An Almost Infinite Sites Model

Mathias Christensen Cronjäger, cronjager@stats.ox.ac.uk
Alejandra Avalos-Pacheco, a.avalos-pacheco@warwick.ac.uk

Introduction

Two popular models of mutation in population-genetics are the infinite-sites and finite-sites models; outlined below. We here present a model of mutation which admits computations of likelihoods in a manner similar to the former model, while retaining the flexibility of the latter.

Finite Sites Model (FSM)
Mutations affect sites distributed uniformly on $\{1, \dots, T\}$. State of population of size n encoded by $S \in \{\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T}\}^{n \times T}$.

+ Can encode any set of sequences; easy to communicate.
- Computing likelihoods scales abysmally; even for "simple" datasets.

Infinite Sites Model (ISM)
Mutations affect sites distributed uniformly on $[0, 1]$. For a countable number of mutations, we only need to track if a site is segregating or not. State of population encoded by $S \in \{0, 1\}^{n \times L}$, where $L \leq T$ denotes the number of segregating sites

+ Compact representation of data when $L \ll T$.
+ Each step will almost surely lead to a lower-rank-state.[a]
+ Small state-space ⇒ traversing state-space easier ⇒ recursive computation of likelihoods more feasible.
- Restrictive modeling assumptions (violated in practice).
- Unable to account for sites with >2 observed nucleotides.

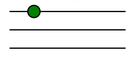
Our **Almost Infinite Sites Model** bridges the gap between these two models of mutation, by applying FSM-analogues of the simplifying assumptions of the ISM–notably the indistinguishability of sites.By imposing an upper bound, $\mathbf{b} \geq L$, on the total number of mutations in a genealogical history, the number of possible ancestral histories of a population may be kept finite, and likelihoods may be recursively computed.

[a] $\mathrm{rank}(S)$ is the sum of the number of distinct haplotypes in S and the number of segregating sites in S.

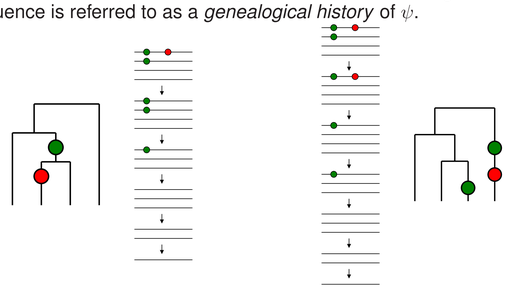
Encoding Sequences

Since populations and mutations are presumed to be exchangeable, it is worth noting that S_1 and S_2 encode the same state (denoted $S_1 \sim S_2$) if and only if permutation-matrices P_r, P_c exist satisfying $S_1 = P_r S_2 P_c$. Hence we distinguish between *encodings* of states S (not unique), and *states* $\psi = [S]_\sim$ (unique).

$$\begin{bmatrix} \mathtt{G} & \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{A} \\ \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{A} \\ \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{A} \\ \mathtt{G} & \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{T} \end{bmatrix} \in \left[\begin{bmatrix} \mathtt{A} & \mathtt{G} & \mathtt{A} & \mathtt{T} & \mathtt{A} & \mathtt{A} \\ \mathtt{A} & \mathtt{G} & \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{A} \\ \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{A} \\ \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{A} & \mathtt{A} \end{bmatrix} \right]_\sim =$$



A sequence of states is referred to as *admissible* if it can be constructed from a coalescent process undergoing mutation by removing events "from the bottom up".
An admissible sequence starting from ψ and terminating in a single sequence is referred to as a *genealogical history* of ψ.



Two distinct genealogical histories of the same state (along with coalescents demonstrating admissibility). Note that the right genealogical history is not admissible under the ISM.

