Australian National University

Algorithms for estimating rates of nucleotide change

Yurong Tang

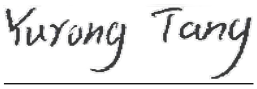
A thesis submitted for the degree of Master of Philosophy

June 2021

Declaration

This thesis is an account of research implemented under the supervision of Professor Gavin Huttley and Associated Professor Conrad Burden at Research School of Biology and Mathematical Sciences Institute, College of Science, Australian National University, Canberra, Australia.

This thesis has not been submitted in whole or part for the purpose of obtaining any other degree at any university. To the best of my knowledge, the material in the thesis is original and has not been previously published or written by another person except where otherwise stated.

Signature: 

Acknowledgments

Foremost, I would like to express my great appreciation to my supervisors Gavin Huttley and Conrad Burden, who have provided valuable and thorough guidance. Without them, I could never finish this thesis. Their broad knowledge and rigorous academic attitude have set an excellent example for me.

I would like to thank my mother, husband, sister and brother for their encouragement and support who are the sources of my strength.

Lastly, I would like to thank all providing help and suggestions to me throughout my studying.

Abstract

The rapidly reducing cost of high throughput sequencing allows for the acquisition of genome-wide data for estimating nucleotide rates not only including mutation rates within species, but including substitution rates between species from multiple sequence alignments.

To address the problem of estimating nucleotide mutation rates within species, Vogl [2014] has developed a general algorithm for the case of bi-allelic neutral evolution. A necessary first step in generalising the Vogl estimator to the multi-allele case is the generalisation of Wright's stationary beta distribution to higher dimensions. This involves finding a stationary solution to the multi-allelic forward Kolmogorov equation. We present an approximate analytic solution to the neutrally evolving multi-allelic forward Kolmogorov equation in the form of a set of line densities defined on the edges of the solution simplex for the general case of K alleles. The solution is obtained in terms of a parameterisation in which the rate matrix Q is decomposed into the sum of a time-reversible part and a non-reversible 'flux' part. The accuracy of the approximate solution with $k = 3$ and $k = 4$ alleles is illustrated using simulation data. The result shows that the agreement between simulation and theory is very close. Based on the approximate analytic solution, 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. We have demonstrated that it is possible, in principle, to estimate an evolutionary rate matrix from the site frequency spectrum of an alignment of genomes sampled from a population, provided certain conditions are met.

Nucleotide substitution rate matrices between species are generally used to calculate the likelihood of phylogenetic trees. In phylogenetic reconstruction, the assumption of heterogeneity in the substitution process across lineages is supported by evidence of compositional heterogeneity between the sequences. However, the total number of possible ways of reducing heterogeneity among lineages is enormous for even a modest number of taxa [Jayaswal et al., 2011]. An efficient strategy for model selection is required to identify an 'optimal' model for a data set. In this study, we address these issues with a novel model selection algorithm we term the Inherited Rate Matrix algorithm (hereafter IRM). This approach is based on the notion that a species inherits the substitution tendencies of its ancestor. We further present the non-stationary heterogeneous across lineages model (hereafter ns-HAL algorithm) which extends the HAL algorithm of Jayaswal et al. [2014] to the general nucleotide Markov process. The IRM algorithm substantially reduces the complexity of identifying a sufficient solution to the problem of time-heterogeneous substitution processes across lineages. We also address the issue of reducing the computing time with development of a constrained-optimisation approach for the IRM algorithm (fast-IRM). From a simulation study based on 2nd codon position genome sequences of yeast, we establish that IRM is significantly more accurate than both ns-HAL and heterogeneity in the substitution process across lineages (hereafter HAL) for close and dispersed sequences. IRM and fast-IRM are faster than ns-HAL. fast-IRM also showed a marked speed improvement over a C++ implementation of HAL. Our

comparison of the accuracy of IRM with fast-IRM showed no difference with identical inferences made for all data sets. These two algorithms greatly improve the compute time for model selection of a non-stationary process, increasing the suite of problems to which this important substitution model class can be applied.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Multiple sequence alignments are completed where three or more protein, DNA or RNA biological sequences are compared. DNA sequence alignments are usually used for comparative genomics analysis, estimation of evolutionary divergence between sequences, and inferring the relationships among genes and species. The rapidly reducing cost of high throughput sequencing now allows for the acquisition of genome-wide data for estimating nucleotide rates not only including specific mutation rates within species, but including substitution rates between species from multiple sequence alignments.

Mutations are changes in the nucleotide sequence of the genome. Mutation rates are the frequency of changes in a single gene or organism over time in DNA sequences [Crow, 1997]. Mutation rates are different in different species and on different regions of the same species. They are nucleotide rates within species, which are of extreme importance in disease progression and in shaping genome evolution and architecture [Nishant et al., 2009].

To estimate mutation rates within species, many methods have been put forward. Counting the number of mutant individuals is a direct and simple mutation rate estimation method [Fu and Huai, 2003]. However, this can lead to an underestimate

of mutation rates due to mutant individuals sharing the same mutations in the same family [Fu and Huai, 2003].Therefore, many mathematical or statistical models have been proposed, in which mutation rates can be measured either with respect to time or with respect to per-generation of pedigree lineage [Xiong et al., 2009]. The former is called the mutation rate in time and the latter is called the mutation rate in replication. The original time method was the fluctuation analysis presented by Luria and Delbruck [Luria and Delbruck, 1943] which was further developed by Stewart et al. [1990] and Qi [2005]. Currently, the most widely used methods are mutation rate in time. Therefore, exploring a method estimating mutation rate in time is one of the primary objects of this thesis.

Generally, physical mutation rates are extremely slow and application to large scale data focuses on small scaled mutation rates. Wright [1931] firstly used a bi-allele model to derive the beta equilibrium distribution of the allelic proportion in a population for scaled mutation rates. The equilibrium distribution can be used for inference of parameters including estimation of mutation rates from site frequency spectra (SFS). Vogl [2014] developed an expectation–maximization (EM) algorithm to obtain the maximum likelihood estimates of both the scaled mutation rate and the mutation bias for the case of bi-allele neutral evolution. Vogl and Bergman [2015] extended the method of Wright [1931] to add selection and analysis of the low mutation rate limit. However, the above methods are all for the bi-allele model and are adequate in certain contexts, but in general they are highly simplified representations of complex underlying systems [Zeng, 2010]. Zeng used a matrix-based approach to construct a multi-allele discrete Wright-Fisher model and has constructed a likelihood-based method for inferring the selection and mutational parameters of the model. Therefore, one object of this thesis is to infer small scaled mutation parameters in neutral mutations by building a multi-allele model in which the rate

matrix is decomposed into the reversible and non-reversible parts.

