Supplementary Information for *Hidden Markov Models Detect Recombination and Ancestry of SARS-CoV-2*

Nobuaki Masaki[1,2] and Trevor Bedford[2,3]

[1]Department of Biostatistics, University of Washington, Seattle, WA, [2]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center, Seattle, WA, [3]Howard Hughes Medical Institute, Seattle, WA

S1 Efficient forward algorithm

We implemented an efficient version of the forward algorithm, reducing the time complexity of the induction step from $\mathcal{O}(M^2)$ to $\mathcal{O}(M)$, where M is the number of unique Pango lineages. Using the notation from the main paper, we define,

$$\alpha_t(i) = P(O_{1:t} = k_{1:t}, Z_t = i|\lambda, \epsilon),$$

which are our forward probabilities. This represents the probability of the observed nucleotide sequence up to position t and the ancestral Pango lineage being lineage i at position t.

In the induction step, we calculate the next time step for the forward probabilities. We have,

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{M} \alpha_t(i) a_{ij}\right) b_{j,t+1}(k_{t+1}).$$

Computing $\alpha_{t+1}(j)$ for one Pango lineage j requires summing over M lineages, which costs $\mathcal{O}(M)$. Thus, computing this for all Pango lineages costs $\mathcal{O}(M^2)$.

In our transition matrix, we have equal diagonal entries and equal off-diagonal entries. Recall,

$$a_{ij} = \begin{cases} 1 - \lambda, & \text{if } i = j, \\ \frac{\lambda}{M-1}, & \text{if } i \neq j. \end{cases}$$

Furthermore, we use the scaled version of the forward probabilities, meaning that $\sum_{i=1}^{M} \alpha_t(i) = 1$. Thus, we can rewrite the induction step as,

$$\begin{aligned}
\alpha_{t+1}(j) &= \left((1 - \alpha_t(j)) \frac{\lambda}{M-1} + \alpha_t(j)(1-\lambda) \right) b_{j,t+1}(k_{t+1}) \\
&= \left(\left(1 - \lambda - \frac{\lambda}{M-1} \right) \alpha_t(j) + \frac{\lambda}{M-1} \right) b_{j,t+1}(k_{t+1}) \\
&= \left(\left(1 - \frac{M}{M-1} \lambda \right) \alpha_t(j) + \frac{\lambda}{M-1} \right) b_{j,t+1}(k_{t+1}),
\end{aligned}$$

which is constant time. Thus, computing this for all Pango lineages now costs $\mathcal{O}(M)$.