Appendix 1: Derivation of the Markov process for a gap-augmented rate matrix

Here we present the differential equations that define a conditional matrix for a gap-augmented rate matrix \( R^g \). We also describe a particular quasi-reversible analytical solution for the extended conditional probabilities which applies for any arbitrary underlying reversible substitutions-only rate matrix \( R \).

For an alphabet of \( K \) residues, those probabilistic models are defined by a \( K \times K \) rate matrix \( R \) such that the matrix of conditional probabilities \( Q_t(i, j) \equiv P(j|i, t) \) is given by

\[
Q_t = e^{tR} = \sum_{l=0}^{\infty} \frac{(tR)^l}{l!}.
\]

(32)

A well behaved rate matrix \( R \) has eigenvalues of the form \( \epsilon_0 = 0 \) and \( -\epsilon_a < 0 \), for \( 1 \leq a \leq A \) (where \( A \) is the number of non-degenerate eigenvalues other than zero), which is at most \( K - 1 \). Assuming that the zero eigenvalue is not degenerate, the model has a saturation behavior given by

\[
\lim_{t \to \infty} Q_t = \left( \begin{array}{cccc} \pi_1 & \cdots & \pi_K \\ \vdots & \ddots & \vdots \\ \pi_1 & \cdots & \pi_K \end{array} \right) = \left( \begin{array}{c} 1 \\ \vdots \\ 1 \end{array} \right) \left( \begin{array}{c} \pi_1 \\ \vdots \\ \pi_K \end{array} \right) = u \pi^T,
\]

where the distribution \( \pi \) are called the equilibrium (or saturation) probabilities, and \( u^T = (1, \ldots, 1) \) is the unitary vector. In addition, if the rate \( R \) is reversible \([\pi_i R(i, j) = \pi_j R(j, i)]\), the conditional probabilities also satisfy the equivalent requirement of time reversibility,

\[
\pi_i Q_t(i, j) = \pi_j Q_t(j, i), \quad \forall t.
\]

(34)

For every diagonalizable rate matrix \( R \), there is a similarity transformation such that for an invertible matrix \( U \) [86]

\[
R = U R_{\text{diag}} U^{-1}.
\]

(35)

Using the similarity transformation, one can cast \( Q_t \) into the form,

\[
Q_t = U e^{t R_{\text{diag}}} U^{-1}.
\]

(36)

For every eigenvalue of \( R \), we can construct a \( K \times K \) matrix \( O_a \) (for \( 0 \leq a \leq A \)), such that if the eigenvalue is at position \((k_a, k_a)\) of \( R_{\text{diag}} \), then

\[
O_a(i, j) = U(i, k_a) U^{-1}(k_a, j).
\]

(37)

Using the \( O_a \) matrices, one can rewrite,

\[
R = \sum_{a=1}^{A} e_a O_a,
\]

(38)
\[ Q_t = O_0 + \sum_{a=1}^{A} O_a e^{-\epsilon_a t}. \] (39)

Since \( O_0 \) controls the behavior at time infinity given by Eq. (33), then one has \( O_0(i, j) = \pi_j \). Thus, \( Q_t \) can be cast into the form,

\[ Q_t = u\pi^T + \sum_{a=1}^{A} O_a e^{-\epsilon_a t}. \] (40)

The advantage of this representation of \( Q_t \) over the one given in terms of the matrix \( U \) in Eq. (36) is that the \( O_a \) matrices, unlike the similarity matrix \( U \), are independent of the particular coordinate system used for \( R_{\text{diag}} \).

The \( O_a \) matrices consist of \( K \times K \) real numbers that satisfy the conditions

\[ \sum_{a=1}^{A} O_a + u\pi^T = U U^{-1} = I, \] (41)

\[ RO_a = -\epsilon_a O_a. \] (42)

In addition, if \( R \) is reversible they also satisfy the conditions

\[ \pi_i O_a(i, j) = \pi_j O_a(j, i). \] (43)

For a general rate matrix \( R \), the eigenvalues \( \{-\epsilon_a\} \) have to be calculated numerically, as well as the \( \{O_a\} \) matrices using Eqs. (41,42). In some simple cases, both the eigenvalues and the matrices \( \{O_a\} \) can be calculated analytically.

