Theoretical Population Biology 113 (2017) 23–33

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Rate matrix estimation from site frequency data

Conrad J. Burden [a,b,*], Yurong Tang [a]

[a] Mathematical Sciences Institute, Australian National University, Canberra, Australia
[b] Research School of Biology, Australian National University, Canberra, Australia

ARTICLE INFO

Article history:
Received 5 July 2016
Available online 4 November 2016

Keywords:
Multi-allele Wright–Fisher
Decoupled Moran
Neutral evolution
Evolutionary rate matrices

ABSTRACT

A procedure is described for estimating evolutionary rate matrices from observed site frequency data. The procedure assumes (1) that the data are obtained from a constant size population evolving according to a stationary Wright–Fisher or decoupled Moran model; (2) that the data consist of a multiple alignment of a moderate number of sequenced genomes drawn randomly from the population; and (3) that within the genome a large number of independent, neutral sites evolving with a common mutation rate matrix can be identified. No restrictions are imposed on the scaled rate matrix other than that the off-diagonal elements are positive, their sum is $\ll 1$, and that the rows of the matrix sum to zero. In particular the rate matrix is not assumed to be reversible. The key to the method is an approximate stationary solution to the diffusion limit, forward Kolmogorov equation for neutral evolution in the limit of low mutation rates.

1. Introduction

This paper is a continuation of previous work (Burden and Tang, 2016) in which an approximate solution to the forward Kolmogorov equation to the multi-allelic neutral Wright–Fisher or decoupled Moran model is derived for the biologically relevant case of low mutation rates. Herein we address the problem of estimating a mutation rate matrix from site frequency data. The data is assumed to take the form of a multiple alignment of independent, neutrally evolving genomic sites sequenced from a moderate number of individuals chosen independently from a large effective population.

For an alphabet of size K alleles the general mutation rate matrix Q has $K(K-1)$ free parameters, which equates to 12 free parameters for the genomic alphabet $\{A, C, G, T\}$. Classical estimates of mutation rates (Watterson, 1975; Ewens, 1974), and more recent treatments of the problem (see RoyChoudhury and Wakeley, 2010, and references therein) have been concerned primarily with estimating an overall mutation rate, generally denoted by θ, whereas the current paper aims to estimate all parameters of the rate matrix Q. The equivalent estimation problem for $K = 2$ alleles has been solved by Vogl (2014) for neutral sites and Vogl and Bergman (2015) when selection is included.

Zeng (2010) has demonstrated that it is feasible to estimate all parameters of an evolutionary rate matrix from site-frequency data via numerical solution of the multi-allelic discrete Wright–Fisher model by assuming the stationary distribution to be restricted to the corners and edges of a simplicial lattice. Our approach is similar, but differs in that we take advantage of our previously determined approximate analytic solution to the forward Kolmogorov equation. Zeng's approach has the advantage that he is able to estimate selection parameters as well as mutation rates. On the other hand, the approach presented here has the following two main advantages. Firstly, the likelihood function takes a relatively simple analytic form entailing very little in the way of numerical calculation for a given observed site-frequency dataset. Secondly, one gains physical insight into the role of the reversible and non-reversible parts of the rate matrix, and hence a simple statistical test of the hypothesis, commonly assumed in phylogenetic analyses, that the rate matrix is reversible.

A 2×2 rate matrix has a total of 2 free parameters to estimate and is necessarily reversible, which simplifies the problem considerably. The innovation which allows us to deal with the $K > 2$ cases is an interpretation of the non-reversible part of the rate matrix as a set of fluxes of probability around closed paths in the solution-space simplex of the forward Kolmogorov equation (Burden and Tang, 2016). Section 2 sets out a convention for parametrising the general $K \times K$ mutation rate matrix Q which exploits this interpretation. When $K = 4$, for instance, we arrive at 3 independent probabilities defining the stationary Markov state, 6 parameters specifying the remaining degrees of freedom in the reversible part of Q, and 3 probability fluxes

* Corresponding author at: Mathematical Sciences Institute, Australian National University, Canberra, Australia.
E-mail addresses: conrad.burden@anu.edu.au (C.J. Burden), yurong.tang@anu.edu.au (Y. Tang).

http://dx.doi.org/10.1016/j.tpb.2016.10.001
0040-5809/

specifying the non-reversible part, which sums to the required 12 parameters. Section 3 summarises our previously reported approximate stationary solution to the diffusion limit, forward Kolmogorov equation for multi-allelic neutral evolution (Burden and Tang, 2016). Because only low mutation rates are considered the solution can be specified as a set of line densities on the edges and point masses at the corners of the $(K-1)$-dimensional simplex over which the stationary distribution is defined.

