Likelihood computation of a phylogenetic tree for dummies

Models of nucleotide substitution

Instantaneous rate matrix $\boldsymbol{Q}$, in which each element $\boldsymbol{Q}_{ij}$ represents the rate of change from base i to base j during some infinitesimal time period dt. The most general form of this matrix is

$$\boldsymbol{Q} = \begin{bmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu g\pi_A & -\mu(g\pi_A + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu h\pi_A & \mu i\pi_C & -\mu(h\pi_A + i\pi_C + f\pi_T) & \mu f\pi_T \\ \mu j\pi_A & \mu k\pi_C & \mu l\pi_G & -\mu(j\pi_A + k\pi_C + l\pi_G) \end{bmatrix}$$

Where μ represents the mean instantaneous substitution rate. This mean rate is modified by the relative rate parameters $a, b, c, \dots, l$, which correspond to each possible transformation from one base to a different base.

In time-reversible model, the overall rate of change from base i to base j in a given length of time is the same as the rate from base j to base i. it's case for all following models, in which $g = a, h = b, i = c, j = d, k = e\,, l = f$.

The mean instantaneous substitution rate can then be computed in this way

$$\mu = \frac{1}{\sum_{i \neq j}^{A,C,T,G} \pi_i \boldsymbol{Q'}_{ij}}$$

Where

$$\boldsymbol{Q'} = \begin{bmatrix} - & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & - & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & - & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & - \end{bmatrix}$$

$\boldsymbol{Q}$ can be decomposed into two matrices $\boldsymbol{R}$ (rate matrix) and $\boldsymbol{\Pi}$ (equilibrium frequencies) :

$$\boldsymbol{Q} = \boldsymbol{R} \times \boldsymbol{\Pi}$$

$$\boldsymbol{R} = \begin{bmatrix} - & \mu a & \mu b & \mu c \\ \mu g & - & \mu d & \mu e \\ \mu h & \mu i & - & \mu f \\ \mu j & \mu k & \mu l & - \end{bmatrix}$$

$$\boldsymbol{\Pi} = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_G & 0 \\ 0 & 0 & 0 & \pi_T \end{bmatrix}$$

$\boldsymbol{P}(t)$ is the transition probability matrix, giving the probabilities of change from any state to any other along a branch of length t. It is calculated as

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

Rate heterogeneity can be incorporated into likelihood analyses by including an additional relative rate component, r, into the transition probability expression. Then, different evolution rates can be assigned for different subsets of the sequences (categories). The relative rate r scaled so that the mean substitution rate remains 1, branch lengths will still reflect the number of substitutions per site. The most commonly used continuous distribution for modeling rate heterogeneity is the gamma distribution. The gamma distribution has 2 parameters, a shape parameter α and a scale parameter β. By setting $\beta = {}^{1}/_{\alpha}$, a distribution with a mean rate of 1 is obtained, and a wide variety of rate distributions can be obtained by varying α. We use k categories to approximate the gamma distribution, with equal probability $1/k$ in each category. We can calculate the percentage point (the cutting point) of the Γ distribution (with a given parameter α, and $\beta = {}^{1}/_{\alpha}$) for category c of the gamma distribution as follows

$$z_{\Gamma}\left(\frac{c}{k}, \alpha\right) = \frac{z_{\chi^2}\left(\frac{c}{k}, 2\alpha\right)}{2\alpha}$$

Where $z_{\chi^2}(p, v)$ is the percentage point of the χ^2 distribution with v degree of freedom, which can be calculated by, say, the algorithm of Best and Roberts (1975). Note that $z_{\Gamma}(0, \alpha) = 0$.

Then the rate of a category c (c going from 0 to $k-1$) is computed this way

$$r(c) = \frac{I\left(\alpha + 1, \alpha \times z_{\Gamma}\left(\frac{c+1}{k}, \alpha\right)\right) - I\left(\alpha + 1, \alpha \times z_{\Gamma}\left(\frac{c}{k}, \alpha\right)\right)}{{}^{1}/_{k}}$$

Matrices for Jukes Cantor model

Simplest model, where all base frequencies are equal ($\pi_A = \pi_C = \pi_G = \pi_T = 0.25$) and all substitutions occur at the same rate ($a = b = c = d = e = f = 1$).

$$\boldsymbol{Q} = \begin{bmatrix} -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu & \frac{1}{4}\mu & -\frac{3}{4}\mu \end{bmatrix}$$

$$P_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-\mu tr} & (i = j) \\ \frac{1}{4} - \frac{1}{4}e^{-\mu tr} & (i \neq j) \end{cases}$$

Matrices for Kimura's 2 parameters model

K2P model takes into account the common observation that transitions and transversions occur at different rates, but still assume equal base frequencies ($\pi_A = \pi_C = \pi_G = \pi_T = 0.25$). Thus we set $a = c = d = f = 1$ and $b = e = \kappa$.

$$\boldsymbol{Q} = \begin{bmatrix} -\frac{3}{4}\mu(\kappa+2) & \frac{1}{4}\mu & \frac{1}{4}\mu\kappa & \frac{1}{4}\mu \\ \frac{1}{4}\mu & -\frac{3}{4}\mu(\kappa+2) & \frac{1}{4}\mu & \frac{1}{4}\mu\kappa \\ \frac{1}{4}\mu\kappa & \frac{1}{4}\mu & -\frac{3}{4}\mu(\kappa+2) & \frac{1}{4}\mu \\ \frac{1}{4}\mu & \frac{1}{4}\mu\kappa & \frac{1}{4}\mu & -\frac{3}{4}\mu(\kappa+2) \end{bmatrix}$$

$$P_{ij}(t) = \begin{cases} \dfrac{1}{4} + \dfrac{1}{4}e^{-\mu tr} + \dfrac{1}{2}e^{-\mu tr\left(\frac{\kappa+1}{2}\right)} & (i = j) \\ \dfrac{1}{4} + \dfrac{1}{4}e^{-\mu tr} - \dfrac{1}{2}e^{-\mu tr\left(\frac{\kappa+1}{2}\right)} & (i \neq j, transition) \\ \dfrac{1}{4} - \dfrac{1}{4}e^{-\mu tr} & (i \neq j, transversion) \end{cases}$$

