J. Appl. Prob. **46,** 497–506 (2009)
Printed in England
© *Applied Probability Trust* 2009

SIMILAR STATES IN CONTINUOUS-TIME MARKOV CHAINS

V. B. YAP,* *National University of Singapore*

Abstract

In a homogeneous continuous-time Markov chain on a finite state space, two states that jump to every other state with the same rate are called similar. By partitioning states into similarity classes, the algebraic derivation of the transition matrix can be simplified, using hidden holding times and lumped Markov chains. When the rate matrix is reversible, the transition matrix is explicitly related in an intuitive way to that of the lumped chain. The theory provides a unified derivation for a whole range of useful DNA base substitution models, and a number of amino acid substitution models.

Keywords: Holding time; uniformisation; lumped Markov chain; reversibility

2000 Mathematics Subject Classification: Primary 60J27; 60J35
Secondary 92D20

1. Introduction

In some applications of homogeneous continuous-time Markov chains on a finite state space, it is important to calculate the transition matrix across the interval $[0, t]$, $\boldsymbol{P}(t)$, from the transition rate matrix $\boldsymbol{Q}$. In molecular evolution, the substitution patterns observed in DNA sequences is often modelled as a Markov chain on the DNA bases {T,C,A,G}. Such models are used routinely to answer important questions in evolutionary biology, such as estimating evolution rates and estimating gene distances, which are both important for reconstructing phylogenies. Early works on molecular evolution focused on gene distances between pairs of DNA sequences or amino acid sequences, based on models with explicit algebraic expressions for $\boldsymbol{P}(t)$. For example, the Kimura distance [5, p. 198], [6, pp. 76–77] is based on the Kimura model [12], with α_1 and α_2 both positive,

$$\boldsymbol{Q} = \begin{bmatrix} -\alpha_1 - 2\alpha_2 & \alpha_1 & \alpha_2 & \alpha_2 \\ \alpha_1 & -\alpha_1 - 2\alpha_2 & \alpha_2 & \alpha_2 \\ \alpha_2 & \alpha_2 & -\alpha_1 - 2\alpha_2 & \alpha_1 \\ \alpha_2 & \alpha_2 & \alpha_1 & -\alpha_1 - 2\alpha_2 \end{bmatrix},$$

$$\boldsymbol{P}(t) = \begin{bmatrix} 1 - r(t) - 2s(t) & r(t) & s(t) & s(t) \\ r(t) & 1 - r(t) - 2s(t) & s(t) & s(t) \\ s(t) & s(t) & 1 - r(t) - 2s(t) & r(t) \\ s(t) & s(t) & r(t) & 1 - r(t) - 2s(t) \end{bmatrix},$$

where $r(t) = (1 + \mathrm{e}^{-4\alpha_2 t} - 2\,\mathrm{e}^{-2(\alpha_1+\alpha_2)t})/4$ and $s(t) = (1 - \mathrm{e}^{-4\alpha_2 t})/4$. This can be derived from the Kolmogorov equations, $\boldsymbol{P}'(t) = \boldsymbol{Q}\boldsymbol{P}(t)$ or $\boldsymbol{P}'(t) = \boldsymbol{P}(t)\boldsymbol{Q}$. In general, the expression

Received 24 January 2008; revision received 28 April 2009.

* Postal address: Department of Statistics and Applied Probability, National University of Singapore, Blk S16 Level 7, 6 Science Drive 2, Singapore 117546. Email address: stayapvb@nus.edu.sg

https://doi.org/10.1239/jap/1245676102 Published online by Cambridge University Press

for an arbitrary four-state process can be found by solving up to twelve differential equations. Most DNA substitution models impose constraints on $\boldsymbol{Q}$, but the manipulation can still be quite complex. See, for example, the eight-parameter model SSL98 [15], where the algebraic expression for $\boldsymbol{P}(t)$ was used to compute the gradient of the likelihood function of a model, based on multiple DNA sequences related by some assumed phylogeny.

The present paper outlines a strategy for obtaining algebraic expressions for $\boldsymbol{P}(t)$ which enables the results to be written down immediately in the simplest cases, such as the Kimura model. In more complex cases there can still be a substantial reduction in the analytical effort. The method depends on there being some similarity in the states. It turns out that such constraints in $\boldsymbol{Q}$ can be revealed in $\boldsymbol{P}(t)$ through the use of hidden holding times, a natural extension of the uniformisation technique, where unobservable transitions from a state to itself are allowed.

The paper is organised as follows. In Section 2, hidden holding times and similar states are introduced, leading to a representation of $\boldsymbol{P}(t)$ in terms of a hidden transition matrix. Then the connection to lumped Markov chains is explained, followed by applications to two substitution models on DNA bases and amino acids. In Section 3 we present the reversible case, where there is an elegant relationship between the transition matrix and the lumped chain. This yields a unified derivation of $\boldsymbol{P}(t)$ in a range of DNA and amino acid substitution models.

2. Markov chains

Consider a homogeneous Markov chain $\{X(t)\}_{t\geq 0}$ on a finite state space $S = \{1, \dots, s\}$ defined by a transition rate matrix $\boldsymbol{Q}$ which has strictly positive off-diagonal entries and diagonal entries such that each row sums to 0. It has a unique equilibrium distribution $\boldsymbol{\pi}$ which is strictly positive. The transition probability over a time interval of length t is $\boldsymbol{P}(t) = \exp(\boldsymbol{Q}t)$. For any $i, j \in S$ and $u, t \geq 0$, $\boldsymbol{P}_{ij}(t) = \mathrm{P}\{X(u+t) = j \mid X(u) = i\}$. The chain can be characterised by a discrete-time Markov chain $\{Y_n\}_{n\geq 0}$ and holding times $\{T_n\}_{n\geq 0}$ [14, pp. 87–90], such that the transition matrix of the jump chain is $\boldsymbol{Q}$ with diagonal elements set to 0, normalised, and, conditional on $(Y_0, \dots, Y_{n-1})$, $(T_1, \dots, T_n)$ are independent exponential random variables with rates $(-\boldsymbol{Q}_{Y_0 Y_0}, \dots, -\boldsymbol{Q}_{Y_{n-1}Y_{n-1}})$. We call this a (Y, T) description of the Markov chain.

