_{ab}, we have

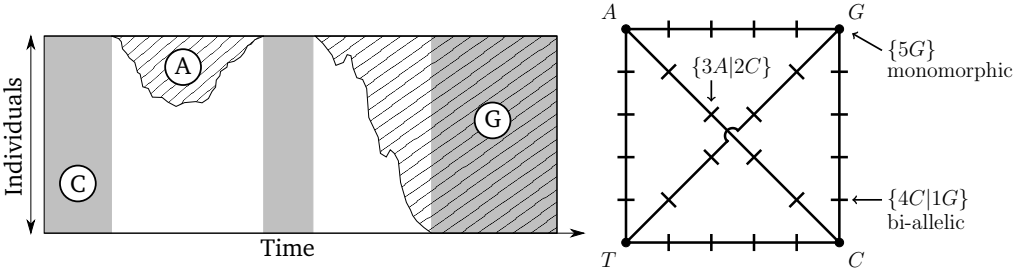
$$(\ldots, z_a, \ldots, z_b, \ldots) \to (\ldots, z_a - 1, \ldots, z_b + 1, \ldots)$$

$$\text{at rate} \quad N\left(\frac{z_b}{N}\frac{z_a}{N} + \frac{z_a}{N}\frac{q_{ab}}{N}\right).$$

[4] Moran (1958).

Approximation for low mutation rates

Drift removes variation fast; disallow mutations when the population is polymorphic



Examples using nucleotides. Cartoon of evolving population with large size and state space for $N = 5$.

Population can only be

Monomorphic $(\ldots, z_a = N, \ldots) \equiv \{Na\}$; K states.

Bi-allelic $(\ldots, z_a = i, \ldots, z_b = N - i, \ldots) \equiv \{ia|(N-i)b\}$; $\binom{K}{2}(N-1)$ states.

For $K = 4$, and $N = 10$, we have 4+54 states.

Discrete multivariate boundary mutation model[5]

From the Moran model with mutation, we have

$$(\ldots, z_a, \ldots, z_b, \ldots) \to (\ldots, z_a - 1, \ldots, z_b + 1, \ldots)$$
$$\text{at rate} \quad \frac{z_a z_b}{N} + z_a \frac{q_{ab}}{N}.$$

Transition rate matrix $\boldsymbol{M}$

Boundary mutation leads to

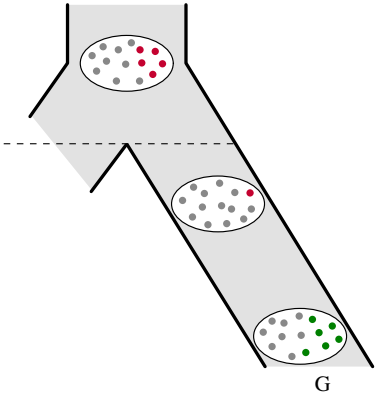
$$m_{\{Na\} \to \{(N-1)a|1b\}} = q_{ab},$$
$$m_{\{ia|(N-i)b\} \to \{(i \pm 1)a|(N-i \mp 1)b\}} = \frac{i(N-i)}{N}.$$

[5] Schrempf and Hobolth (2017).

Polymorphism-aware phylogenetic Model (PoMo)

Use discrete multivariate boundary mutation model with

- $K = 4$ nucleotides;
- *virtual* population size N;
- transition rate matrix $\boldsymbol{M}$.

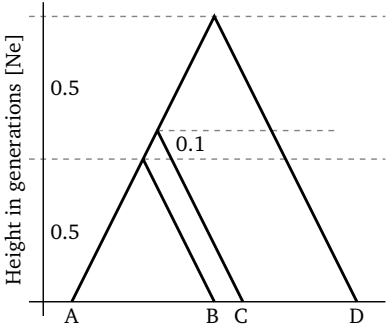


Likelihood calculation similar to DNA substitution models, for example,

$$\Pr\left(X_t = \{(N-1)a|1b\}\right) X_0 = \{Na\} = \left(e^{t\boldsymbol{M}}\right)_{\{Na\}\{(N-1)a|1b\}}.$$

Assessment of tree estimation error

Incomplete lineage sorting



Simulate

1. Up to 1000 gene trees with the multispecies coalescent model; 10 samples per species[6].
2. Sequences with 1000 base pairs per gene (HKY[7] model); $\theta = 0.025$[8].