In a species' genome, there are many sites with mutations with effectively zero fitness effects, which are usually named neutral sites. Theoretically, those neutral sites become fixed between different species at a precise mutation rate [Yi, 2013]. Fixed synonymous mutations are also synonymous substitutions, which change the sequence of a gene but they do not change the protein encoded by that gene. They are often used to estimate the mutation rate although some synonymous mutations contain fitness effects. Therefore, for neutral mutations, the mutation rate can use the substitution rate for estimation. Wielgoss et al. [2011] directly infer the point-mutation rate from the accumulation of synonymous substitutions on 19 Escherichia coli genomes. They compared their results with published genome sequence datasets which are for other bacterial evolution experiments. The results showed that their estimate was the most accurate measure available. This proves inferring mutation rate from substitution rate is reasonable and feasible when some conditions are satisfied. The solution for multi-allele neutral diffusion proposed in this thesis is obtained in terms of a parameterisation in which the rate matrix is decomposed into the sum of a general time-reversible part and a non-reversible 'flux' part. Here, the time-reversible part we used is the general time-reversible substitution rate matrix GTR. The non-reversible part is calculated based on the reversible part.

Another important rates of nucleotide change are substitution rates between species. Nucleotide substitutions are changes resulting in the replacement of a nucleotide in DNA or RNA. Nucleotide substitution rates represent the rates of change between two nucleotides. They are the basis of the evolutionary analysis at the molecular level in phylogenetics.

Nucleotide substitution models describe the rates of change between two nucleotides.

At present, stationary nucleotide substitution models are employed nearly exclusively. These models have the advantage of being relatively simple with a small number of parameters. However, they incorrectly assume that the composition of nucleotides remains roughly the same across a tree [Felsenstein, 2003]. It has been demonstrated that violation of this assumption results in phylogenetic errors (e.g. Foster [2004]).Non-stationary nucleotide substitution models are able to represent divergence in sequence composition. This class of models have been demonstrated to fit sequence data extremely well in comparison to the widely employed stationary substitution models [Kaehler et al., 2015]. Therefore, the nucleotide substitution models used in this thesis are the non-stationary General Nucleotide (GN) model.

Compositional heterogeneity across sequences is very widespread which indicates that stationary, reversible and homogeneous models can lead to bias in phylogenetic estimates [Jayaswal et al., 2014]. Therefore, some models accounting for heterogeneity in the substitution process across lineages have been proposed. For example, Jayaswal et al. [2005] assume a distinct rate matrix on each edge of the phylogenetic tree. Yang and Roberts [1995] assign a distinct vector of equilibrium nucleotide frequencies on each edge of the tree. The above two kinds of methods both possibly overparameterize the data. In order to reduce model complexity, Foster [2004] assigns the same rate matrix to two or more edges in which the number of distinct rate matrices is specified a priori. Jayaswal et al. [2014] proposed a HAL algorithm based on stationary GTR in which optimal models are decided by information criteria from candidate models. However, that algorithm allows disjoint edges sharing the same rate matrix. IRM and fast-IRM described in this thesis only allow joint edges sharing the same rate matrix, which indicates that lineages inherent their mutation tendencies (and thus substitution processes) from their immediate ancestor.

For heterogeneity in substitution process across lineages, it is essential to implement a model complexity approach to identify optimal and near-optimal time-heterogeneous models [Jayaswal et al., 2011]. One model complexity reducing approach is top-down algorithm which starts with the most general model (a distinct rate matrix on every edge) and then gradually decreases the model complexity until a stop condition is satisfied, producing a near-optimal model. Another model complexity reducing approach is a bottom-up approach which starts with the simplest model (one rate matrix) and gradually increase model complexity (number of rate matrices) of the model until a stop condition is satisfied. For a tree with large taxa number, bottom-up approaches should be faster than top-down approaches because the latter start with the most complex (parameter rich) model. IRM, fast-IRM and ns-HAL described in this thesis are all bottom-up approaches.

The structure of this thesis is as follows. Chapter 2 and chapter 3 address the problem of estimating nucleotide mutation rates within species. Chapter 2 describes an approximate stationary solution for the neutrally evolving multi-allele forward Kolmogorov equation for slow mutation rates, which is a method estimating mutation rates in time and inferring the rate from substitution rate. This work has been published. Based on the above approximate solution, site frequency data from multiple alignment is obtained and then maximum likelihood is used for estimating mutation rate from the site frequency data, which is described in Chapter 3. This project is also published. Chapter 4 addresses the problem of estimating nucleotide substitution rates between species. In this chapter, three bottom-up algorithms IRM, fast-IRM and ns-HAL combined with non-stationary GN model, implementing heterogeneous substitution process across lineages, are developed and used for selection of time-heterogeneous non-stationary models. In the process of model selection, maximum likelihood is used to estimate the substitution rates of non-stationary

models which are heterogeneous across lineages. The manuscript for publishing is in preparation.

Chapter 2

An approximate stationary solution for multi-allele neutral diffusion with low mutation rates

Conrad J. Burden, Yurong Tang

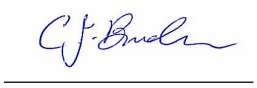
This chapter has been published in Theoretical Population Biology. The reference for the publication is:

Burden CJ, Tang Y. An approximate stationary solution for multi-allele neutral diffusion with low mutation rates. Theor Popul Biol. 112(2016):22-32.

My Contributions For this paper, my supervisor Associate Professor Conrad Burden devised and conducted the project. I attended the discussion of proofs of formulae and GTR selection. I wrote R codes for generating synthetic data, and contributed to drafting the manuscript under the supervision of Associate Professor Conrad Burden. I used the mathematical tool Mathematica to derive a concise formula for the correct normalisation of the solution presented in this paper.

Acknowledgments The authors wish to thank Claus Vogl and Juraj Bergman for encouraging us to investigate the asymptotic behaviour the forward Kolmogorov

equation near the simplex boundary. The results of this investigation are reported in Appendix A. We also wish to thank Asger Hobolth for a very thorough proof-reading of the original manuscript.