We extend the above formulation as follows. For $i = 1, \dots, s$, let $\lambda_i \geq -\boldsymbol{Q}_{ii}$, so that $\boldsymbol{P} = \boldsymbol{I} + \mathrm{diag}(\lambda_1^{-1}, \dots, \lambda_s^{-1})\boldsymbol{Q}$ is a stochastic matrix. Define a process with a discrete-time Markov chain $\{Z_n\}_{n\geq 0}$ with transition matrix $\boldsymbol{P}$, and holding times $\{\tau_n\}_{n\geq 0}$ such that, conditional on $(Z_0, \dots, Z_{n-1})$, $(\tau_1, \dots, \tau_n)$ are independent exponential variables with rates $(\lambda_{Z_0}, \dots, \lambda_{Z_{n-1}})$. In general, the τs are hidden, due to possible self-transitions. This process is equivalent to the original Markov chain, because (a) the associated jump process has the same distribution as Y; (b) conditional on the first n jumps, the first n holding times are sums of independent random numbers of independent hidden holding times and, therefore, are independent; and (c) the holding time between Y_{n-1} and Y_n is the sum of N independent exponential variables of rate $\lambda_{Y_{n-1}}$, where N has a geometric distribution with success probability $-\boldsymbol{Q}_{Y_{n-1}Y_{n-1}}/\lambda_{Y_{n-1}}$, so the nth holding time is exponential with rate $-\boldsymbol{Q}_{Y_{n-1}Y_{n-1}}$ [3, p. 54]. There are two special cases: (I) if each $\lambda_i = -\boldsymbol{Q}_{ii}$ then (Z, τ) is identical to the jump chain (Y, T); (II) if the λ_i are equal then the τs correspond to a homogeneous Poisson process (this is uniformisation [10, pp. 20–21]). Let $\rho_i = \lambda_i + \boldsymbol{Q}_{ii} = \lambda_i - \sum_{j\neq i} \boldsymbol{Q}_{ij}$ be the hidden self-transition rate. The (Z, τ) description is valid for any choice of $\rho_i \geq 0$.

For $t \geq 0$, we define a hidden transition matrix $\boldsymbol{W}(t)$ as follows. Start the process at $X(0) = i$, so that τ_1 has rate λ_i. For an event E, we write $\mathrm{P}\{E \mid X(0) = i\}$ as $\mathrm{P}_i\{E\}$. Set

$$\boldsymbol{W}_{ij}(t) = \mathrm{P}_i\{\tau_1 < t,\ X(t) = j\}. \tag{1}$$

https://doi.org/10.1239/jap/1245676102 Published online by Cambridge University Press

Since $T_1 \geq \tau_1$ with probability 1, if $\tau_1 > t$ then $X(t) = i$. Hence,

$$\begin{aligned}\boldsymbol{P}_{ij}(t) &= \mathrm{P}_i\{\tau_1 > t, X(t) = j\} + \boldsymbol{W}_{ij}(t)\\ &= \mathrm{P}_i\{\tau_1 > t\}\delta_{ij} + \boldsymbol{W}_{ij}(t)\\ &= \mathrm{e}^{-\lambda_i t}\delta_{ij} + \boldsymbol{W}_{ij}(t)\end{aligned}$$

gives the fundamental representation

$$\boldsymbol{P}(t) = \mathrm{diag}(\mathrm{e}^{-\lambda_1 t}, \ldots, \mathrm{e}^{-\lambda_s t}) + \boldsymbol{W}(t).$$

2.1. Similar states

Now we define similar states in a Markov chain.

Definition 1. Two distinct states i_1 and i_2 are similar if their rates to every other state agree, i.e. $\boldsymbol{Q}_{i_1 j} = \boldsymbol{Q}_{i_2 j}$ whenever $j \neq i_1, i_2$. A nonempty subset of S is a similarity class if every pair of states in the set are similar.

For example, if $S = \{1, 2\}$ then

$$\boldsymbol{Q} = \begin{bmatrix} -\alpha_2 & \alpha_2 \\ \alpha_1 & -\alpha_1 \end{bmatrix} \tag{2}$$

and the states are formally similar. In the following rate matrix on $\{1, 2, 3\}$, 1 and 2 are similar, but not any other pair of states, unless $\alpha_2 = \alpha_5$ or $\alpha_1 = \alpha_4$:

$$\boldsymbol{Q} = \begin{bmatrix} -\alpha_2 - \alpha_3 & \alpha_2 & \alpha_3 \\ \alpha_1 & -\alpha_1 - \alpha_3 & \alpha_3 \\ \alpha_4 & \alpha_5 & -\alpha_4 - \alpha_5 \end{bmatrix}. \tag{3}$$

The relation of similarity is not transitive. For example, if $\alpha_4 = \alpha_1$ but $\alpha_5 \neq \alpha_2$ in (3), then 1 is similar to 2, 2 is similar to 3, but 1 is not similar to 3. This point is general. Suppose that i and j are similar, and that j and k are similar. If $l \neq i, j, k$ then $\boldsymbol{Q}_{il} = \boldsymbol{Q}_{jl} = \boldsymbol{Q}_{kl}$. So the similarity of i and k hinges on whether $\boldsymbol{Q}_{ij} = \boldsymbol{Q}_{kj}$. Although similarity is not an equivalence relation, it is always possible to partition S into similarity classes, although there may not be a unique maximal partition having the least number of classes. In the above example, there are two maximal partitions. It will be shown later that, under reversibility, similarity is an equivalence relation.