The extended \( (K + 1) \times (K + 1) \) rate matrix \( R^e \) considered in this work depends on a rate of deletion \( \mu \geq 0 \) and a rate of insertion \( \lambda \geq 0 \) and is given by

\[
R^e = \begin{pmatrix}
R - \mu I & \mu \\
\vdots & \mu \\
\lambda p_1 & \ldots & \lambda p_K & -\lambda
\end{pmatrix},
\] (44)

where \( I \) is the \( K \times K \) identity matrix, and \( p \) is the distribution of inserted residues (\( \sum_{i=1}^{K} p_i = 1 \)). This extended rate matrix \( R^e \) has the property that each row adds up to zero. The extended conditional probabilities are defined by

\[ Q^e_t \equiv e^{t R^e} = \sum_{l=0}^{\infty} \frac{(t R^e)^l}{l!}. \] (45)
Generally, the extended conditionals $Q^ε_t$ can be cast to the form,

$$Q^ε_t = \begin{pmatrix} M_t \\ \vdots \\ ξ^ε_t \end{pmatrix},$$

(46)

with the conditions $σ_t + \sum_{j=1}^K ξ^j_t = 1$ and $γ_t + \sum_{j=1}^K M_t(i, j) = 1$, for each row $i$. The analytic expression for $Q^ε_t$ can be obtained by solving the linear differential equation $d\left(e^{tR^ε}\right)/dt = R^ε e^{tR^ε}$, which corresponds to

\[
\begin{align*}
\dot{M}_t &= (R - μI)M_t + μu ξ^T_t, \\
\dot{γ}_t &= μ (σ_t - γ_t), \\
\dot{σ}_t &= λ (γ_t - σ_t), \\
\dot{ξ}^i_t &= λ (p^T M_t)^i - λξ^i_t,
\end{align*}
\]

(47)

(48)

(49)

(50)

where $ξ^T_t = (ξ^1_t, \ldots, ξ^K_t)$, and $(p^T M_t)^i = \sum_{j=1}^K p_j M_t(j, i)$. With the initialization conditions at $t = 0$

$$M_0 = I, \quad γ_0 = 0, \quad σ_0 = 1, \quad ξ^i_0 = 0.$$  

(51)

The linear differential equations for $γ_t$ and $σ_t$ are independent of the underlying substitution process and given by,

$$\begin{align*}
γ_t &= \frac{μ}{λ + μ} \left(1 - e^{-(λ + μ)t}\right), \\
σ_t &= \frac{1}{λ + μ} \left(μ + λe^{-(λ + μ)t}\right).
\end{align*}$$

(52)

(53)

We will not attempt to solve the general case analytically, but rather the restricted case in which the underlying substitution process is reversible, and where the distribution of insertions introduced in Eq. (44) is the same as the equilibrium distribution of the substitution rate matrix (i.e. $p = π$). In that case, the reversibility condition for $R$ together with the property $π_i R^ε(i, -)R^ε(-, j) = π_j R^ε(j, -)R^ε(-, i) = π_i π_j λμ$ guarantee that reversibility is maintained in the substitution subspace and therefore $(π^T M_t)^i = π_i (1 - γ_t)$. In this particular case, the equation for $ξ^i_t$ becomes

$$\dot{ξ}^i_t + λξ^i_t = λπ_i (1 - γ_t),$$

(54)

which has the solution $ξ^i_t = π_i ξ_t$ for

$$ξ_t = 1 - σ_t = \frac{λ}{λ + μ} \left(1 - e^{-(λ + μ)t}\right).$$

(55)

We call this model quasi-reversible since for a reversible rate matrix $R$, the extended model maintains
reversibility in the substitution subspace, that is, \( \pi_i M_t(i,j) = \pi_j M_t(j,i) \).

