Model details

Here, the stochastic model on the sequence propagation and the mutaition / substitution in PRESUME is described.

- Model on propagation

  The propagation of sequences was modeled to describe cell division/death during development, and speciation/extinction during evolution.

  - Parameters
    - e : deletion probability of a sequence
    - σ : standard deviation of propagation rate of children
  - Model
    - Propagation

      Sequences propagate with their own rate. Suppose a mother sequence (M) and its two daughter sequences (D_1, D_2), we define d_M, d_{D1}, and d_{D2} as the doubling time of the M, D_1, and D_2. Here, the propagation rate of D_i ($i = 1, 2$), $1/d_{Di}$ independently follows the normal distribution of which the mean is $1/d_M$ and the variance is σ^2. That is,

      $1/d_{Di} \sim Norm(1/d_M, \sigma^2)$ ($i = 1, 2$)

      This simulates that the propagation rate is inherited from a mother to daughters, strongly when σ is small, and weakly when σ is large.

    - Death/extinction

      The death/extinction of a cell/species was also modeled. Di is deleted if $1/d_{Di} < 0$ ($i = 1, 2$), which means sequences whose propagation rate is too small (correspondsto too low compatibility) dies before the next doubling. In addition, Di is randomly deleted at probability of e ($0 < e < 1$), which describes the accidental death of the sequence.

- Model on mutation/substitution
  - Here, we just write "substitution" for describing mutation or substitution.
  - GTR-GAMMA model (option: —model)

    This model is a commonly used in evolutionary biology to model accumulating substitution in sequences. This model consists of modeling site heterogeneity and time-dependent subsitution rate.

    - Parameters
      - g : shape of the gamma distribution which mutation rate of every site follows
      - m : mean of the gamma distribution which mutation rate of every site follows
      - parameters of Q : There are 9 parameters for defining Q
    - Site heterogeneiry

In GTR-GAMMA model, the rate of substitution, γ, varies among sites of each sequence, which independently follows the gamma distribution of which shape parameter is g and the mean is m. That is,

$\gamma \sim Gamma(g, g/m)$ (g : shape, g/m : rate)

In other words, the mutation rate of most sites are distributed around m when g is high, while mutation rates of most site are around 0 except only small fraction of sites with exceptionally high mutation rate when g is small.

- Time dependent substitution

  Let $P(t)$ be a 4×4 matrix which means the transition probability from a certain character x to y ($x, y \in \{A, C, G, T\}$) during the time interval of t (ex. When $d_M = t$, the probability that an A in M becomes C in D_1 is $P(t)_{AC}$.).

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

  Here, Q means the substitution rate matrix which has 9 parameters. γ is the substitution rate of each site in a sequence stated above.

  $$Q = \begin{pmatrix} - & a_1 & a_2 & a_3 \\ a_1 & - & a_4 & a_5 \\ a_2 & a_4 & - & a_6 \\ a_3 & a_5 & a_6 & - \end{pmatrix} \begin{pmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_G & 0 \\ 0 & 0 & 0 & \pi_T \end{pmatrix}$$

  Diagonal values of left matrix are determined to fulfill that sum of each row in Q is 0. Also, $\pi_A + \pi_C + \pi_G + \pi_T = 1$.

- Model option format (—model, -m)

  When you specify the parameters of GTRGAMMA model please type as following:

  Example:

  PRESUME.py -m m --model GTR{a_1/a_2/a_3/a_4/a_5/a_6}+FU{π_A/π_C/π_G/π_T}+G{g}

- Time-independent model (option: —delta)

  Time-independent model is much simpler than the GTR-GAMMA model. Let $\delta \in [0, 1]$ be the substitution probability, $P(t)$ explained above is calculated in this way:

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

  That is, $P(t)$ does not depend on t.

  When you specify the parameters of time-independent model please type as following:

  Example:

  PRESUME.py —delta δ