The (Z, τ) description of the process gives us the freedom to choose ρ_i to exploit similar states in a very natural way. For example, let $\rho_i = \alpha_i$ in (2), so that $\lambda_i = \alpha_1 + \alpha_2$. Then $\tau_1 \mid \{X(0) = 1\}$ and $\tau_1 \mid \{X(0) = 2\}$ are both exponential with rate $\alpha_1 + \alpha_2$, $\mathrm{P}_i\{X(\tau_1) = j\} = \alpha_j/(\alpha_1 + \alpha_2)$, and it follows from (1) that the two rows of $\boldsymbol{W}(t)$ are identical.

More generally, suppose that S is partitioned into m nonempty similarity classes $S_1, \ldots, S_m$, some of which are not singletons. Let S_g be a nonsingleton similarity class. For $j \in S$, let

$$\alpha_j^g = \boldsymbol{Q}_{ij} \quad \text{for any } i \in S_g, i \neq j. \tag{4}$$

The rate α_j^g is well defined because it is independent of the choice of i. For $i \in S_g$, set the self-transition rate as $\rho_i = \alpha_i^g$. Since $\lambda_i = \rho_i + \sum_{j \neq i} \boldsymbol{Q}_{ij} = \sum_j \alpha_j^g$ is independent of the choice of $i \in S_g$, it will be denoted by λ_g^*. For any $i \in S_g$, $\tau_1 \mid \{X(0) = i\}$ is exponential with rate λ_g^*, $\mathrm{P}_i\{X(\tau_1) = j\} = \alpha_j^g/\lambda_g^*$, and again by (1) the $\boldsymbol{W}(t)$ rows corresponding to S_g are identical. The results are summarised below.

https://doi.org/10.1239/jap/1245676102 Published online by Cambridge University Press

Definition 2. If $i \in S_g$, a nonsingleton similarity class, α_i^g in (4) is the canonical self-transition rate of i, and $\lambda_i = \sum_j \alpha_j^g$ is the canonical hidden holding rate of i, denoted by λ_g^*. If $S_g = \{i\}$ is a singleton then any value $\lambda_i \geq -\boldsymbol{Q}_{ii}$ will be called canonical. The matrix $\boldsymbol{W}(t) = \boldsymbol{P}(t) - \operatorname{diag}(\mathrm{e}^{-\lambda_1 t}, \dots, \mathrm{e}^{-\lambda_s t})$ is a canonical hidden transition matrix.

Theorem 1. *Suppose that S is partitioned into similar classes $S_1, \dots, S_m$. Then in a canonical hidden transition matrix $\boldsymbol{W}(t)$, rows corresponding to any similarity class are identical. In particular, if i is similar to j then, for any $k \neq i, j$, $\boldsymbol{P}_{ik}(t) = \boldsymbol{P}_{jk}(t)$.*

There is a converse to Theorem 1: if $\boldsymbol{P}_{ik}(t) = \boldsymbol{P}_{jk}(t)$ for every $k \neq i, j$ and every t in a neighbourhood of 0, then i and j are similar. This easily follows from calculating the derivative of $\boldsymbol{P}(t)$ at $t = 0$. Ignoring similarity, $\boldsymbol{P}(t)$ can be derived by solving $s(s-1)$ differential equations. Taking similarity into account, only $m(s-1)$ are needed, where m is the number of similarity classes. Thus, a significant reduction is obtained if the number of states is large relative to the number of similarity classes. In fact, the required work can sometimes be less, which can be understood via lumped Markov chains.

2.2. Lumped Markov chains

Let S be partitioned into m similarity classes $S_1, \dots, S_m$. It is intuitive that the chain obtained by lumping nonsingleton similar classes is Markovian. Indeed, this is true by direct verification of the criteria in [1]. Denote the rate matrix of the lumped Markov chain by $\boldsymbol{Q}^*$. We have

$$\boldsymbol{Q}^*_{S_g S_h} = \boldsymbol{Q}^*_{gh} = \sum_{j \in S_h} \boldsymbol{Q}_{ij} \quad \text{for any } i \in S_g, \tag{5}$$

which is well defined because of similarity. It is also readily concluded from the Kolmogorov forward equation $\boldsymbol{P}'(t) = \boldsymbol{P}(t)\boldsymbol{Q}$ that the lumped transition matrix $\boldsymbol{P}^*(t)$ is given by

$$\boldsymbol{P}^*_{gh}(t) = \sum_{j \in S_h} \boldsymbol{P}_{ij}(t) \quad \text{for any } i \in S_g. \tag{6}$$

The lumped transition matrix $\boldsymbol{P}^*(t)$ can be found by solving $m(m-1)$ differential equations. Next, (6) and the fact that $\boldsymbol{W}(t)$ is completely determined by certain m rows, one from each similarity class, implies that solving an additional $m(s-m)$ differential equations will yield $\boldsymbol{P}(t)$. Although the total $m(s-1)$ is the same as using just Theorem 1, the lumped chain clarifies certain cases which need less work. For example, (6) implies that if $m-1$ similarity classes are lumped then the columns of $\boldsymbol{P}(t)$ corresponding to the unlumped class is given directly by $\boldsymbol{P}^*(t)$. Thus, if the similarity classes are all of equal size, say k, then $\boldsymbol{P}(t)$ is obtained by applying m times the expression for the case where one class is of size k, and the other $m-1$ classes are singletons, which only requires solving $m(k+m-1)$ differential equations. Furthermore, as will be seen later, if $m = 1$ then no differential equation is required. Also, if $m = 2$ then with $\boldsymbol{P}^*(t)$ given by (11), below, at most $2s-4$ differential equations need to be solved.