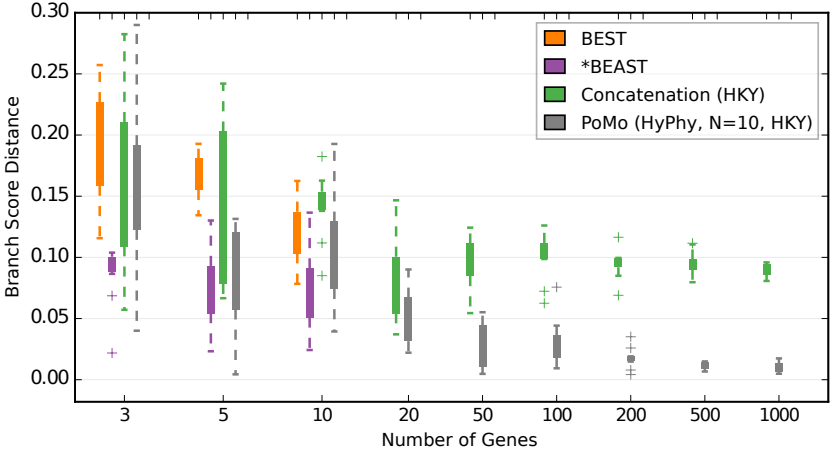
Infer phylogeny from data.

Measure branch score distance between original and estimated species tree.

[6] MSMS, Ewing and Hermisson (2010).
[7] Hasegawa et al. (1985).
[8] SeqGen, Rambaut and Grassly (1997).

Tree estimation error

Incomplete lineage sorting, $1N_e$ generations height, 10 samples per species



BEST (Liu 2008), *BEAST (Heled and Drummond 2010), and HyPhy (Pond et al. 2005).

Exchangeabilities, stationary distributions, and reversibility

Some mathematical prerequisites

The mutation rates can be separated into

$$q_{ab} = r_{ab}\pi_b,$$

where

π_a is the stationary distribution of allele frequencies, and

r_{ab} are the exchangeabilities.

If the mutation model is reversible, the exchangeabilities are symmetric $r_{ab} = r_{ba}$.

Theorem (Retention of reversibility of mutation model)

The discrete multivariate boundary mutation model is reversible if and only if the underlying mutation model is reversible.

Stationary distribution (reversible mutation model)

Theorem

For $K, N > 1$ and reversible mutation models, the discrete multivariate boundary mutation model defined by the transition rate matrix $\boldsymbol{M}$ has a stationary distribution of

$$\varphi_{\{Na\}} = \frac{1}{Z}\pi_a,$$

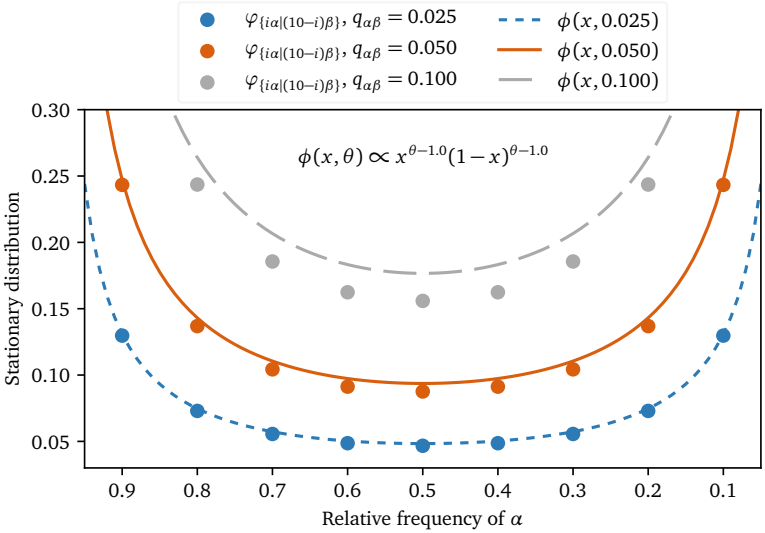
$$\varphi_{\{ia|(N-i)b\}} = \frac{1}{Z}\pi_a\pi_b r_{ab}\left(\frac{1}{i} + \frac{1}{N-i}\right)$$

with normalization constant

$$Z = 1 + \sum_{k=1}^{N-1}\frac{1}{k}\sum_{\substack{a,b\\a\neq b}} r_{ab}\pi_a\pi_b.$$