Matrices for Hasegawa-Kishino-Yano 1985 model

HKY85 is a generalization of K2P model allowing unequal equilibrium base frequencies.

$$\boldsymbol{Q} = \begin{bmatrix} -\mu(\pi_C + \kappa\pi_G + \pi_T) & \mu\pi_C & \mu\kappa\pi_G & \mu\pi_T \\ \mu\pi_A & -\mu(\pi_A + \pi_G + \kappa\pi_T) & \mu d\pi_G & \mu\kappa\pi_T \\ \mu\kappa\pi_A & \mu\pi_C & -\mu(\kappa\pi_A + \pi_C + \pi_T) & \mu\pi_T \\ \mu\pi_A & \mu\kappa\pi_C & \mu\pi_G & -\mu(\pi_A + \kappa\pi_C + \pi_G) \end{bmatrix}$$

$$P_{ij}(t) = \begin{cases} \pi_j + \pi_j\left(\dfrac{1}{\Pi_j} - 1\right)e^{-\mu tr} + \left(\dfrac{\Pi_j - \pi_j}{\Pi_j}\right)e^{-\mu tr\left(1+\Pi_j(\kappa-1)\right)} & (i = j) \\ \pi_j + \pi_j\left(\dfrac{1}{\Pi_j} - 1\right)e^{-\mu tr} - \left(\dfrac{\pi_j}{\Pi_j}\right)e^{-\mu tr\left(1+\Pi_j(\kappa-1)\right)} & (i \neq j, transition) \\ \pi_j(1 - e^{-\mu tr}) & (i \neq j, transversion) \end{cases}$$

With $\Pi_j = \pi_A + \pi_G$ is base j is a purine (A or G) and $\Pi_j = \pi_C + \pi_T$ if base j is a pyrimidine (C or T).

Matrices for General Time Reversible model

GTR is the most general model that is time reversible.

$$\boldsymbol{Q} = \begin{bmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu a\pi_A & -\mu(a\pi_A + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu b\pi_A & \mu d\pi_C & -\mu(b\pi_A + d\pi_C + f\pi_T) & \mu f\pi_T \\ \mu c\pi_A & \mu e\pi_C & \mu f\pi_G & -\mu(c\pi_A + e\pi_C + f\pi_G) \end{bmatrix}$$

Substitution probabilities can be calculated by numerical evaluation of the eigenvalues and eigenvectors of $\boldsymbol{Q}$ using standard algorithms.

$$\boldsymbol{P}(t) = \Omega \times \Phi(\mathrm{t}) \times \Omega^{-1}$$

Where Ω is the matrix containing the right eigenvectors of $\boldsymbol{Q}$, and

$$\Phi_{\mathrm{ij}}(\mathrm{t}) = \begin{cases} e^{tr\Psi_{ii}} & (\mathrm{i} = \mathrm{j}) \\ 0 & (i \neq j) \end{cases}$$

Where Ψ is the diagonal matrix containing the eigenvalues of $\boldsymbol{Q}$.

Likelihood computation

The method for evaluating the likelihood of a given tree proceeds from a hypothetical root node at any convenient location in the tree, and combines the likelihoods of each of its daughter trees. As we work with time reversible models, the choice of the root location will not change the likelihood of the tree. We will sum the likelihood of each individual site, and as likelihood of a site can be a very small number, it is more convenient to use the negative log likelihood. As we can also define a proportion of invariant sites p_i (proportion of the sites that cannot change), it leads to this calcultation

$$ML(tree) = -\sum_{s=1}^{nSites} \ln\big((1-p_i)\Lambda_{\mathrm{V}}(\mathrm{s}) + p_i\Lambda_{\mathrm{I}}(\mathrm{s})\big)$$

Where $nSites$ is the number of sites and

$$\Lambda_I(s) = \begin{cases} \pi_i & if\ s\ is\ an\ invariant\ site\ with\ state\ i \\ 0 & if\ s\ is\ not\ an\ invariant\ site \end{cases}$$

$$\Lambda_V(s) = \sum_{i}^{A,C,G,T} \frac{\sum_{c=0}^{k} L_c(x_{Rs} = i)}{k} \pi_i$$

Where k is the total number of category for rate heterogeneity, and R is the root node.
$L_c(x_{Rs} = i)$ is the conditional likelihood of state i at site s for category c at the root node, detailed below.

If node A is an ancestor that gave rise to sequences B and C, then the conditional likelihood of state i at site s for category c in A is

$$L_c(x_{As} = i) = \left[\sum_{j}^{A,C,T,G} P_{ij}\left(\frac{v_{AB}}{1-p_i}\right).L_c(x_{Bs} = j)\right] \times \left[\sum_{k}^{A,C,T,G} P_{ik}\left(\frac{v_{AC}}{1-p_i}\right).L_c(x_{Cs} = k)\right]$$

Where v_{xy} is the length of the branch joining sequence x to sequence y.