The procedure for estimating the parameters of Q from site frequency data is described in Section 4. Maximum likelihood estimates are obtained assuming the data to consist of counts of allele frequencies observed in a finite sample of individuals assumed to be chosen at random from the population. Interestingly, Roy-Choudhury and Wakeley (2010) come close to providing the equivalent estimate for the restricted case of a parent-independent rate matrix, but only specify the overall scale θ and not the complete rate matrix, which, for their restricted case, has K parameters and is reversible. Our estimates are tested using synthetic data for $K = 3$ and $K = 4$ rate matrices in Section 5. Conclusions are summarised in Section 6.

2. Parametrisation of the rate matrix Q

Suppose we are given any $K \times K$ rate matrix Q whose elements Q_{ab}, where $a, b = 1, \ldots, K$, must satisfy

$$Q_{ab} \geq 0, \quad \text{for } a \neq b, \quad \text{and} \quad \sum_{b=1}^{K} Q_{ab} = 0. \tag{1}$$

These constraints imply that $K(K-1)$ parameters are necessary to specify Q. Inspired by the results of Burden and Tang (2016) we begin our analysis by constructing a parametrisation consistent with the decomposition of Q into a reversible part (Lanave et al., 1984; Tavaré, 1986) and a flux part, that is,

$$Q = Q^{\text{GTR}} + Q^{\text{flux}}. \tag{2}$$

The flux part represents a set of fluxes of probability around closed paths between subsets of 3 alleles once the Markovian process has settled into its stationary state.

Let us assume that Q has a unique stationary state $\pi^{\text{T}} = (\pi_1 \ldots \pi_K)$ satisfying

$$\pi_a \geq 0, \quad \sum_{a=1}^{K} \pi_a = 1, \qquad \sum_{a=1}^{K} \pi_a Q_{ab} = \pi_b. \tag{3}$$

A sufficient condition for a unique π^{T} to exist is that $Q_{ab} > 0$ for all $a \neq b$. One would expect this to include any biologically realistic model. For an evolving population in its stationary state, the rate of mutations from allele-a to allele-b at any genomic site is $\pi_a Q_{ab}$.

Define parameters C_{ab} and Φ_{ab} by

$$C_{ab} = \pi_a Q_{ab} + \pi_b Q_{ba}, \qquad \Phi_{ab} = \pi_a Q_{ab} - \pi_b Q_{ba}. \tag{4}$$

It is easy to check that

$$Q_{ab} = \tfrac{1}{2}(C_{ab} + \Phi_{ab})/\pi_a. \tag{5}$$

Hence Q can be decomposed according to Eq. (2) where

$$Q^{\text{GTR}}_{ab} = \tfrac{1}{2} C_{ab}/\pi_a, \tag{6}$$

satisfies the time-reversible condition $\pi_a Q^{\text{GTR}}_{ab} = \pi_b Q^{\text{GTR}}_{ba}$, and

$$Q^{\text{flux}}_{ab} = \tfrac{1}{2} \Phi_{ab}/\pi_a. \tag{7}$$

It is clear from Eq. (4) that Φ_{ab} is the net flux of probability per unit time from allele-a to allele-b.

