Tailoring BADTRIP or PoMo for SARS-CoV-2

Graham Jones, www.indriid.com

2020-04-06, April 8, 2020

1 Introduction

THIS IS A WORK IN PROGRESS

Section 2 is my understanding of the models. Section 3 is about applying them to SARS-CoV-2 and especially speeding up the tree likelihood calculation.

Two paragraphs where I am working things out as I write are in 2.3 and 3.3, in large font.

2 The PoMo/BADTRIP models

2.1 PoMo model of evolution

In the gene pool of a population/species, at one site, there is some proportion of each of A, C, G, and T/U. When drift dominates ($N_e\mu \ll 1$, no selection) it is rare for the mixture to be far from just a single nucleotide. If is far away, it's most likely to be almost entirely a mixture of 2 nucleotides. In PoMo, this situation is modelled as a virtual population of small fixed size (say 5 to 50). If the pure nucleotides are imagined as the vertices of a tetrahedron, the mixtures of 2 nucleotides lie on the edges. PoMo models drift and mutation using the Moran model. Figure 1. See supplement to PoMo paper for full details.

2.2 PoMo model of sampling

TODO... dates, sequencing errors

2.3 BADTRIP model of evolution

BADTRIP applies the PoMo model to infectious diseases. Splits are asymmetric, with a severe population bottleneck near the start of a new infection. Figure 2. Also, limits can be placed on the infectious period for each host.

On the left of Fig 2 is how I imagined BADTRIP *would* be implemented. The tree is bifurcating, and nodes (and their branches) are labelled by their hosts, and as colonising/noncolonising. But it is not this way. On the right is closer to how it *is* implemented, but I'm unclear about details. The tree is a 'host tree' or 'transmission tree' which is multifurcating. The likelihood is the same either way, but the two options require different organisation of the likelihood calculation, different tree initialisation, different operators, different analysis of results.

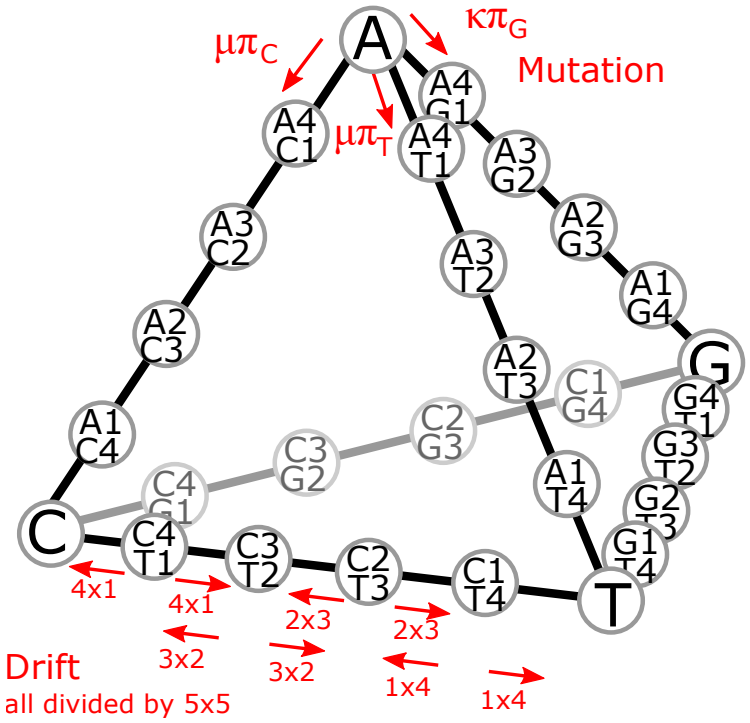

Figure 1: The PoMo model with a virtual population of 5 and the HKY substitution model. The drift model for one edge, and the mutation model for one vertex is shown.

