Mutation models used in PRESUME

Substitutions

In PRESUME, the substitution probabilities at different positions in each sequence are defined in a time-dependent manner using GTR-Gamma model or set to a certain rate as follows.

1. GTR-Gamma model (executed by `--gtrgamma`)

This model is commonly used in evolutionary biology to describe sequence diversification by modeling of heterogeneity in substitution rates across different sequence positions.

When the GTR-Gamma model is chosen, a position-specific relative substitution parameter γ_i for every different position i is first determined by a gamma distribution with user-defined parameters μ and α as follows:

$$\gamma_i \sim \Gamma(\alpha, \mu/\alpha).$$

Let $P_i(\Delta t)$ be a 4 × 4 matrix, where $P_i(\Delta t)_{x \to y}$ is the transition probability from a certain source nucleotide x to a destination nucleotide y ($x, y \in$ {A, C, G, T}) within the time interval Δt. $P_i(\Delta t)$ is then defined using γ_i and the substitution rate matrix Q, as follows:

$$P_i(\Delta t) = e^{\gamma_i \Delta t Q},$$

where Q is given by

$$\mathrm{Q} = \begin{pmatrix} - & a_{\mathrm{A}\to\mathrm{C}} & a_{\mathrm{A}\to\mathrm{G}} & a_{\mathrm{A}\to\mathrm{T}} \\ a_{\mathrm{C}\to\mathrm{A}} & - & a_{\mathrm{C}\to\mathrm{G}} & a_{\mathrm{C}\to\mathrm{T}} \\ a_{\mathrm{G}\to\mathrm{A}} & a_{\mathrm{G}\to\mathrm{C}} & - & a_{\mathrm{G}\to\mathrm{T}} \\ a_{\mathrm{T}\to\mathrm{A}} & a_{\mathrm{T}\to\mathrm{C}} & a_{\mathrm{T}\to\mathrm{G}} & - \end{pmatrix} \begin{pmatrix} \pi_{\mathrm{A}} & 0 & 0 & 0 \\ 0 & \pi_{\mathrm{C}} & 0 & 0 \\ 0 & 0 & \pi_{\mathrm{G}} & 0 \\ 0 & 0 & 0 & \pi_{\mathrm{T}} \end{pmatrix}.$$

The sum of the diagonal values of the right-side matrix of Q must be 1 ($\pi_{\mathrm{A}} + \pi_{\mathrm{C}} + \pi_{\mathrm{G}} + \pi_{\mathrm{T}} = 1$), and the left-side matrix needs to be symmetrical (i.e., the same values are assigned to the symmetrical nucleotide transition patterns), wherein the diagonal missing values are given to satisfy the condition that every row sum of Q becomes 0. Here, $a_{x \to y}$ and π_x are user-defined parameters.

In PRESUME, the GTR-Gamma model is executed by `--gtrgamma` with the following format to specify the parameters mentioned above:

$$\mathrm{GTR}\{\mathrm{a_{A,C}}/\mathrm{a_{A,G}}/\mathrm{a_{A,T}}/\mathrm{a_{C,G}}/\mathrm{a_{C,T}}/\mathrm{a_{G,T}}\}+\mathrm{FU}\{\pi_{\mathrm{A}}/\pi_{\mathrm{C}}/\pi_{\mathrm{G}}/\pi_{\mathrm{T}}\}+\mathrm{G}\{\alpha\}$$

2. Time-independent model (executed by `--editprofile`)

Users can define time-independent substitutions per branch (or generation) using the substitution probability matrix $\Phi_{sub,i}$ for every sequence position which original position in the root sequence is i as follows:

$$\Phi_{sub,i} = \begin{pmatrix} - & \Phi_{\mathrm{i,A\to C}} & \Phi_{\mathrm{i,A\to G}} & \Phi_{\mathrm{i,A\to T}} \\ \Phi_{\mathrm{i,C\to A}} & - & \Phi_{\mathrm{i,C\to G}} & \Phi_{\mathrm{i,C\to T}} \\ \Phi_{\mathrm{i,G\to A}} & \Phi_{\mathrm{i,G\to C}} & - & \Phi_{\mathrm{i,G\to T}} \\ \Phi_{\mathrm{i,T\to A}} & \Phi_{\mathrm{i,T\to C}} & \Phi_{\mathrm{i,T\to G}} & - \end{pmatrix},$$

where all elements in the matrix are non-negative, and the diagonal missing values are given to satisfy the condition that every row sum becomes 1.

Indels

The time-independent indel simulation is as follows. Sequences are propagated along the template propagating unit (PU) tree and progressively accumulate indels at each generation. Let *i'* be the nucleotide position of a sequence at a given branch of the PU tree, where its corresponding position in the root sequence is *i*. Insertion events in each generation (branch) are modeled by insertion probability parameters $\Phi_{ins,i}$, representing the chance of having an insertion event at the 3' side of nucleotide position *i'* and a position-independent size probability distribution $\psi_{ins}(L)$. The sequence of the insert length L at every event is randomly generated with equal probability for each nucleotide at every position. Deletion events are modeled by deletion probability parameters $\Phi_{del,i}$ and a position-independent size probability distribution $\psi_{del}(L)$, where a deletion of size L centering at a nucleotide position *i'* spanning from position $(i' - \lfloor L/2 \rfloor)$ through $(i' - \lfloor L/2 \rfloor + L - 1)$ occurs with probability $\Phi_{del,i}$, where L is stochastically determined according to $\psi_{del}(L)$. Note that $\Phi_{ins,i}$ and $\Phi_{del,i}$ are defined only for nucleotide positions *i* of the root sequence, reflecting that genome editing occurs only for predefined targeting sequences in cell lineage tracing.