Supervisor's signature: 

ABSTRACT We address the problem of determining the stationary distribution of the multi-allelic, neutral-evolution Wright-Fisher model in the diffusion limit. A full solution to this problem for an arbitrary $K \times K$ mutation rate matrix involves solving for the stationary solution of a forward Kolmogorov equation over a $(K-1)$-dimensional simplex, and remains intractable. In most practical situations mutations rates are slow on the scale of the diffusion limit and the solution is heavily concentrated on the corners and edges of the simplex. In this paper we present a practical approximate solution for slow mutation rates in the form of a set of line densities along the edges of the simplex. The method of solution relies on parameterising the general non-reversible rate matrix as the sum of a reversible part and a set of $(K-1)(K-2)/2$ independent terms corresponding to fluxes of probability along closed paths around faces of the simplex. The solution is potentially a first step in estimating non-reversible evolutionary rate matrices from observed allele frequency spectra.

2.1 Introduction

The rapidly reducing cost of high throughput sequencing now allows for the acquisition of genome-wide data for detecting nucleotide allele frequencies extracted from multiple alignments within a population across large numbers of genomic sites [Pool et al., 2010]. The existence of such data raises the possibility of estimating not only specific mutation rates, but complete evolutionary rate matrices from the current observed state of allele frequencies with the genome.

In a recent paper Vogl [2014] has developed a general algorithm and, in the limit of slow scaled mutation rates, a maximum likelihood estimate, of the two parameters defining the scaled instantaneous rate matrix for the case of bi-allelic neutral

evolution. The estimator is similar in style to Watterson's estimator for the infinite allele case [Watterson, 1975], and assumes the data to consist of a site-frequency spectrum (or allele-frequency spectrum) obtained from genotyping a finite number of individuals at a relatively large number of independent sites whose evolution is subject only to genetic drift and identical-rate mutations. It is derived by assuming the data has a beta-binomial distribution as a result of being sampled from the well-known beta-distribution solution to the diffusion limit of the neutral Wright-Fisher model [Wright, 1931]. The method is extended to include selection and the analysis of the low mutation rate limit developed further by Vogl and Bergman [2015].

A necessary first step in generalising the Vogl estimator to the multi-allele case, and in particular to the 4-allele case relevant to genomic rate matrices, is the generalisation of Wright's stationary beta distribution to higher dimensions. This involves finding a stationary solution to the multi-allelic forward Kolmogorov equation (see Eq. (2.3) in the next section). There is no known general solution to this partial differential equation for an arbitrary instantaneous rate matrix.

However, physical mutation rates are extremely slow on the scale relevant to the diffusion limit, and therefore we argue that for practical purposes it is not necessary to solve the forward Kolmogorov equation in its entirety over the full volume of the 3-dimensional simplex on which its solution is defined. Consider for instance the numerical stationary solution to the discrete Wright-Fisher defined by Eqs. (2.1) to (2.2) below, shown in Fig. 2.1. For the purposes of illustration we have simulated this solution using the popular Hasegawa-Kishino-Yano matrix (HKY85) [Hasegawa et al., 1985] with a small population in order to render the simulation numerically tractable, and mutation rates which are unrealistically high by at least two orders of magnitude to enable the distribution to be visible over the entire simplex on the

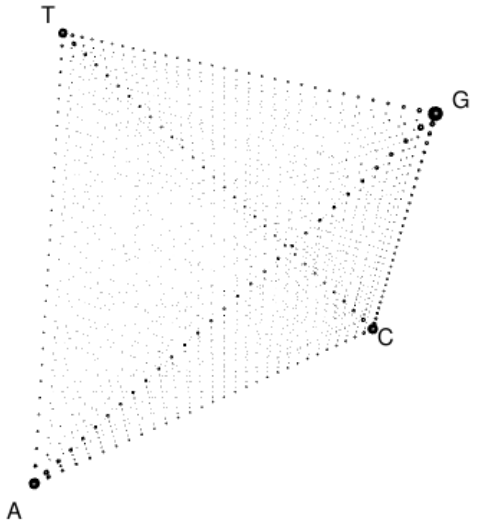


Figure 2.1: Stationary distribution of allele frequencies for the HKY85 model for a haploid population of size $N = 30$ with parameters $\alpha = 0.2$, $\beta = 0.1$, $\pi_A = \pi_T = 0.2$ and $\pi_C = \pi_G = 0.3$, using the parameterisation defined in Hasegawa et al. [1985]. The corners labelled A, C, G and T correspond to allele frequencies $\mathbf{i} = (N, 0, 0, 0)$, $(0, N, 0, 0)$, $(0, 0, N, 0)$ and $(0, 0, 0, N)$ respectively, and the volume of the sphere at each coordinate point is proportional to the probability mass function.

scale of the plot.

The distribution is clearly dominated by the corners of the tetrahedron, indicating that the majority of genomic sites are not polymorphisms (SNPs). This effect is explained in Vogl and Bergman [2015] in the context of the 2-allele Moran model as a strong dominance of genetic drift over mutations for polymorphic sites. Most of the remaining support of the distribution lies on the edges of the tetrahedron, which correspond to 2-allele SNPs. The interiors of the four faces, corresponding to 3-allele SNPs, and the interior volume of the tetrahedron, corresponding to 4-allele SNPs, account for only a small fraction of the total probability. Consistent with observation of the human genome [Hodgkinson and Eyre-Walker, 2010, Cao et al.,

2015, Phillips et al., 2015], the multi-allele neutral Wright-Fisher model predicts that 3- and 4-allele SNPs are extremely rare when scaled mutation rates are low. In fact, when tri-allelic SNPs are observed, the least frequent allele is generally observed in only 1% or 2% of the population (see Table S1 of Hodgkinson and Eyre-Walker [2010]), corresponding to points very close to an edge of the tetrahedron.

Zeng [2010] has demonstrated that it is feasible to estimate mutation rates and selection parameters 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 the simplicial lattice. However an analytic solution to the continuum diffusion limit would facilitate far more computationally efficient maximum likelihood estimate of parameters, and also provide physical insight into the dynamics of mutation.

