Nucleotide substitution models used in PRESUME

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

Let γ be a relative substitution rate at each sequence position i that follows a gamma distribution whose shape and mean are defined by α and μ

$$\gamma \sim \mathrm{Gamma}(\alpha, \alpha/\mu)\ (\alpha : \mathrm{shape}, \alpha/\mu\text{: rate})$$

Let $\mathrm{P(t)}$ be a 4×4 matrix, in which $\mathrm{P(t)}_{\mathrm{x,y}}$ is the transition probability from a certain source nucleotide x to a destination nucleotide $\mathrm{y}\ (\mathrm{x}, \mathrm{y} \in \{\mathrm{A, C, G, T}\})$ within the time interval t. In GTR-Gamma model, $\mathrm{P}(t)$ of sequence position i is defined using the relative substitution rate γ and a constant matrix of Q

$$\mathrm{P(t)} = \mathrm{e}^{\mathrm{t}\gamma\mathrm{Q}}$$

In this formula, Q is the substitution rate matrix

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

where sum of the diagonal values of the right-side matrix are required to be 1 ($\pi_{\mathrm{A}} + \pi_{\mathrm{C}} + \pi_{\mathrm{G}} + \pi_{\mathrm{T}} = 1$), and the left-side matrix are required to be symmetric (i.e. same element values are assigned to symmetric nucleotide transition patterns) whose diagonal missing values fulfill that every row sum of Q becomes 0.

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

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

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

PRESUME allows user to use a time-independent model where $\mathrm{P(t)}$ is set as follows

$$\mathrm{P}(t) = \Phi = \begin{pmatrix} 1-\phi & \phi/3 & \phi/3 & \phi/3 \\ \phi/3 & 1-\phi & \phi/3 & \phi/3 \\ \phi/3 & \phi/3 & 1-\phi & \phi/3 \\ \phi/3 & \phi/3 & \phi/3 & 1-\phi \end{pmatrix}$$

where $\delta \in [0,1]$

In PRESUME, the time-independent model is executed by `--constant` with the single parameter ϕ.