By relabelling the states if necessary, assume that states $1, \dots, k$ are in the similarity class S_1 with the canonical hidden first passage rate $\lambda_1^* = \lambda_1 = \cdots = \lambda_k$. By looking at the form of $\boldsymbol{Q} + \lambda_1^* \boldsymbol{I}$, we conclude that $-\lambda_1^*$ is an eigenvalue of $\boldsymbol{Q}$, and its geometric multiplicity is at least $k-1$. Thus, the characteristic polynomial of $\boldsymbol{Q}$, $p_{\boldsymbol{Q}}(x) = (x + \lambda_1^*)^{k-1} f(x)$ for some polynomial $f(x)$. We claim that in fact $f(x)$ is the characteristic polynomial of $\boldsymbol{Q}^*$, the rate matrix for the lumped chain with only S_1 lumped. Since every eigenvalue of $\boldsymbol{Q}^*$ is an eigenvalue of $\boldsymbol{Q}$, this is clear if the eigenvalues of $\boldsymbol{Q}^*$ are distinct, and distinct from λ_1^*. Otherwise, the

https://doi.org/10.1239/jap/1245676102 Published online by Cambridge University Press

fact that matrices with distinct eigenvalues are dense in the set of matrices implies that, for any $\varepsilon > 0$, there exists A within ε of $\boldsymbol{Q}$, not necessarily a rate matrix, having the following properties: it has $n-k+2$ distinct eigenvalues, one of them, μ, having geometric multiplicity equal to $k-1$, and its first k rows are 'similar' in the same sense as in $\boldsymbol{Q}$. If A^* is obtained from A in the same way $\boldsymbol{Q}^*$ is obtained from $\boldsymbol{Q}$, then $p_A(x) = (x-\mu)^{k-1} p_{A^*}(x)$. Taking limits on both sides yields the result. By induction on the number of nonsingleton similarity classes, we obtain the following result.

Theorem 2. *Suppose that S is partitioned into similar classes $S_1, \dots, S_m$ such that, for $i = 1, \dots, l \le m$, S_i is nonsingleton: $s_i = |S_i| > 1$, with canonical hidden first passage rate λ_i^*. Let $\boldsymbol{Q}^*$ be the rate matrix for the lumped process with $S_1, \dots, S_l$ lumped. Then*

$$p_{\boldsymbol{Q}}(x) = \prod_{i=1}^{l} (x+\lambda_i^*)^{s_i - 1}\, p_{\boldsymbol{Q}^*}(x).$$

2.3. Certain Markov chains with two similarity classes

Consider a Markov chain on $s \ge 3$ states having maximal similarity partition with two components $\{1, \dots, s-1\}$ and $\{s\}$. Then there exist $\alpha_{11}, \dots, \alpha_{1s}$ and $\alpha_{21}, \dots, \alpha_{2,s-1}$ such that, for some $k = 1, \dots, s-1$, $\alpha_{1k} \ne \alpha_{2k}$, and $i \ne j$,

$$\boldsymbol{Q}_{ij} = \begin{cases} \alpha_{1j}, & 1 \le i \le s-1, \\ \alpha_{2j}, & i = s. \end{cases}$$

For $i = 1, \dots, s-1$, $\lambda_i = \sum_{j=1}^{s} \alpha_{1j} = \lambda_1^*$. By Theorem 1, $\boldsymbol{P}(t)$ is determined by the first and last rows of $\boldsymbol{W}(t)$. They can be derived by solving $2(s-2)$ differential equations, in view of (6) and the fact that the lumped Markov chain has only two states.

Indeed, the differential equations can be reduced to linear equations. First, $-\lambda_1^*$ is an eigenvalue of $\boldsymbol{Q}$, with geometric multiplicities at least $s-2$. Next, the lumped Markov chain has two eigenvalues, 0 and $-\lambda_2^* = -\alpha_{1s} - \sum_{j=1}^{s-1} \alpha_{2j}$, which are also eigenvalues of $\boldsymbol{Q}$. If $\lambda_2^* \ne \lambda_1^*$, by Theorem 2, $\boldsymbol{Q}$ is diagonalisable, and $\boldsymbol{P}(t)$ is a linear combination of 1, $\mathrm{e}^{-\lambda_1^* t}$, and $\mathrm{e}^{-\lambda_2^*}$. If $\lambda_2^* = \lambda_1^*$ then the algebraic multiplicity of λ_1^* exceeds its geometric multiplicity by 1, and $\boldsymbol{Q}$ is not diagonalisable. In this case, $\boldsymbol{P}(t)$ is a linear combination of 1, $\mathrm{e}^{-\lambda_1^* t}$, and $t\,\mathrm{e}^{-\lambda_1^* t}$.

For example, the rate matrix (3) has $\lambda_1^* = \lambda_1 = \lambda_2 = \alpha_1 + \alpha_2 + \alpha_3$ and $\lambda_2^* = \alpha_3 + \alpha_4 + \alpha_5$. The following are found by solving just two linear equations. The equilibrium distribution is

$$\boldsymbol{\pi} = \left[\frac{\alpha_1(\alpha_4+\alpha_5) + \alpha_3\alpha_4}{\lambda_1^*\lambda_2^*}, \frac{\alpha_2(\alpha_4+\alpha_5) + \alpha_3\alpha_5}{\lambda_1^*\lambda_2^*}, \frac{\alpha_3}{\lambda_2^*} \right].$$

If $\lambda_1^* \ne \lambda_2^*$, define

$$\boldsymbol{\kappa} = \frac{1}{\lambda_1^* - \lambda_2^*} [\alpha_1 - \alpha_4, \alpha_2 - \alpha_5],$$

$$\boldsymbol{A} = \begin{bmatrix} \pi_1 & \pi_2 & \pi_3 \\ \pi_1 & \pi_2 & \pi_3 \\ \pi_1 & \pi_2 & \pi_3 \end{bmatrix}, \qquad \boldsymbol{B} = \begin{bmatrix} \pi_3\kappa_1 & \pi_3\kappa_2 & -\pi_3 \\ \pi_3\kappa_1 & \pi_3\kappa_2 & -\pi_3 \\ (\pi_3 - 1)\kappa_1 & (\pi_3-1)\kappa_2 & -\pi_3 \end{bmatrix}.$$