Below we present an approximate analytic solution to the neutrally evolving multi-allelic forward Kolmogorov equation in the form of a set of line densities defined on the edges of the solution simplex for the general case of K alleles. The basis of our solution is a novel parameterisation of the most general form of the instantaneous rate matrix Q. The parameterisation consists of writing Q as the sum of a time-reversible part [Tavaré, 1986] plus a non-reversible part parametrised by $(K-1)(K-2)/2$ 'probability fluxes' corresponding to a set of independent closed triangular paths following edges of the solution simplex. Note that in this paper the terms "reversible" and "non-reversible" are used in the sense of a continuous-time Markov chain. A non-reversible rate matrix allows mutations among all states, but the system is not in detailed balance, that is to say the mutation rate from allele-1 to allele-2 need not balance the mutation rate from allele-2 to allele-1 at equilibrium. The assumption that rate matrices are reversible is popular in the

phylogenetics literature because the pulley principle [Felsenstein, 1981] simplifies calculations. However there is no biochemical justification for this assumption. We find that in the limit of low mutation rates, and if neutral evolution is assumed, asymmetry in the allele frequency spectrum along edges of the solution simplex can only be explained by the non-reversible part of Q. Equivalently, if Q is reversible, the allele frequency spectrum is symmetric along each edge.

The structure of this paper is as follows. Section 2.2 contains a review of the multi-allelic neutral Wright-Fisher model and sets out the statement of the problem. Section 2.3 reviews the $K = 2$ solution to the forward Kolmogorov equation with a focus on non-standard boundary conditions. Sections 2.4, 2.5 and 2.6 contain our approximate solutions for the $K = 3$, 4 and arbitrary K cases respectively. Section 2.7 discusses the strand-symmetric case. Conclusions are summarised in Section 2.8. Appendix A is devoted to deriving the asymptotic behaviour of the solution to Eq. (2.3) near the simplex boundary in the limit of low mutation rates. Appendix B is devoted to technical details of obtaining marginal distributions of the stationary K-allele solution in terms of effective 2-allele models.

2.2 Review of the multi-allelic neutral Wright-Fisher model

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 diploid population of size $N/2$), let the number of individuals of type A_a at time step τ be $Y_a(\tau)$ 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 $\mathbf{Y}(\tau) = (Y_1(\tau), \ldots, Y_K(\tau)))$, the

multi-allele neutral Wright-Fisher model is defined by the transition matrix from an allele frequency $\mathbf{i} = (i_1, \ldots, i_K)$ to an allele frequency $\mathbf{j} = (j_1, \ldots, j_K)$ in the population given by the multinomial distribution

$$\text{Prob}\left(\mathbf{Y}(\tau+1) = \mathbf{j} | \mathbf{Y}(\tau) = \mathbf{i}\right) = \frac{N!}{\prod_{a=1}^{K} j_a!} \prod_{a=1}^{K} \left(\sum_{b=1}^{K} \frac{i_b}{N} u_{ba} \right)^{j_a}. \tag{2.1}$$

This transition matrix defines a finite state Markov chain with a state space of dimension $\binom{N+K-1}{K-1}$.

The usual diffusion limit is obtained by defining random variables $X_a(t) = Y_a(\tau)/N$ equal to the relative proportion of type-A_a alleles within the population at continuous time $t = \tau/N$. The limit $N \to \infty$ and $u_{ab} \to 0$ for $a \neq b$ is taken in such a way that the $K \times K$ instantaneous rate matrix Q, whose elements are defined by

$$Q_{ab} = N(u_{ab} - \delta_{ab}), \tag{2.2}$$

remains finite. Here δ_{ab} is the Kronecker delta, equal to 1 if $a = b$ and 0 otherwise. This limit gives the forward Kolmogorov equation

$$\frac{\partial f}{\partial t} = -\sum_{a=1}^{K-1} \frac{\partial}{\partial x_a} \sum_{b=1}^{K} x_b Q_{ba} f + \frac{1}{2} \sum_{a,b=1}^{K-1} \frac{\partial^2}{\partial x_a \partial x_b} \left\{ (\delta_{ab} x_a - x_a x_b) f \right\}, \tag{2.3}$$

for the density function $f(x_1, \ldots, x_{K-1}; t)$ of the vector of continuous random variables $X_1(t), \ldots, X_{K-1}(t)$. The function f is defined over the simplex

$$\left\{ (x_1, \ldots, x_{K-1}) : x_1, \ldots, x_{K-1} \geq 0, \sum_{a=1}^{K-1} x_a \leq 1 \right\}. \tag{2.4}$$

For notational convenience, we have defined $x_K = 1 - \sum_{a=1}^{K-1} x_a$ in Eq. (2.3). For details of the derivation of Eq. (2.3) see Lemma 4.2 of Etheridge [2011] or Eq. (5.125) of Ewens [2004].

The stationary distribution $f(x_1, \ldots, x_{K-1})$ to the forward Kolmogorov equation is obtained by setting $\partial f/\partial t$ to zero. The general solution to this problem for an arbitrary rate matrix Q is unknown. However, it is known for the special case in which the elements of the rate matrix take the form $Q_{ab} = Q_b$ (independent of a), in which case the solution is a Dirichlet distribution [Wright, 1969, Tier and Keller, 1978, Griffiths, 1979]. Rate matrices of this form are termed 'parent independent' and constitute a subset of the set of general time-reversible rate matrices.

Our aim is to explore the stationary solution in the limit of slow but otherwise arbitrary mutation rates, that is for rate matrices Q constrained only by the requirements that $0 \leq Q_{ab} << 1$ for $a \neq b$ and $\sum_{b=1}^{K} Q_{ab} = 0$. Our analysis is based on two observations. First, in the limit $Q_{ab} \to 0$, the probability distribution f is concentrated along the edges of the simplex, and therefore the solution to Eq. (2.3) can be represented accurately as a set of line densities defined along the edges of the simplex, as demonstrated in Appendix A. Second, we assume that marginal distributions of the stationary distribution to Eq. (2.3) corresponding to partitioning the set of alleles into two distinct subsets (as described in Appendix B) are Wright's well-known beta function solutions to the $K = 2$ case.