Note that there are certain dependences between the parameters π_a, C_{ab} and Φ_{ab}. Firstly, the normalisation in Eq. (3) implies that only $K-1$ components of π_a are independent, i.e.

$$\pi_K = 1 - \sum_{i=1}^{K-1} \pi_i. \tag{8}$$

Secondly, $C_{ab} = C_{ba}$, and it follows from the properties of Q that $\sum_{b=1}^{K} C_{ab} = 0$. Thus C_{ab} is a symmetric matrix whose diagonal elements are given in terms of its off-diagonal elements via

$$C_{aa} = -\sum_{b \neq a} C_{ab}, \quad a, b = 1, \ldots, K. \tag{9}$$

Thirdly, $\Phi_{ab} = -\Phi_{ba}$, and it follows from the properties of Q that $\sum_{b=1}^{K} \Phi_{ab} = 0$. Thus Φ_{ab} is an antisymmetric matrix whose rows sum to zero, that is, the final row and column of Φ_{ab} are given in terms of the remaining elements via

$$\Phi_{iK} = -\Phi_{Ki} = -\sum_{j \neq i} \Phi_{ij}, \quad i, j = 1, \ldots, K-1. \tag{10}$$

Eq. (10) is a statement that, in the steady state, the net flux of probability from any allele is zero. For $K = 3$ alleles there is only one independent flux, Φ_{12}, and the elements of Q are

$$Q = \frac{1}{2} \begin{pmatrix} \dfrac{-C_{12} - C_{13}}{\pi_1} & \dfrac{C_{12} + \Phi_{12}}{\pi_1} & \dfrac{C_{13} - \Phi_{12}}{\pi_1} \\ \dfrac{C_{12} - \Phi_{12}}{\pi_2} & \dfrac{-C_{12} - C_{23}}{\pi_2} & \dfrac{C_{23} + \Phi_{12}}{\pi_2} \\ \dfrac{C_{13} + \Phi_{12}}{1 - \pi_1 - \pi_2} & \dfrac{C_{23} - \Phi_{12}}{1 - \pi_1 - \pi_2} & \dfrac{-C_{13} - C_{23}}{1 - \pi_1 - \pi_2} \end{pmatrix}. \tag{11}$$

For $K = 4$ alleles here are three independent fluxes Φ_{12}, Φ_{23} and Φ_{31} as illustrated in Fig. 1.

To summarise, the general rate matrix Q can be parametrised via Eqs. (2), (6) and (7) using the following minimal set of parameters:

$$\begin{aligned} &\pi_i, && i = 1, \ldots, K-1 : && K-1 \text{ parameters;} \\ &C_{ab} = C_{ba}, && 1 \leq a < b \leq K : && \tfrac{1}{2}K(K-1) \text{ parameters;} \\ &\Phi_{ij} = -\Phi_{ji}, && 1 \leq i < j \leq K-1 : && \tfrac{1}{2}(K-1)(K-2) \text{ parameters,} \end{aligned} \tag{12}$$

with the remaining, unspecified parameters given by Eqs. (8)–(10). The total number of independent parameters listed in Eq. (12) is $K(K-1)$, as required. The requirement that the off-diagonal elements of Q be positive implies the further constraints on the parameter space that

$$\pi_a \geq 0, \qquad C_{ab} \geq 0, \qquad |\Phi_{ab}| \leq C_{ab}, \quad 1 \leq a < b \leq K. \tag{13}$$

The remainder of this paper is concerned with estimating the $K(K-1)$ parameters of a genomic evolutionary rate matrix from site frequency data assuming a population whose genome includes a large number of independent sites that have evolved to stationarity according to a neutral evolution Wright–Fisher model.

3. Approximate solution to multi-allelic neutral diffusion

We consider the neutral evolution Wright–Fisher model for K alleles, labelled $A_1, \ldots, A_K$ (see, for example, Section 4.1 of Etheridge, 2011). Given a haploid population of size N (or monoecious diploid population of size $N/2$), let the number of individuals of type A_a at time step τ be $Z_a(\tau)$ for discrete times $\tau = 0, 1, 2, \ldots$. Also, let u_{ab} be the probability of an individual making a transition from A_a to A_b in a single time step, where $u_{ab} \geq 0$ and $\sum_{b=1}^{K} u_{ab} = 1$. Writing $\mathbf{Z}(\tau) = (Z_1(\tau), \ldots, Z_K(\tau))$, the multi-allele neutral Wright–Fisher model is defined by the transition matrix