Then

$$\boldsymbol{P}(t) = \mathrm{diag}(\mathrm{e}^{-\lambda_1^* t}, \mathrm{e}^{-\lambda_1^* t}, \mathrm{e}^{-\lambda_2^* t}) + \boldsymbol{A} + \mathrm{e}^{-\lambda_2^* t}\boldsymbol{B} - \mathrm{e}^{-\lambda_1^* t}(\boldsymbol{A}+\boldsymbol{B}). \tag{7}$$

https://doi.org/10.1239/jap/1245676102 Published online by Cambridge University Press

For the nondiagonalisable case, $\lambda_1^* = \lambda_2^*$ but $\alpha_1 \neq \alpha_4$, $\boldsymbol{\pi}$ is still as above,

$$\boldsymbol{P}(t) = \operatorname{diag}(\mathrm{e}^{-\lambda_1^* t}, \mathrm{e}^{-\lambda_1^* t}, \mathrm{e}^{-\lambda_1^* t}) + (1 - \mathrm{e}^{-\lambda_1^* t})\boldsymbol{A} + t\,\mathrm{e}^{-\lambda_1^* t}\boldsymbol{C}, \tag{8}$$

where

$$\boldsymbol{C} = \begin{bmatrix} \kappa_1 & -\kappa_1 & 0 \\ \kappa_1 & -\kappa_1 & 0 \\ \kappa_2 & -\kappa_2 & 0 \end{bmatrix}, \qquad \boldsymbol{\kappa} = [\pi_3(\alpha_5 - \alpha_2), \pi_1(\alpha_2 - \alpha_5) - \pi_2(\alpha_1 - \alpha_4)].$$

DNA bases are classified by chemical similarity into pyrimidines Y = {T, C} and purines R = {A, G}. It is well known that intra-class substitutions, or transitions, occur at higher rates than inter-class substitutions, or transversions. The SSL98 model of DNA base substitution models [15] takes this into account, where pyrimidines and purines are similarity classes. For the bases ordered as T, C, A, G,

$$\boldsymbol{Q} = \begin{bmatrix} - & \mu_2 & \mu_3 & \mu_4 \\ \mu_1 & - & \mu_3 & \mu_4 \\ \mu_5 & \mu_6 & - & \mu_8 \\ \mu_5 & \mu_6 & \mu_7 & - \end{bmatrix}.$$

Let $\boldsymbol{P}^{*1}(t)$ and $\boldsymbol{P}^{*2}(t)$ respectively denote the transition matrices of the lumped chains on the ordered states {T, C, R} and {G, A, Y}. Both are of the form (7) or (8). By (6),

$$\boldsymbol{P}(t) = \begin{bmatrix} \boldsymbol{P}^{*1}_{11}(t) & \boldsymbol{P}^{*1}_{12}(t) & \boldsymbol{P}^{*2}_{32}(t) & \boldsymbol{P}^{*2}_{31}(t) \\ \boldsymbol{P}^{*1}_{21}(t) & \boldsymbol{P}^{*1}_{22}(t) & \boldsymbol{P}^{*2}_{32}(t) & \boldsymbol{P}^{*2}_{31}(t) \\ \boldsymbol{P}^{*1}_{31}(t) & \boldsymbol{P}^{*1}_{32}(t) & \boldsymbol{P}^{*2}_{22}(t) & \boldsymbol{P}^{*2}_{21}(t) \\ \boldsymbol{P}^{*1}_{31}(t) & \boldsymbol{P}^{*1}_{32}(t) & \boldsymbol{P}^{*2}_{12}(t) & \boldsymbol{P}^{*2}_{11}(t) \end{bmatrix}.$$

The new derivation is conceptually and computationally easier than the original work, resulting in a much simpler expression in the diagonalisable case, where $\mu_1 + \mu_2 \neq \mu_5 + \mu_6$ and $\mu_3 + \mu_4 \neq \mu_7 + \mu_8$. The authors of [15] seemed unaware that otherwise $\boldsymbol{P}(t)$ assumes three different expressions, though this may not be too important in practice.

2.4. An amino acid substitution model

The 20 naturally occurring amino acids are encoded by 61 codons, or triplets of DNA bases; the other three codons act as signals to stop transcription. Like DNA bases, the substitution process on amino acids is often modelled by Markov chains. Indeed, early research on molecular evolution focused on amino acid sequences, rather than DNA sequences. In a poster entitled 'Parametric models of amino acid evolution' presented by Raspe *et al.* at the 2008 Annual Meeting of the Society for Molecular Biology and Evolution, the 'Bachelor Model 2' treats the amino acids as belonging to three similarity classes of sizes 7, 4, and 9. By the previous discussion, the algebraic expression for $\boldsymbol{P}(t)$ can be derived by solving $3(20 - 1) = 57$ differential equations, instead of 380.

2.5. A Markov chain with two maximal similarity partitions

For the rate matrix (3) with $\alpha_4 = \alpha_1$ and $\alpha_5 \neq \alpha_2$, discussed briefly after the definition of similarity, 1 and 2 are similar, and so are 2 and 3, but 1 and 3 are not. The lumped Markov chain on $\{\{1, 2\}, 3\}$ gives the third column of $\boldsymbol{P}(t)$ as

$$\begin{bmatrix} \pi_3(1-e_3) \\ \pi_3(1-e_3) \\ e_3 + \pi_3(1-e_3) \end{bmatrix},$$

where

$$\pi_3 = \frac{\alpha_3}{\alpha_1+\alpha_5+\alpha_3}, \qquad e_3 = \mathrm{e}^{-(\alpha_1+\alpha_5+\alpha_3)t}.$$

The lumped Markov chain on $\{1, \{2, 3\}\}$ gives the first column as

$$\begin{bmatrix} e_1 + \pi_1(1-e_1) \\ \pi_1(1-e_1) \\ \pi_1(1-e_1) \end{bmatrix},$$

where

$$\pi_1 = \frac{\alpha_1}{\alpha_1+\alpha_2+\alpha_3}, \qquad e_1 = \mathrm{e}^{-(\alpha_1+\alpha_2+\alpha_3)t},$$

so the second column follows.