2.3 $K = 2$ case

The solution to the $K = 2$ case is of course well known (see for example Section 5.6 of Ewens [2004]). However we summarise the solution in order to establish notation and

to draw attention to properties associated with non-standard boundary conditions. Set x equal to the proportion of A_1 alleles and $(1-x)$ equal to the proportion of A_2 alleles present in a population. Setting $K = 2$ in Eq.(2.3) and integrating once yields

$$\{Q_{12}x - Q_{21}(1-x)\}f(x) + \frac{1}{2}\frac{d}{dx}\{x(1-x)f(x)\} = \Phi, \tag{2.5}$$

where the constant of integration Φ represents a flux of probability per unit time across any point in the interval $[0,1]$. This flux is generally set equal to zero since, when only 2 alleles are present, probability cannot flow across the boundary points at $x = 0$ and 1.However it will be necessary in subsequent sections to consider the solution to Eq. (2.5) when Φ is non-zero.

The most general solution can be written in a form which is symmetric with respect to the two alleles A_1 and A_2 as

$$\begin{aligned} f(x;C,\Phi) &= [C + \{B(x;1-2Q_{21},1-2Q_{12}) \\ &\quad -B(1-x;1-2Q_{12},1-2Q_{21})\}\Phi]x^{2Q_{21}-1}(1-x)^{2Q_{12}-1}, \end{aligned} \tag{2.6}$$

where C is a constant of integration and $B(x;\alpha,\beta) = \int_0^x \xi^{\alpha-1}(1-\xi)^{\beta-1}d\xi$ is the incomplete beta function, defined for any two parameters α and β. The usual solution, first quoted by Wright [1931], is obtained by setting $\Phi = 0$ and $C = 1/B(2Q_{21},2Q_{12})$, where $B(\alpha,\beta) = B(1;\alpha,\beta)$ is the complete beta function.

Equation (2.6) takes a particularly simple form in the limit $0 \le Q_{12}, Q_{21} < \epsilon << 1$ provided x is not close to the boundaries 0 or 1. In this case we have

$$x^{2Q_{21}-1}(1-x)^{2Q_{12}-1} = \frac{1}{x(1-x)}\{1 + O(\epsilon\,|\log x + \log(1-x)|)\}, \tag{2.7}$$

as $\epsilon \to 0$. Furthermore

$$\begin{aligned} B(x; 1-2Q_{21}, 1-2Q_{12}) &= \int_0^x \xi^{-2Q_{21}}(1-\xi)^{-2Q_{12}}d\xi \\ &= \int_0^x \xi^{-2Q_{21}}d\xi\{1 + O(\epsilon)\} \\ &= x^{1-2Q_{21}}\{1 + O(\epsilon)\} \\ &= x\{1 + O(\epsilon(1 + |\log x|)\}, \end{aligned} \tag{2.8}$$

and similarly

$$B(1-x; 1-2Q_{12}, 1-2Q_{21}) = (1-x)\{1 + O(\epsilon(1 + |\log(1-x)|)\}. \tag{2.9}$$

Thus

$$f(x; C, \Phi) \approx \frac{C}{x(1-x)} - \Phi\left(\frac{1}{x} - \frac{1}{1-x}\right), \tag{2.10}$$

provided $\max(Q_{12}, Q_{21}) \times |\log x + \log(1-x)| << 1$. Note that the constant C depends on the 2×2 rate matrix Q, but the flux Φ arises as a constant of integration and is independent of Q, which is necessarily reversible when $K = 2$. Note also that the first term in Eq. (2.10) is encapsulated within the boundary-selection-mutation model of Vogl and Bergman [2015]. In this model the distribution is accounted for by only genetic drift and selection in the region $x \in [1/N, 1-1/N]$ and the local effects of mutation are safely ignored.

2.4 $K = 3$ case

Consider the most general form of a 3×3 instantaneous rate matrix Q, the only restrictions on its elements being that $Q_{ab} \geq 0$ for $a \neq b$ and $\sum_{b=1}^{3} Q_{ab} = 0$. It will prove convenient in what follows to parameterise Q as

$$\begin{aligned} Q = & \begin{pmatrix} -(\alpha_{12}\pi_2 + \alpha_{13}\pi_3) & \alpha_{12}\pi_2 & \alpha_{13}\pi_3 \\ \alpha_{21}\pi_1 & -(\alpha_{21}\pi_1 + \alpha_{23}\pi_3) & \alpha_{23}\pi_3 \\ \alpha_{31}\pi_1 & \alpha_{32}\pi_2 & -(\alpha_{31}\pi_1 + \alpha_{32}\pi_2) \end{pmatrix} \\ & + \frac{\Phi}{2} \begin{pmatrix} 0 & 1/\pi_1 & -1/\pi_1 \\ -1/\pi_2 & 0 & 1/\pi_2 \\ 1/\pi_3 & -1/\pi_3 & 0 \end{pmatrix} \\ = & Q^{\mathrm{GTR}} + Q^{\mathrm{flux}}, \end{aligned} \tag{2.11}$$

subject to the constraints $\alpha_{ab} = \alpha_{ba} \geq 0$, $\pi_a \geq 0$ and $\sum_{a=1}^{3} \pi_a = 1$. The requirement that the off-diagonal elements of Q are non-negative implies the further constraint that

$$|\Phi| \leq \min_{1 \leq a < b \leq 3} (2\alpha_{ab}\pi_a\pi_b). \tag{2.12}$$

The first term, Q^{GTR}, is the general time-reversible rate matrix [Lanave et al., 1984, Tavaré, 1986], with stationary distribution $\pi^T = (\pi_1, \pi_2, \pi_3)$ satisfying $\pi^T Q = \pi^T Q^{\mathrm{GTR}} = 0$. The defining property of a time-reversible rate matrix, namely that its elements Q_{ab}^{GTR} satisfy

$$\pi_a Q_{ab}^{\mathrm{GTR}} = \pi_b Q_{ba}^{\mathrm{GTR}}, \quad a, b = 1, 2, 3 \tag{2.13}$$

implies that, for a Markov chain in its stationary state, transitions from allele A_a to allele A_b occur with the same frequency as transitions from allele A_b to allele A_a. The second term, Q^{flux}, represents a net rate Φ of transitions around the closed loop $A_1 \to A_2 \to A_3 \to A_1$. The factor of $1/2$ ensures that, when the stationary state is achieved, the number of transitions from A_a to A_b minus the number of transitions from A_b to A_a per unit time, namely $\pi_a Q_{ab} - \pi_b Q_{ba}$, is Φ. As required for a 3×3 rate matrix the total number of independent parameters is six: α_{12}, α_{13} and α_{23}, any two of π_1, π_2 and π_3, and Φ.