Recursion – Computing Likelihoods

By summing the probabilities of all genealogical histories of a dataset, we obtain its associated likelihood. Using the Markovian structure of the coalescent, we may express the likelihood recursively.

$$\mathbb{P}\left(\text{ }\right) = \sum_{\substack{\Psi \text{ potential} \\ \text{history of } \text{ }}} \mathbb{P}(\Psi) \qquad , \qquad \mathbb{P}(\psi) = \sum_{\psi' : \mathbb{P}(\psi' \to \psi) > 0} \mathbb{P}(\psi') \mathbb{P}(\psi' \to \psi)$$

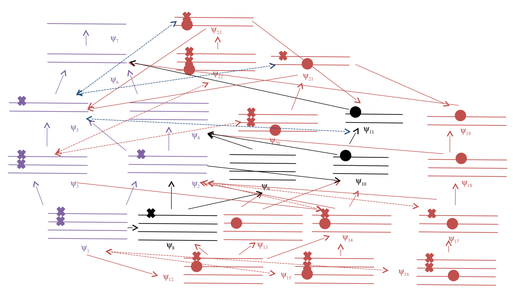
We aim to determine $\mathbb{P}(\psi, B \leq b)$; the joint probability of observing sample configuration $\psi = [S]_\sim = (A, \mathbf{n}^{\mathrm{r}}, \mathbf{n}^{\mathrm{c}})$ and the number of mutations in the genealogical history of ψ (denoted by $\mathbf{B}$) being bounded from above by $\mathbf{b} > 0$. To this end we make the following assumptions:

- Mergers in the genealogical history of ψ are governed by an underlying Kingman-coalescent process.
- Mutation-events occur at rate $\frac{\theta}{2}$, and affect a site uniformly chosen from $\{1 \dots T\}$

Taking a "forward-backward" approach similar to Griffiths (1989), we obtain the following recurrence:

$$\mathbb{P}((A, \mathbf{n}^{\mathrm{r}}, \mathbf{n}^{\mathrm{c}}), B \leq b) = \frac{N-1}{N-1+\theta} \sum_{i : \mathbf{n}^{\mathrm{r}}_i > 1} \frac{n_i - 1}{N-1} \mathbb{P}((A, \mathbf{n}^{\mathrm{r}} - e_i, \mathbf{n}^{\mathrm{c}}), B \leq b)$$

$$+ \frac{\theta}{N-1+\theta} \sum_{\substack{1 \leq i \leq l \\ 1 \leq j \leq L \\ \mathtt{x} \in \{\mathtt{A},\mathtt{G},\mathtt{T},\mathtt{C}\}}} \frac{\mathcal{N}^{\mathrm{r}}_{i,j,\mathtt{x}}(A, \mathbf{n}^{\mathrm{r}})}{N} \frac{\mathcal{N}^{\mathrm{c}}_{i,j,\mathtt{x}}(A, \mathbf{n}^{\mathrm{c}})}{T} P_{\mathtt{x}, a_{i,j}} \mathbb{P}(\mathcal{M}_{i,j,\mathtt{x}}(A, \mathbf{n}^{\mathrm{r}}, \mathbf{n}^{\mathrm{c}}), B \leq b-1) \qquad (1)$$



Ancestral Configurations with upper bounds $\mathbf{b} = 2, 3, 4$ (purple, black and red respectively) for a simple dataset.

Algorithm – Solving the recursion (1)

Although the recursion (1) can be used directly for computing $\mathbb{P}(\psi, \mathbf{B} \leq \mathbf{b})$ via naïve tail-recursion, this is highly inefficient. We explore the following approach based on storing all values already computed.