3. Reversibility

The utility of similar states is more remarkable when the rate matrix is reversible, i.e. for $i \neq j$, $\boldsymbol{Q}_{ij} = \boldsymbol{\beta}_{ij}\pi_j$, where $\boldsymbol{\beta}_{ij} = \boldsymbol{\beta}_{ji}$ and $\boldsymbol{\pi}$ is the equilibrium distribution of the Markov chain (see [11, p. 7]). Under reversibility, the similarity between two states can be checked by looking at the symmetric frequency-adjusted rate matrix $\boldsymbol{\beta}$, and it is an equivalence relation. Indeed, suppose that i and j are similar and that j and k are similar. Then $\boldsymbol{\beta}_{ij} = \boldsymbol{\beta}_{ji} = \boldsymbol{\beta}_{ki} = \boldsymbol{\beta}_{ik} = \boldsymbol{\beta}_{jk} = \boldsymbol{\beta}_{kj}$. By the discussion in Subsection 2.1, this implies that i and k are similar. Therefore, a reversible chain has a unique maximal partition with the smallest number of components. Suppose that S is partitioned maximally into similarity classes $S_1, \ldots, S_m$. It turns out that $\boldsymbol{P}(t)$ is completely determined in a simple manner by $\boldsymbol{P}^*(t)$.

Theorem 3. *Let $\boldsymbol{Q}$ be a reversible rate matrix on $\{1, \ldots, s\}$, the states of which are ordered such that each similarity class consists of consecutive states. Let λ_i be the canonical hidden holding rate for state i, and let λ_g^* be the common rate for similarity class S_g. Let $\boldsymbol{W}(t) = \boldsymbol{P}(t) - \operatorname{diag}(\mathrm{e}^{-\lambda_1 t}, \ldots, \mathrm{e}^{-\lambda_s t})$ be the canonical hidden transition matrix. Denote the transition matrix of the lumped chain by $\boldsymbol{P}^*(t)$. Then, for $i \in S_g$ and $j \in S_h$,*

$$\boldsymbol{W}_{ij}(t) = \begin{cases} \dfrac{\pi_j}{\pi_g^*}[\boldsymbol{P}_{gg}^*(t) - \mathrm{e}^{-\lambda_g^* t}], & h = g, \\[2ex] \dfrac{\pi_j}{\pi_h^*}\boldsymbol{P}_{gh}^*(t), & h \neq g, \end{cases} \tag{9}$$

where $\pi_h^* = \sum_{i \in S_h} \pi_i$. Equivalently,

$$\boldsymbol{P}(t) = \operatorname{diag}(\mathrm{e}^{-\lambda_1 t}, \ldots, \mathrm{e}^{-\lambda_s t}) + \begin{bmatrix} (\boldsymbol{P}_{11}^*(t) - \mathrm{e}^{-\lambda_1^* t})\boldsymbol{J}_1 & \boldsymbol{P}_{12}^*(t)\boldsymbol{J}_2 & \cdots & \boldsymbol{P}_{1m}^*(t)\boldsymbol{J}_m \\ \boldsymbol{P}_{21}^*(t)\boldsymbol{J}_1 & (\boldsymbol{P}_{22}^*(t) - \mathrm{e}^{-\lambda_2^* t})\boldsymbol{J}_2 & \cdots & \boldsymbol{P}_{2m}^*(t)\boldsymbol{J}_m \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{P}_{m1}^*(t)\boldsymbol{J}_1 & \boldsymbol{P}_{m2}^*(t)\boldsymbol{J}_2 & \cdots & (\boldsymbol{P}_{mm}^*(t) - \mathrm{e}^{-\lambda_m^* t})\boldsymbol{J}_m \end{bmatrix},$$

https://doi.org/10.1239/jap/1245676102 Published online by Cambridge University Press

where the square matrix $\boldsymbol{J}_g$ is the Kronecker product of a column vector of 1s with the subvector of the row vector $\boldsymbol{\pi}$ corresponding to g, divided by π_g^*.

A row vector $\boldsymbol{\mu}$ of length s is said to be partially stationary with respect to the reversible process in Theorem 3 if, for each similarity class S_g, the corresponding entries are proportional to the corresponding entries in the stationary distribution $\boldsymbol{\pi}$. If the proportionality constants are equal across all similarity classes, then $\boldsymbol{\mu}$ is stationary, i.e. $\boldsymbol{\mu Q} = \mathbf{0}$. The strategy is to show that every row of $\boldsymbol{W}(t)$ is partially stationary. Then (9) follows from (6) and the proof is complete.

It is useful to define diagonal entries of the frequency-adjusted rate matrix $\boldsymbol{\beta}$. For $i \in S_g$, let $\boldsymbol{\beta}_{ii} = \alpha_i^g/\pi_i$, where α_i^g is the canonical self-transition rate of i. Consequently, for $i \in S_g$,

$$\mathrm{P}_i\{X(\tau_1) = j\} = \frac{\alpha_j^g}{\lambda_g^*} = \frac{\boldsymbol{\beta}_{ij}\pi_j}{\lambda_g^*}, \qquad j \in S. \tag{10}$$

Lemma 1. (a) *Partition the matrix $\boldsymbol{\beta}$ according to the similarity classes. Then each submatrix is constant, i.e. if $i \in S_g$ and $j \in S_h$, $\boldsymbol{\beta}_{ij} = \boldsymbol{\beta}_{gh}^*$.*

(b) *For each i, $X(\tau_1) \mid \{X(0) = i\}$ is partially stationary.*

(c) *If $\boldsymbol{\mu}$ is partially stationary then so is $\boldsymbol{\mu P}(t)$ for any $t \geq 0$.*