Now assume that $\max(\alpha_{12}, \alpha_{13}, \alpha_{23}, |\Phi|) << 1$, so that the off-diagonal elements of Q are small and the solution $f(x_1, x_2)$ to Eq. (2.3) is concentrated on the boundary of the 2-simplex illustrated in Fig. 2.2. For $K = 3$, we demonstrate in Appendix A that the stationary solution to the full forward Kolmogorov equation, Eq. (2.3), is well approximated by a line density of the form Eq. (2.10) along each edge of the 2-simplex. Thus we define line densities $f_{12}(x)$, $f_{23}(x)$ and $f_{31}(x)$ of the site frequency spectrum as shown, where the argument x refers to the proportion of A_1, A_2 and A_3-type alleles respectively at a given genomic site, and hence $(1 - x)$ is the proportion of A_2, A_3 and A_1-type alleles respectively.

The form of the rate matrix Q entails that there is a net flux Φ of probability circulating clockwise around the boundary of the simplex. Thus the line density along each edge takes the form of the approximate $K = 2$-allele solution Eq (2.10),

$$f_{ab}(x) = f(x; C_{ab}, \Phi), \quad 1 \leq a \neq b \leq 3, \tag{2.14}$$

where the normalisation constants C_{12}, C_{23} and C_{31} are yet to be determined.

In order to determine the unknown constants we partition the three alleles into two

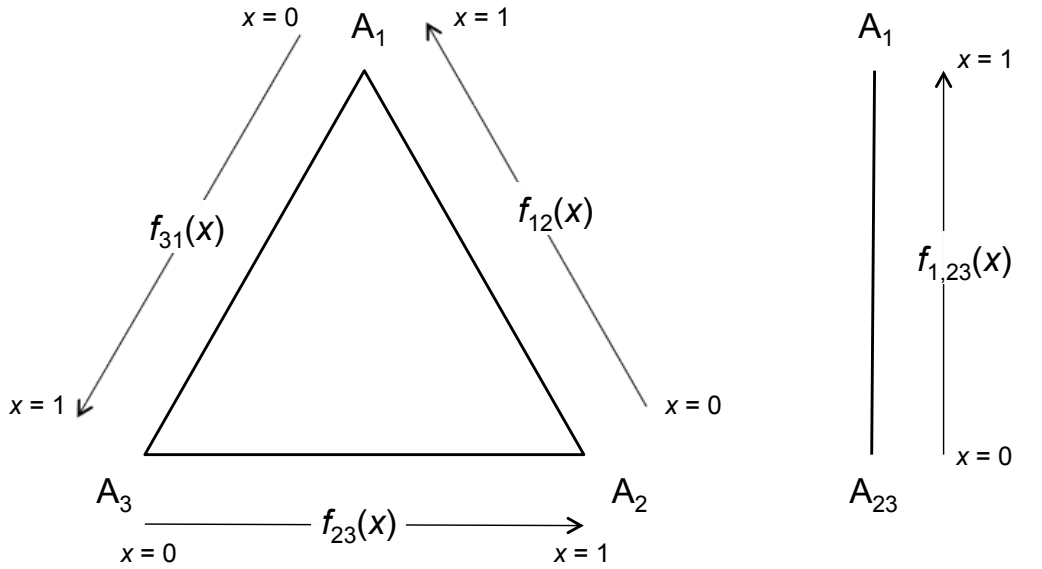


Figure 2.2: The left-hand diagram is the 2-simplex over which the solution to the forward Kolmogorov equation for $K = 3$ alleles is defined. The vertices labelled A_1, A_2 and A_3 correspond to the states in which the entire population has allele A_1, A_2 or A_3 respectively at the genomic site in question. The functions f_{12}, f_{23} and f_{31} are line densities approximating the stationary distribution to Eq. (2.3) on the edges indicated for a rate matrix with transition rates $<< 1$. The right-hand diagram is the 1-simplex supporting the effective 2-allele model corresponding to the allele partitioning (1)(23).

subsets and relabel all alleles within a given subset as an effective single allele as described in Appendix B. For instance, the partitioning (1)(23) illustrated in Fig. 2.2 yields a $K = 2$-state Markov model with transition matrix

$$Q^{(1)(23)} = \begin{pmatrix} -Q_{1,23} & Q_{1,23} \\ Q_{23,1} & -Q_{23,1} \end{pmatrix}, \tag{2.15}$$

where (see Eq. (B.3))

$$Q_{1,23} = Q_{12} + Q_{13}, \qquad Q_{23,1} = \frac{\pi_2 Q_{21} + \pi_3 Q_{31}}{\pi_2 + \pi_3}. \tag{2.16}$$

The stationary state of this matrix is $(\pi_1, \pi_2 + \pi_3)$. As pointed out in the discussion following Eq. (B.2), an effective 2-allele partitioning of the original 3-allele model is not Markovian and therefore does not have a time-dependent behaviour equivalent to that of the Markov chain defined by Eqs. (2.15) and (2.16). However its stationary distribution *is* equivalent to that of the 2-allele Markov chain in the sense that transitions between effective states occur with the same frequency in both models once the stationary state is achieved.

The effective 2-allele model leads to a stationary distribution

$$\begin{aligned} f_{1,23}(x) &= \frac{1}{B(2Q_{23,1}, 2Q_{1,23})} x^{2Q_{23,1}-1}(1-x)^{2Q_{1,23}-1} \\ &\approx \frac{2Q_{1,23}Q_{1,23}}{Q_{1,23}+Q_{1,23}} \times \frac{1}{x(1-x)}. \end{aligned} \tag{2.17}$$

The approximation is valid away from the end points of the interval $0 \leq x \leq 1$, and relies on the approximation

$$\frac{1}{B(\epsilon,\eta)} = \frac{\epsilon\eta}{\epsilon+\eta}\left(1 + O(\epsilon) + O(\eta)\right), \quad \epsilon, \eta \to 0+ . \tag{2.18}$$

No Φ term is present because no flux can cross the boundary points at $x = 0, 1$ of the 2-allele model. In fact this distribution depends only on Q^{GTR} and is independent of Q^{flux}. As suggested by Fig. 2.2 we identify this density with the marginal distribution of the 3-allele site frequency spectrum, and thus

$$\begin{aligned} f_{1,23}(x) &\approx f_{12}(x) + f_{31}(1-x), \\ &\approx \frac{C_{12}+C_{31}}{x(1-x)}, \end{aligned} \tag{2.19}$$

where we have used Eqs. (2.10) and (2.14).