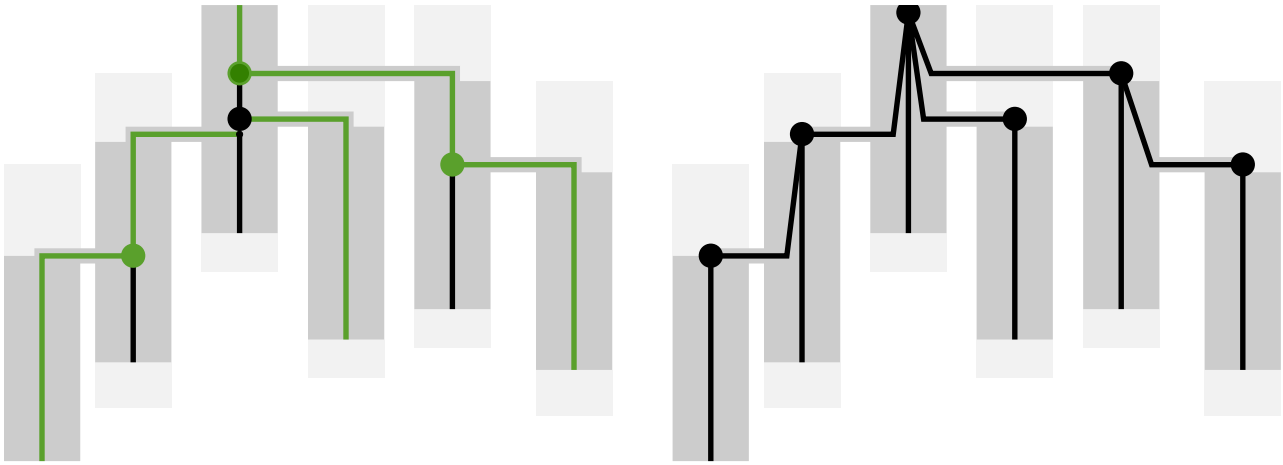

Figure 2: Mid-grey tree shows bottlenecks and within-host sections, with asymmetric splits. Pale grey rectangles indicate assumed limits on infectious periods.

2.4 BADTRIP model of sampling

TODO... dates, sequencing errors, start and end dates.

3 Applying to SARS-CoV-2

3.1 Rough picture of virus in host

This is in the spirit of the book *Cell Biology by the numbers*. If the numbers are correct within a factor of 2, that's a success.

- The mutation rate ν per site per replication is estimated as between 1e-7 and 1e-6 for coronaviruses. (Thinking Outside the Triangle:Replication Fidelity of theLargest RNA Viruses, Everett Clinton Smith, Nicole R. Sexton, and Mark R. Denison. https://www.annualreviews.org/doi/pdf/10.1146/annurev-virology-031413-085507).
- I can only find a few vague estimates for the replication time. Some bacteriophages take less than 0.5h. Some viruses take more than 24h. There's a cauliflower virus which take 21h. If I read this paper (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4947482/) correctly, influenza takes 4h-20h. Say the typical time is T hours.
- In a year, expect $365 * 24\nu/T$ mutations per site. Estimates so far are around 0.001 mutations per site per year. So $\nu/T = 0.001/365/24 \approx$1e-7. This could happen if $\nu =$1e-6 and $T = 10$, or $\nu =$5e-7 and $T = 5$.

3.2 Rough estimates of computational cost

I will do some profiling, but I think that the main time is spent doing vector-matrix multiplications. Assume a virtual population N of 10, so 64 states. Need to do 64*64 mul-and-adds per site for each within-host branch section, between colonisation time, any transmissions out of the branch, and any sampling times. And then once for each transmission bottleneck. Also need to exponentiate the rate matrix for each within-host branch section, and once for the bottlenecks (same bottleneck for all transmissions). I don't think exponentiation is a major time user when there are 30000 sites. Maybe for large N it would be.

3.3 Possible speedup: drift only, mutation only

The idea is to approximate the bottlenecks with drift only (as now in BADTRIP) and the within-host evolution as mutation only. Drift will only be important within a host while N_e is small, maybe up to a 100 or so, and this can be viewed as part of an extended bottleneck. There will be a period where $N_e\mu$ is still much less than 1, so that drift would dominate in the long term, but this period is very short, so drift won't actually do much after the extended bottleneck. Within the hosts, the most important thing happening is new mutations from the 4 pure nucleotide states. The values of μ and κ will be biased upwards a little to cover the $N_e \lesssim 100$ period.

The mutation-only rate matrix U only has 16 non-zero entries. The branch lengths are tiny for SARS-CoV-2 (a week is $\approx$ 1e-3/50 = 2e-5), so $(I + tU)$ is a good approximation to the exponential $\exp(Ut)$. It will just require 4 operations like

$$P_{A\bar{N}C1} += t\mu\pi_C P_A, \quad P_{A\bar{N}G1} += t\kappa\pi_G P_A, \quad P_{A\bar{N}T1} += t\mu\pi_T P_A$$

$$P_A -= t(\mu\pi_C + \kappa\pi_G + \mu\pi_T)P_A$$

where t is the duration and $\bar{N} - N - 1$.

The drift-only stage will then probably be the slow part, though I am not sure what other calculations are needed besides exponentiating matrices and the vector-matrix multiplications. For the drift-only rate matrix, the 4 pure nucleotide states are absorbing, and (I think) the $(6 * N - 2)^2$ matrix reduces to 6 identical matrices size $(N + 1)^2$.

I don't know how easy this is to implement. The calculations are straightforward: the most complicated thing is exponentiating the $(N + 1)^2$ matrix for drift-only. However, it is no longer a standard calculation as done by BEAST's TreeLikelihood.