Proof. (a) Since the matrix is symmetric, it suffices to show that, for any i, $\boldsymbol{\beta}_{ij_1} = \boldsymbol{\beta}_{ij_2}$ whenever j_1 is similar to j_2. There are three cases: (i) if $j_1 \neq i$ and $j_2 \neq i$, then $\boldsymbol{Q}_{j_1i} = \boldsymbol{Q}_{j_2i}$, or $\boldsymbol{\beta}_{j_1i} = \boldsymbol{\beta}_{j_2i}$, and by reversibility, $\boldsymbol{\beta}_{ij_1} = \boldsymbol{\beta}_{ij_2}$; (ii) if $j_1 = i$ then, by definition and reversibility, $\boldsymbol{\beta}_{ii} = \boldsymbol{\beta}_{j_2i} = \boldsymbol{\beta}_{ij_2}$; (iii) if $j_2 = i$ then similarly $\boldsymbol{\beta}_{ij_1} = \boldsymbol{\beta}_{j_1i} = \boldsymbol{\beta}_{ii}$.

(b) Let $i \in S_g$, and let S_h be a nonsingleton similarity class. It follows from (10) and (a) that, for $j \in S_h$,

$$\mathrm{P}_i\{X(\tau_1) = j\} = \frac{\boldsymbol{\beta}_{ij}\pi_j}{\lambda_g^*} = \frac{\boldsymbol{\beta}_{gh}^*}{\lambda_g^*}\pi_j.$$

(c) We claim that if $\boldsymbol{\mu}$ is partially stationary then so is $\boldsymbol{\nu} = \boldsymbol{\mu Q}$. Let $\boldsymbol{\mu} = (\kappa_1\pi_1, \dots, \kappa_s\pi_s)$, where there are $\kappa_1^*, \dots, \kappa_m^*$ such that $\kappa_i = \kappa_g^*$ if $i \in S_g$. Let $\pi_g^* = \sum_{i \in S_g} \pi_i$. Then, for $j \in S_h$,

$$\nu_j = \sum_{i \neq j} \mu_i \boldsymbol{Q}_{ij} - \mu_j \sum_{i \neq j} \boldsymbol{Q}_{ji} = \pi_j \sum_{i \neq j} \boldsymbol{\beta}_{ij}\pi_i(\kappa_i - \kappa_j) = \pi_j \sum_{g \neq h} \boldsymbol{\beta}_{gh}^*\pi_g^*(\kappa_g^* - \kappa_h^*).$$

It follows that $\boldsymbol{\mu Q}^k$ is partially stationary for any positive integer k. Since $\boldsymbol{P}(t)$ is a power series in $\boldsymbol{Q}$, $\boldsymbol{\mu P}(t)$ is partially stationary.

Now we show that row i of $\boldsymbol{W}(t)$ is partially stationary. Start the chain at $X(0) = i$. For $t > x$, the distribution of $X(t) \mid \{\tau_1 = x\}$ is obtained by multiplying the distribution of $X(x) \mid \{\tau_1 = x\}$ on the right by $\boldsymbol{P}(t - x)$. It follows from Lemma 1(b) and (c) that $X(t) \mid \{\tau_1 = x\}$ is partially stationary. The result follows from

$$\boldsymbol{W}_{ij}(t) = \int_0^t \mathrm{P}_i\{X(t) = j \mid \tau_1 = x\}\lambda_i \,\mathrm{e}^{-\lambda_i x}\,\mathrm{d}x.$$

Equation (9) reduces the derivation of $\boldsymbol{P}(t)$ to that of $\boldsymbol{P}^*(t)$, for which the rate matrix is also reversible, for, by (5) and Lemma 1(a), $\boldsymbol{Q}_{gh}^* = \boldsymbol{\beta}_{gh}^*\pi_h^*$ if $g \neq h$. Thus, the result can be very useful for dealing with large reversible rate matrices with a fair amount of similarity.

https://doi.org/10.1239/jap/1245676102 Published online by Cambridge University Press

3.1. All states belong to one similarity class

Suppose that in a Markov chain every pair of states are similar. Then $\boldsymbol{Q}_{ij} = \alpha_j$ for any $i \neq j$, so that the canonical hidden holding rates $\lambda_i = \sum_{j=1}^{s} \alpha_j = \lambda^*$ are independent of i. It is easy to check that $\boldsymbol{Q}$ is reversible, and the equilibrium distribution $\boldsymbol{\pi}$ is proportional to the αs. Here $\boldsymbol{P}^*(t) \equiv \mathbf{1}$, so, by Theorem 3,

$$\boldsymbol{P}(t) = \mathrm{e}^{-\lambda^* t}\boldsymbol{I} + (1 - \mathrm{e}^{-\lambda^* t})\boldsymbol{J}, \qquad \boldsymbol{J} = \frac{1}{\lambda^*}\begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_s \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1 & \alpha_2 & \cdots & \alpha_s \end{bmatrix}.$$

Since there is only one similarity class, the hidden holding times immediately bring the system into stationarity. The hidden holding times are identically distributed, so the Markov chain has in fact been uniformised.

In the F81 DNA substitution model [4], all states belong to a similarity class. The JC69 model [9] is a special case where all substitution rates are equal, so that the equilibrium distribution is uniform. A heuristic version of hidden holding times is used in [5, pp. 156–157, 200–204] to derive $\boldsymbol{P}(t)$ for these models. The analogues of JC69 and F81 on the amino acids have been described [2, p. 382], [7].