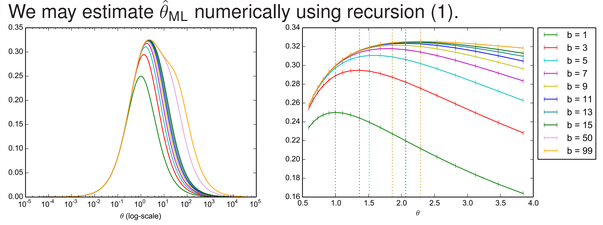
Input S, b, and model-specific constants for (1)
Output $\mathbb{P}([S]_\sim, \mathbf{B} \leq \mathbf{b})$
Initialise empty hash table $\mathbf{H}$
if $S = [0]$ (No segregating sites; 1 active lineage)
  Return 1
Else If $\exists S' \sim S : (S', \mathbf{b}) \in \mathrm{Keys}(\mathbf{H})$ [a]
  Look up $\mathbb{P}([S]_\sim, \mathbf{B} \leq \mathbf{b}) = \mathbf{H}(S', \mathbf{b})$.
  Return $\mathbb{P}([S]_\sim, \mathbf{B} \leq \mathbf{b})$
Else if $\mathbf{b} \geq b_{\min}([S]_\sim)$
  Compute $\mathbb{P}([S]_\sim, \mathbf{B} \leq \mathbf{b})$ using recursion in formula (1)
  Add key-value pair $((S, \mathbf{b}), \mathbb{P}([S]_\sim, \mathbf{B} \leq \mathbf{b}))$ to $\mathbf{H}$
Else
  Return 0
End

The algorithm has been implemented in Python; source, code is available at: `https://github.com/Cronjaeger/almost-infinite-sites-recursions`

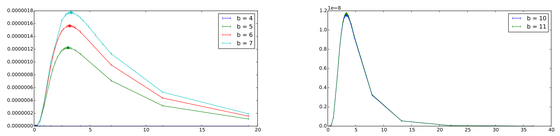
[a] Checking this condition quickly is non-trivial. In our implementation it is done by applying a $\sim$-invariant hash-function and resolving collisions by checking if suitable $P_{\mathrm{r}}, P_{\mathrm{c}}$ can be constructed using a backtracking algorithm.

Results & Benchmarking

We may estimate $\hat{\theta}_{\mathrm{ML}}$ numerically using recursion (1).

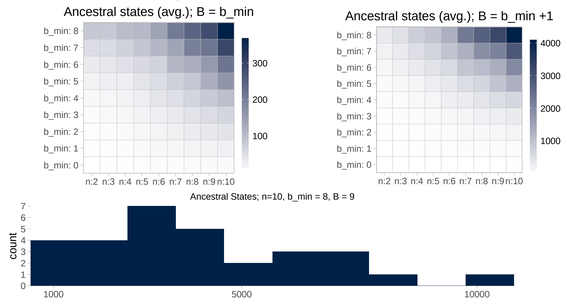


Likelihoods for a very simple dataset $S = \left[\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right]$ and accompanying MLE-estimates for varying b.



The number of non-segregating sites impacts the mass gap significantly: on the left 50% of sites are segregating; on the right 1%.

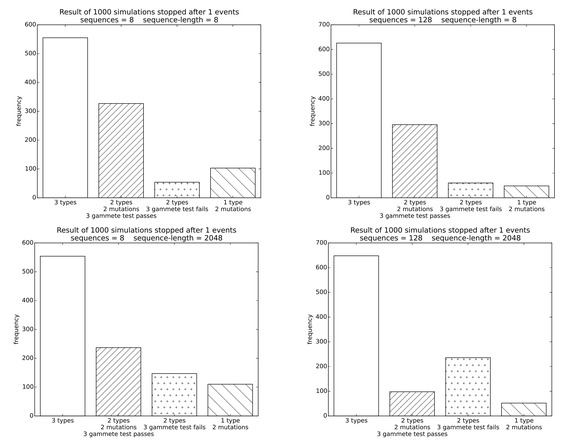
Some benchmarking-results for simulated data:



Simulation – How is the infinite sites assumption violated?