A key observation regarding obtaining the conditional probabilities associated to substitutions in the presence of insertions and substitutions \((M_t)\) is to notice that if the eigenvalues of \( R \) are \( \{0, -(e_a + \mu), -(\lambda + \mu)\} \). One can see that the conditional substitutions that satisfy the linear non-homogeneous equation (47) are given by

\[
M_t = u \pi^T \frac{\lambda}{\lambda + \mu} + \sum_a O_a e^{-(e_a + \mu)t} + u \pi^T \frac{\mu}{\lambda + \mu} e^{-(\lambda + \mu)t},
\]

provided that either the rate of insertions or the rate of deletion is positive. In the particular case \( \lambda = \mu = 0 \), then

\[
M_t = u \pi^T + \sum_a O_a e^{-e_a t}.
\]

In summary, for any reversible \( K \times K \) substitution rate matrix \( R \) and using the stationary distribution of residues for \( R \) as the distribution of inserted residues, the solution for the extended conditional probabilities is quasi-reversible and given by

\[
Q^\varepsilon_t = \begin{pmatrix}
\xi_t & \cdots & \xi_t & 1 - \xi_t \\
M_t & \cdots & M_t & \xi_t \\
\vdots & \ddots & \vdots & \vdots \\
\xi_t & \cdots & \xi_t & \xi_t \\
\end{pmatrix},
\]

where the functions \( \gamma_t, \xi_t, \) and the \( K \times K \) matrix \( M_t \) are given by Eqs. (52), (55), and (56) respectively.

For instance, for the Jukes-Cantor model [7] with substitution rate \( R(i \neq j) = \alpha \), there is only one non-zero eigenvalue, \( e_1 = 4\alpha \), and the stationary probabilities are \( \pi_i = \frac{1}{4} \). The matrix \( O_1 \) is given by \( O_1(i,j) = \delta_{ij} - \frac{1}{4} \), where \( \delta_{ij} \) is one if \( i \) and \( j \) are equal or zero otherwise. Introducing those values in Eq. (56), one obtains for the extended Jukes-Cantor model

\[
M_t^{JC}(i,j) = \frac{1}{4} \frac{\lambda}{\lambda + \mu} + \left( \delta_{ij} - \frac{1}{4} \right) e^{-(4\alpha + \mu)t} + \frac{1}{4} \frac{\mu}{\lambda + \mu} e^{-(\lambda + \mu)t},
\]

such that for \( \lambda = \mu = 0 \) is given by the original Jukes-Cantor model

\[
Q_t^{JC}(i,j) = \frac{1}{4} + \left( \delta_{ij} - \frac{1}{4} \right) e^{-4\alpha t}.
\]

The asymptotic behavior for the quasi-reversible solution is as follows: for infinitesimally small times \( \tau \) we have,

\[
M_\tau \simeq I + \tau (R - \mu I),
\]

\[
\gamma_\tau \simeq \mu \tau,
\]

\[
\xi_\tau \simeq \lambda \tau,
\]

\[
\sigma_\tau \simeq 1 - \lambda \tau,
\]
which is the expected behavior since $\lambda$ and $\mu$ have been introduced as the rate of insertions and deletions respectively. The asymptotic behavior at time infinity is given by,

$$Q^\varepsilon_\infty = \begin{pmatrix}
\frac{\lambda}{\lambda+\mu} \pi_1 & \cdots & \frac{\lambda}{\lambda+\mu} \pi_K & \frac{\mu}{\lambda+\mu} \\
\vdots & \ddots & \vdots & \vdots \\
\frac{\lambda}{\lambda+\mu} \pi_1 & \cdots & \frac{\lambda}{\lambda+\mu} \pi_K & \frac{\mu}{\lambda+\mu} \\
\frac{\lambda}{\lambda+\mu} \pi_1 & \cdots & \frac{\lambda}{\lambda+\mu} \pi_K & \frac{\mu}{\lambda+\mu}
\end{pmatrix}.$$  

(65)

Thus, the saturation (or equilibrium) distribution for the gap-augmented rate matrix $R^\varepsilon$ is given by

$$\pi_i^\varepsilon = \begin{cases} 
\frac{\lambda}{\lambda+\mu} \pi_i & \text{if } i \leq K, \\
\frac{\mu}{\lambda+\mu} & \text{if } i = -. 
\end{cases}$$  

(66)

For more general nonreversible substitution rate matrices, the extended conditional probabilities can always be calculated numerically using the Taylor expansion in (45).