Specialising the result to (2) gives the familiar

$$\boldsymbol{P}(t) = \frac{1}{\alpha_1 + \alpha_2}\begin{bmatrix} \alpha_1 + \alpha_2\,\mathrm{e}^{-(\alpha_1+\alpha_2)t} & \alpha_2 - \alpha_2\,\mathrm{e}^{-(\alpha_1+\alpha_2)t} \\ \alpha_1 - \alpha_1\,\mathrm{e}^{-(\alpha_1+\alpha_2)t} & \alpha_2 + \alpha_1\,\mathrm{e}^{-(\alpha_1+\alpha_2)t} \end{bmatrix}. \tag{11}$$

3.2. Reversible chains with two similarity classes

A reversible rate matrix with exactly two distinct similarity classes of sizes s_1 and $s - s_1$ has the form

$$\boldsymbol{Q} = \left[\begin{array}{cccc|cccc} - & \beta_1\pi_2 & \cdots & \beta_1\pi_{s_1} & \beta\pi_{s_1+1} & \beta\pi_{s_1+2} & \cdots & \beta\pi_s \\ \beta_1\pi_1 & - & \cdots & \beta_1\pi_{s_1} & \beta\pi_{s_1+1} & \beta\pi_{s_1+2} & \cdots & \beta\pi_s \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta_1\pi_1 & \beta_1\pi_2 & \cdots & - & \beta\pi_{s_1+1} & \beta\pi_{s_1+2} & \cdots & \beta\pi_s \\ \hline \beta\pi_1 & \beta\pi_2 & \cdots & \beta\pi_{s_1} & - & \beta_2\pi_{s_1+2} & \cdots & \beta_2\pi_s \\ \beta\pi_1 & \beta\pi_2 & \cdots & \beta\pi_{s_1} & \beta_2\pi_{s_1+1} & - & \cdots & \beta_2\pi_s \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta\pi_1 & \beta\pi_2 & \cdots & \beta\pi_{s_1} & \beta_2\pi_{s_1+1} & \beta_2\pi_{s_1+2} & \cdots & - \end{array}\right].$$

where π is a positive probability vector, and β_1, β_2, and β are positive constants. Let $\pi_1^* = \sum_{j=1}^{s_1} \pi_j$ and $\pi_2^* = \sum_{j=s_1+1}^{s} \pi_j$. The canonical hidden holding rates are

$$\lambda_i = \begin{cases} \beta_1\pi_1^* + \beta\pi_2^*, & 1 \le i \le s_1, \\ \beta\pi_1^* + \beta_2\pi_2^*, & s_1 + 1 \le i \le s. \end{cases}$$

Equation (11) gives $\boldsymbol{P}^*(t)$, with $\alpha_i = \beta\pi_i^*$. Theorem 3 then allows $\boldsymbol{P}(t)$ to be written down immediately. In particular, this applies to the TN93 model of DNA substitution [17], which is obtained by imposing reversibility on SSL98, with $s = 4$ and $s_1 = 2$. The present approach is simpler and more general than the heuristic treatment of [5, pp. 200–204], which can be formalised via hidden holding times. Special cases of TN93 include the HKY85, Felsenstein's

https://doi.org/10.1239/jap/1245676102 Published online by Cambridge University Press

F84, T92, and K80 [8], [12], [13], [16]. Even though these models are mathematically quite simple, they have been used extensively in applications, and will continue to be very useful for understanding the dynamics of evolutionary forces at the molecular level.

3.3. Amino acid substitution models

The poster discussed in Subsection 2.4 also dealt with a reversible version of the 'Bachelor Model 2'. The new approach will significantly simplify the computation, since it only requires the numerical evaluation of the transition matrix of a reversible Markov chain on three states. Finally, substitution models on codons, which are often used to detect evidence of Darwinian selection at the level of DNA, provide opportunities for simplification in the presence of similar states, for example, codons for the same amino acid.

Acknowledgements

This work was supported by the Singapore Ministry of Education Academic Research Fund R-155-050-054-133/R-155-050-054-101. The author is indebted to Bruce Brown for much generous advice and the reviewer for helpful comments.

References

[1] Ball, F. and Yeo, G. F. (1993). Lumpability and marginalisability for continuous-time Markov chains. *J. Appl. Prob.* **30,** 518–528.

[2] Ewens, W. J. and Grant, G. R. (2005). *Statistical Methods in Bioinformatics: An Introduction*, 2nd edn. Springer, New York.

[3] Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, Vol. 2, 2nd edn. John Wiley, New York.

[4] Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Molec. Evol.* **17,** 368–376.

[5] Felsenstein, J. (2004). *Inferring Phylogenies.* Sinauer, New York.

[6] Graur, D. and Li, W.-H. (2000). *Fundamentals of Molecular Evolution*. Sinauer, Sunderland, MA.

[7] Hasegawa, M. and Fujiwara, M. (1993). Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. *Molec. Phylogenet. Evol.* **2,** 1–5.

[8] Hasegawa, M., Kishino, H. and Yano, T. (1985). Phylogenetic relationships among eukaryotic kingdoms inferred from ribosomal RNA sequences. *J. Molec. Evol.* **22,** 32–38.

[9] Jukes, T. H. and Cantor, C. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism.* Academic Press, New York, pp. 21–132.

[10] Keilson, J. (1979). *Markov Chain Models—Rarity and Exponentiality* (Appl. Math. Sci. **28**). Springer, New York.

[11] Kelly, F. P. (1979). *Reversibility and Stochastic Networks.* John Wiley, Chichester.

[12] Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Molec. Evol.* **16,** 111–120.

[13] Kishino, H. and Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Molec. Evol.* **29,** 170–179.

[14] Norris, J. R. (1997). *Markov Chains.* Cambridge University Press.

[15] Schadt, E. E., Sinsheimer, J. S. and Lange, K. (1989). Computational advances in maximum likelihood methods for molecular phylogeny. *Genome Res.* **8,** 222–233.

[16] Tamura, K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C content biases. *Molec. Biol. Evol.* **9,** 678–687.

[17] Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molec. Biol. Evol.* **10,** 512–526.

https://doi.org/10.1239/jap/1245676102 Published online by Cambridge University Press