Equating coefficients of Eq. (2.17) and (2.19) gives

$$C_{12} + C_{31} = \frac{2Q_{1,23}Q_{1,23}}{Q_{1,23} + Q_{1,23}}. \tag{2.20}$$

The other two possible partitionings, namely (2)(13) and (3)(12), give similar equations with the indices cyclicly permuted, giving a total of three equations in the three unknowns C_{12}, C_{23} and C_{31}. Using Eqs. (2.11), (2.16) and the symbolic manipulation package Mathematica [Wolfram Research Inc., 2015] we have solved these three equations to obtain the lowest order approximate normalisations

$$C_{ab} = 2\alpha_{ab}\pi_a\pi_b. \tag{2.21}$$

To summarise, given any general rate matrix with off-diagonal elements $Q_{ab} << 1$ and left eigenvector π_a, the stationary distribution to the neutral Wright-Fisher model can be represented as a line density

$$f_{ab}(x) \approx (C_{ab} - \Phi)\frac{1}{x} + (C_{ab} + \Phi)\frac{1}{1-x}, \tag{2.22}$$

along each edge a-b, where x is the relative proportion of A_a alleles. The coefficients are obtained via Eq. (2.11) and (2.21) as

$$C_{ab} = \pi_a Q_{ab} + \pi_b Q_{ba}, \quad \Phi = \pi_a Q_{ab} - \pi_b Q_{ba}. \tag{2.23}$$

The properties of Q ensure that this final formula for Φ is independent of which

edge a-b is chosen. The restriction Eq. (2.12) on Φ ensures that the approximate form of the line density on each edge is non-negative.

To test the accuracy of the analytic solution we have performed a simulation of the neutral Wright-Fisher model with $K = 3$ alleles and a full 3×3 mutation rate matrix. In this simulation the population size is $N = 30$, and mutation rates from allele a to allele b for $a \neq b$ are Q_{ab}/N, where Q_{ab} are elements of a rate matrix of the form Eq. (2.11) with parameters $\alpha_{12} = 1/1000$, $\alpha_{13} = 2/1000$, $\alpha_{23} = 5/1000$, $(\pi_1, \pi_2, \pi_3) = (0.5, 0.3, 0.2)$ and $\Phi = 0.2/1000$. Results are shown in Fig. 2.3. It is clear from the figure that the stationary distribution is almost completely concentrated on the boundary of the simplex and particularly heavily weighted at the corners, as expected. The theoretical line densities f_{12}, f_{23} and f_{31} using the approximation Eq. (2.22) are plotted together with the numerically determined stationary distribution, suitably normalised for comparison. In general, by experimenting with a range of parameter values, we have found agreement between simulation and theory to be very good provided the elements of the rate matrix Q are less than 10^{-2}.

Note that the line densities f_{12}, f_{23} and f_{31} are not suitable for evaluating the stationary distribution close to the corners of the simplex. The probability that a genomic site is a non-segregating site with allele of type A_1, say, can be calculated instead by integrating the effective 2-allele distribution Eq. (2.17) from $1 - 1/N$ to 1, where N is the population size. This gives

