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 by $S_i$ ($i = 1, 2$) if and only if permutation-matrices $P_1$, $P_2$ exist satisfying $S_i = P_i S_j P_i^{-1}$. Hence we distinguish between encodings of states $S$ (not unique), and states $S_i$ ($i = 1, 2$, unique).

A sequence of states is regarded as admissible if it can be constructed from a coalescent process under mutation by removing events "from the bottom up". An admissible sequence starting from $\psi$ and terminating in a single state is referred to as a genealogical history of $\psi$.

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

ALGORITHM – SOLVING THE RECURRENCE (1)

Input $S$, and model-specific constants for (1)
Output $P(S[S_i], b) = P_i$

Start

Initialise empty hash table $H$
If $S = \emptyset$ (No segregating sites; 1 active lineage)
Else if $\exists \psi \sim S : (\mathcal{S}, b) \in \text{Keev}(H)$
Lookup $P(S[S_i], b) \leq b) \Rightarrow H(\mathcal{S}, b)$.
Return $P(S[S_i], b) \leq b$
Else if $\exists \psi \sim S$ (No segregating sites; 1 active lineage)
Compute $P(S[S_i], b) \leq b)$ using recursion in formula (1)
Add key-value pair $(\mathcal{S}, b), (P(S[S_i], b) \leq b))$ to $H$
Else
Return 0
End

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

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 occurred, $\psi \sim |S|$$^2$.
4. Two mutations "cancel out"

RESULTS & BENCHMARKING

We may estimate $L$ numerically using recursion (1).

Likelihoods for a very simple dataset $S = \{1\}$ 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%.

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

REFERENCES


ACKNOWLEDGEMENTS

The work outlined here grew out if a 10 Week joint mini-project by Mr. Cronjaeger 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.