We 1000 times simulate a Kingman-coalescent, and add mutations under the finite sites hypothesis, until one of the following events have occurred twice:

1. A site with > 2 nucleotides occurs.
2. A site with 2 nucleotides has been affected by > 2 mutations.
3. An *incompatibility* has ocurred, eg. $S = \left[\begin{smallmatrix} \mathtt{A} & \mathtt{T} & \mathtt{T} & \mathtt{A} \\ \mathtt{A} & \mathtt{T} & \mathtt{A} & \mathtt{T} \end{smallmatrix}\right]^T$
4. Two mutations "cancel out"



Code for simulation and plot-generation code available at: `https://github.com/Cronjaeger/coalescent-simulations`.

Future Work and Interesting Open Questions

- This work was mainly motivated by wanting to analyse viral sequence-data for sampled regions with sufficiently high mutation-rates that the ISA might be violated at a significant number of sites. This still remains to be done.
- Upper bounds the on run-time of the suggested algorithm remain unknown. Even for the case when b is equal to the number of segregating sites (recursion reduces to that of Griffiths '89), no literature could be found. The main factor of interest is the total number of states considered, given an initial input ψ.
- Being able to give bounds on $\mathbb{P}(\mathbf{B} \leq \mathbf{b} | \psi)$ would allow a more informed way of picking $\mathbf{b}$ when applying our algorithm.
- Verifying if
$$\exists S' \sim S : (S', \mathbf{b}) \in \mathrm{Keys}(\mathbf{H})$$
holds is currently slow.
- A range of combinatorial and asymptotic questions regarding the set of all configurations remain open; solving them might bring us closer to understanding the nature and limitations of the algorithm.
  1. Given ψ, how many distinct genealogical histories of ψ with up to $\mathbf{b}$ mutation-events exist?
  2. Given ψ, how many distinct states ψ' exist, which occur in at least one such genealogical history?
  3. What are the maxima and/or maximizers of the above counts taken over the set of all ψ with $\mathrm{rank}(\psi) < R$?
  4. Can asymptotic bounds on the above quantities be obtained?

  It is worth noting that answering questions 1. and 2. cosresponds to counting paths and nodes respectively in a graph such as the one depicted on this poster.

References

Árnason, E., 2004. Mitochondrial Cytochrome b DNA Variation in the High-Fecundity Atlantic Cod: Trans-Atlantic Clines and Shallow Gene Genealogy. Genetics, 166(4), pp.1871–1885.
Dialdestoro, K. et al., 2016. Coalescent Inference Using Serially Sampled, High-Throughput Sequencing Data from Intrahost HIV Infection. Genetics, 202(4), pp.1449–72.
Ethier, S.N. & Griffiths, R.C., 1987. The Infinitely-Many-Sites Model as a Measure-Valued Diffusion. The Annals of Probability, 15(2), pp.515–545.
Griffiths, R.C., 1987. Counting genealogical trees. Journal of Mathematical Biology, 25(4), pp.423–431.
Griffiths, R.C., 1989. Genealogical-tree probabilities in the infinitely-many-site model. Journal of Mathematical Biology, (27), pp.667–680.
Hobolth, A. & Wiuf, C., 2009. The genealogy, site frequency spectrum and ages of two nested mutant alleles. Theoretical population biology, 75(4), pp.260–5.
Jenkins, P.A. & Song, Y.S., 2011. The effect of recurrent mutation on the frequency spectrum of a segregating site and the age of an allele. Theoretical Population Biology, 80(2), pp.158–173.

Aknowledgements

The work outlined here grew out if a 10 Week joint mini-project by Mr. Cronjäger and Ms. Avalos-Pacheco, who are both students in the Oxford-Warwick Statistical Programme; a Centre for Doctoral Training in Next Generational Statistical Science supported by the Engineering and Physical Sciences Research Council and the Medical Research Council. Ms. Avalos Pacheco is furthermore supported by the Mexican National Council of Science and Technology (CONACYT). The project was supervised by Dr. Jotun Hein (Oxford) and Dr. Paul Jenkins (Warwick), who acting as supervisors both contributed substantially.