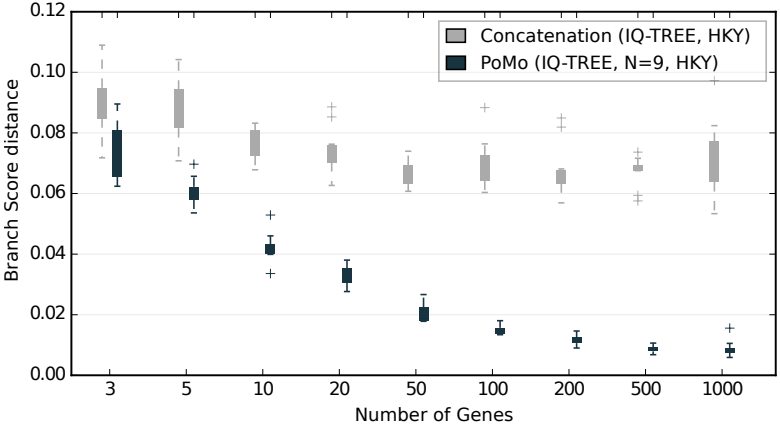
Stationary distribution

Alleles α and β ($K = 2$), $N = 10$, $q_{\alpha\beta} = q_{\beta\alpha} = \theta$; comparison to diffusion theory



Tree estimation error

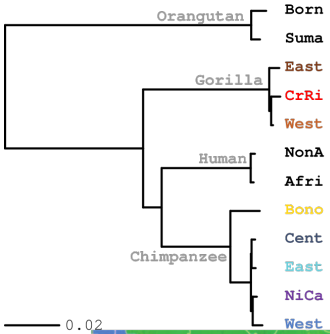
Yule[9] tree with 60 species, $3N_e$ generations height, 10 samples per species



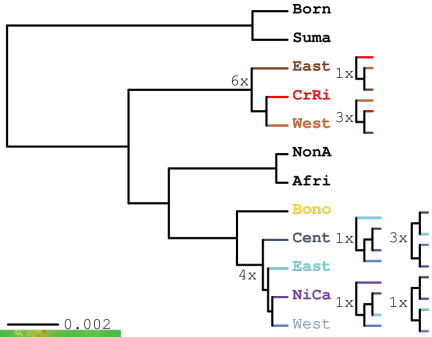
IQ-TREE (Minh et al. 2020).

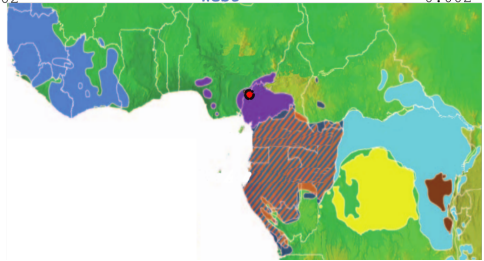
[9] Yule (1925).

PoMo IQ-TREE (HKY, N=10)



Concatenation (HKY)





Prado-Martinez et al. (2013)

Summary

Idea of PoMo

Improve phylogenetic inference by modeling the evolution of populations and not of individuals.

Discrete multivariate boundary mutation model

1. Moran model with mutations.
2. Approximation for low mutation rates.

Stationary distribution fits well if $\theta < 0.1$.

Suggestions for further reading

Non-reversible mutation models[10]

The stationary distribution is also known for non-reversible mutation models. Then, the mutation rate matrix $\boldsymbol{Q}$ can be separated into a reversible part, and a part describing circular probability flux.

Advanced mutation models[11]

Partition models

$$\overbrace{\text{ACCTTGAAGG}}^{\mathbf{Q_1}}\ \overbrace{\text{ATGTTTCTGT}}^{\mathbf{Q_2}}$$
$$\text{ACCTTCAAGG ATGTTTGTGT}$$

Mixture models

$$m_1\mathbf{Q_1} + m_2\mathbf{Q_2}$$

ACCTTGAAGG
ACCTTCAAGG

[10] Burden and Tang (2016) and Schrempf and Hobolth (2017).
[11] Schrempf et al. (2019) and Borges et al. (2019).

