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

$$\gamma \sim Gamma(\alpha, \alpha/\mu)\ (\alpha : shape, \alpha/\mu : rate)$$

Let $P(t)$ be a 4×4 matrix, in which $P(t)_{x,y}$ 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 t. In GTR-Gamma model, $P(t)$ of sequence position i is defined using the relative substitution rate γ and a constant matrix of Q

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

In this formula, Q is the substitution rate matrix

$$Q = \begin{pmatrix} - & a_{A\to C} & a_{A\to G} & a_{A\to T} \\ a_{C\to A} & - & a_{C\to G} & a_{C\to T} \\ a_{G\to A} & a_{G\to C} & - & a_{G\to T} \\ a_{T\to A} & a_{T\to C} & a_{T\to G} & - \end{pmatrix} \begin{pmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_G & 0 \\ 0 & 0 & 0 & \pi_T \end{pmatrix}$$

where sum of the diagonal values of the right-side matrix are required to be 1 ($\pi_A + \pi_C + \pi_G + \pi_T = 1$), 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:

$$\text{GTR}\{a_{A,C}/a_{A,G}/a_{A,T}/a_{C,G}/a_{C,T}/a_{G,T}\}\text{+FU}\{\pi_A/\pi_C/\pi_G/\pi_T\}\text{+G}\{\alpha\}$$

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

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

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

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

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