Acknowledgments

Carolin Kosiol

Arndt von Haeseler

Claus Vogl

Bui Quang Minh

Asger Hobolth

Nicola De Maio

Gergely Szöllősi

Nicolas Lartillot

Eötvös Loránd University

Bibliography I

- Borges, Rui, Gergely J. Szöllősi, and Carolin Kosiol (2019). "Quantifying GC-Biased Gene Conversion in Great Ape Genomes Using Polymorphism-Aware Models." In: *Genetics* 212.4, pp. 1321–1336. DOI: 10.1534/genetics.119.302074.
- Burden, Conrad J. and Yurong Tang (2016). "An approximate stationary solution for multi-allele neutral diffusion with low mutation rates." In: *Theoretical Population Biology* 112, pp. 22–32. DOI: 10.1016/j.tpb.2016.07.005.
- De Maio, Nicola, Dominik Schrempf, and Carolin Kosiol (2015). "PoMo: An Allele Frequency-Based Approach for Species Tree Estimation." In: *Systematic Biology* 64.6, pp. 1018–1031. DOI: 10.1093/sysbio/syv048.
- Ewing, Gregory and Joachim Hermisson (2010). "MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus." In: *Bioinformatics* 26.16, pp. 2064–2065. DOI: 10.1093/bioinformatics/btq322.
- Fisher, Ronald (1930). *The genetical theory of natural selection*.

Bibliography II

- Hasegawa, Masami, Hirohisa Kishino, and Taka-aki Yano (1985). "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA." In: *Journal of Molecular Evolution* 22.2, pp. 160–174. DOI: 10.1007/BF02101694.
- Heled, Joseph and Alexei J. Drummond (2010). "Bayesian Inference of Species Trees from Multilocus Data." In: *Molecular Biology and Evolution* 27.3, pp. 570–580. DOI: 10.1093/molbev/msp274.
- Liu, Liang (2008). "BEST: Bayesian estimation of species trees under the coalescent model." In: *Bioinformatics* 24.21, pp. 2542–2543. DOI: 10.1093/bioinformatics/btn484.
- Minh, Bui Quang, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear (2020). "IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era." In: *Molecular Biology and Evolution* 37.5. Ed. by Emma Teeling, pp. 1530–1534. DOI: 10.1093/molbev/msaa015.

Bibliography III

- Moran, P. a. P. (1958). "Random processes in genetics." In: *Mathematical Proceedings of the Cambridge Philosophical Society* 54.01, pp. 60–71. DOI: 10.1017/S0305004100033193.
- Pond, Sergei L. Kosakovsky, Simon D. W. Frost, and Spencer V. Muse (2005). "HyPhy: hypothesis testing using phylogenies." In: *Bioinformatics* 21.5, pp. 676–679. DOI: 10.1093/bioinformatics/bti079.
- Rambaut, Andrew and Nicholas C. Grassly (1997). "Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees." In: *Computer Applications in the Biosciences : CABIOS* 13.3, pp. 235–238. DOI: 10.1093/bioinformatics/13.3.235.
- Rannala, Bruce and Ziheng Yang (2003). "Bayes Estimation of Species Divergence Times and Ancestral Population Sizes Using DNA Sequences From Multiple Loci." In: *Genetics* 164.4, pp. 1645–1656.

Bibliography IV

- Schrempf, Dominik, Bui Quang Minh, Nicola De Maio, Arndt von Haeseler, and Carolin Kosiol (2016). "Reversible polymorphism-aware phylogenetic models and their application to tree inference." In: *Journal of Theoretical Biology* 407, pp. 362–370. DOI: 10.1016/j.jtbi.2016.07.042.
- Schrempf, Dominik and Asger Hobolth (2017). "An alternative derivation of the stationary distribution of the multivariate neutral Wright–Fisher model for low mutation rates with a view to mutation rate estimation from site frequency data." In: *Theoretical Population Biology* 114, pp. 88–94. DOI: 10.1016/j.tpb.2016.12.001.
- Schrempf, Dominik, Bui Quang Minh, Arndt von Haeseler, and Carolin Kosiol (2019). "Polymorphism-Aware Species Trees with Advanced Mutation Models, Bootstrap, and Rate Heterogeneity." In: *Molecular Biology and Evolution* 36.6. Ed. by Naruya Saitou, pp. 1294–1301. DOI: 10.1093/molbev/msz043.
- Wright, Sewall (1931). "Evolution in Mendelian Populations." In: *Genetics* 16.2, pp. 97–159.

Bibliography V

Yule, G. U. (1925). "A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 213.402-410, pp. 21–87. DOI: 10.1098/rstb.1925.0002.