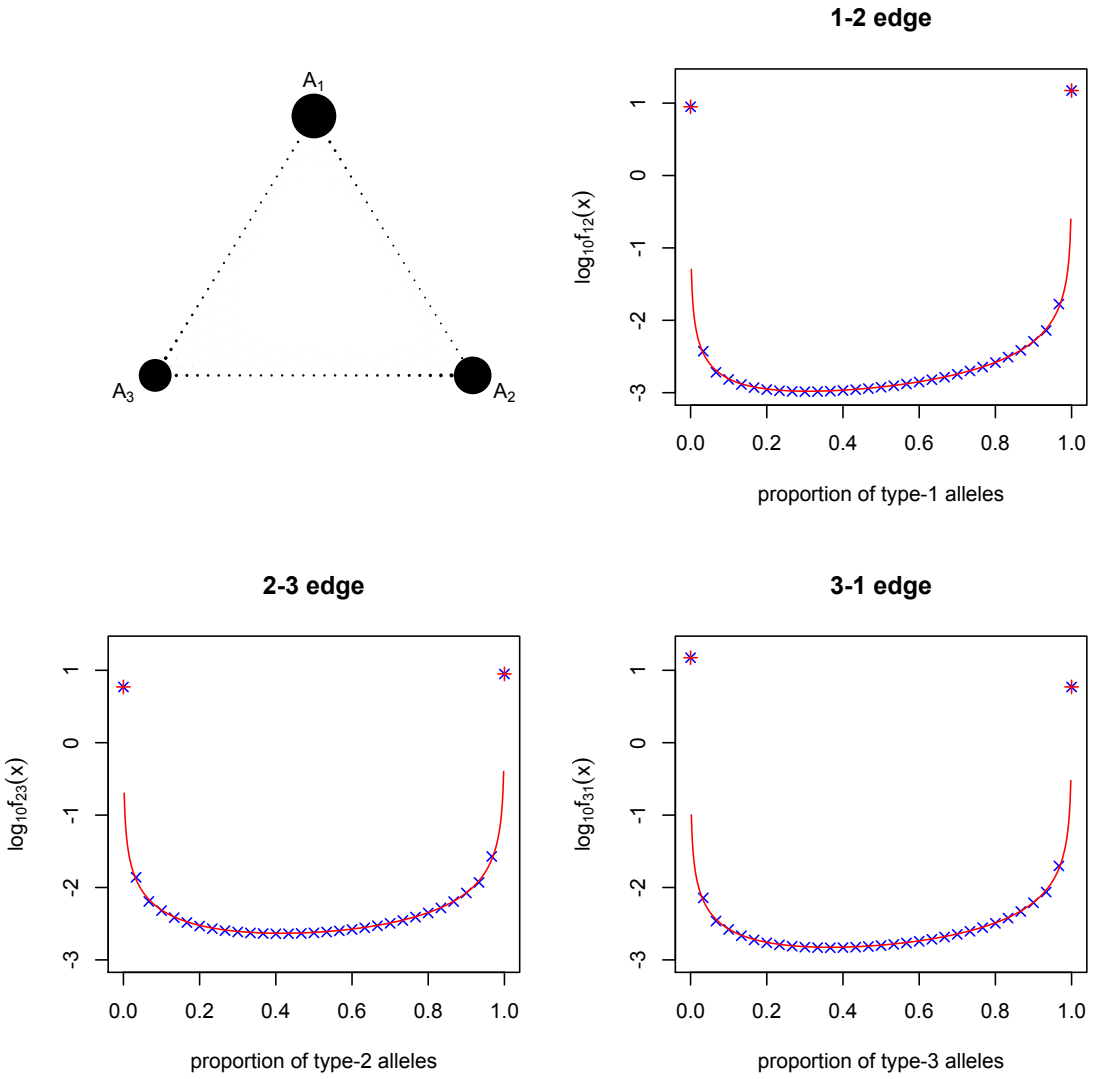


Figure 2.3: Simulation of the neutral Wright-Fisher model with $K = 3$ alleles and a haploid population size $N = 30$. The triangular pattern is the numerically determined stationary site frequency spectrum of a 3-allele neutral Wright-Fisher with mutations. Parameter values are $(\alpha_{12}, \alpha_{13}, \alpha_{23}) = (1, 2, 5)/1000$, $(\pi_1, \pi_2, \pi_3) = (0.5, 0.3, 0.2)$ and $\phi = 0.2/1000$. The full distribution over all $(N + 1)(N + 2)/2$ points is calculated, though points in the interior of the triangle are too small to be easily visible. The remaining three plots are the site frequency spectra on the boundary. Blue crosses are the numerically determined stationary distribution multiplied by N. The red curves are the analytic solutions from Eq. (2.22) and the red plus signs are N times the theoretical probabilities that a site is non-segregating, Eq. (2.24) and cyclic permutations.(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$\begin{aligned} P(A_1) &= \text{Prob}\,(\text{site is non-segregating of type } A_1) \\ &= \int_{1-1/N}^{1} f_{1,23}(x)\,dx \\ &\approx \frac{1}{2Q_{1,23}N^{2Q_{1,23}}B(2Q_{23,1}, 2Q_{1,23})} \\ &\approx \pi_1 N^{-2(\alpha_{12}\pi_2+\alpha_{13}\pi_3)}, \\ &\approx \pi_1 - (C_{12}+C_{31})\log N, \end{aligned} \tag{2.24}$$

where we have used the approximation Eq.(2.18) to the beta-function together with Eqs. (2.11) and (2.16) in the second last line and Eq. (2.21) in the last line. The probability that a site is non-segregating of type A_2 or A_3 is found in a similar manner by cyclicly permuting the indices. These probabilities are also plotted in Fig. 2.3, and agree very well with the numerical simulation.

To complete the $K = 3$ case we check the normalisation of the approximate first order solution. The total probability of the stationary distribution is

$$\sum_{1\le a<b\le 3} \int_{1/N}^{1-1/N} f_{ab}(x)\,dx + \sum_{a=1}^{3} P(A_a). \tag{2.25}$$

From Eq. (2.22), the first term is

$$\sum_{1\le a<b\le 3} \int_{1/N}^{1-1/N} f_{ab}(x)\,dx \approx 2(C_{12}+C_{23}+C_{31})\log N, \tag{2.26}$$

while the second term is, from Eq. (2.24),

$$\begin{aligned}\sum_{a=1}^{3} P(A_a) &\approx \pi_1 - (C_{12} + C_{31}) \log N + \text{cyclic permutations} \\ &= 1 - 2(C_{12} + C_{23} + C_{31}) \log N.\end{aligned} \tag{2.27}$$

The two terms sum to 1, as required.

2.5 $K = 4$ case

The $K = 4$ case is relevant to the DNA alphabet $\{A, C, G, T\}$ and therefore to analysis of genomic site frequency spectra. As for the $K = 3$ case, we consider the most general rate matrix, the only restrictions being that off diagonal elements are non-negative and rows sum to zero. Any such rate matrix is specified by 12 independent parameters can be split into a reversible part and a 'flux' part:

$$Q = Q^{\text{GTR}} + Q^{\text{flux}}. \tag{2.28}$$

The reversible part has 9 independent parameters and can be written as Tavaré [1986]

$$Q^{\text{GTR}} = \begin{pmatrix} \bullet & \alpha_{12}\pi_2 & \alpha_{13}\pi_3 & \alpha_{14}\pi_4 \\ \alpha_{21}\pi_1 & \bullet & \alpha_{23}\pi_3 & \alpha_{24}\pi_4 \\ \alpha_{31}\pi_1 & \alpha_{32}\pi_2 & \bullet & \alpha_{34}\pi_4 \\ \alpha_{41}\pi_1 & \alpha_{42}\pi_2 & \alpha_{43}\pi_3 & \bullet \end{pmatrix} \tag{2.29}$$

where $\alpha_{ab} = \alpha_{ba}$ for $a, b = 1, \ldots, 4$, $a \neq b$, and the diagonal elements are set by ensuring the rows sum to 0. The row vector $(\pi_1, \pi_2, \pi_3, \pi_4)$ is the stationary

distribution of the continuous time Markov model, normalised so that $\sum_{i=1}^{4} \pi_i = 1$. The flux part has 3 parameters and takes the form:

$$Q^{\text{flux}} = \frac{1}{2} \sum_{i=1}^{3} \Phi_i F^{(i)} \tag{2.30}$$

where

$$F^{(1)} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1/\pi_2 & -1/\pi_2 \\ 0 & -1/\pi_3 & 0 & 1/\pi_3 \\ 0 & 1/\pi_4 & -1/\pi_4 & 0 \end{pmatrix}, \quad F^{(2)} = \begin{pmatrix} 0 & 0 & -1/\pi_1 & 1/\pi_1 \\ 0 & 0 & 0 & 0 \\ 1/\pi_3 & 0 & 0 & -1/\pi_3 \\ -1/\pi_4 & 0 & 1/\pi_4 & 0 \end{pmatrix},$$

$$F^{(3)} = \begin{pmatrix} 0 & 1/\pi_1 & 0 & -1/\pi_1 \\ -1/\pi_2 & 0 & 0 & 1/\pi_2 \\ 0 & 0 & 0 & 0 \\ 1/\pi_4 & -1/\pi_4 & 0 & 0 \end{pmatrix}. \tag{2.31}$$

The parameters Φ_i represent a net rate of transitions around each of three independent triangular paths along edges of the 4-allele simplex over which the solution of the forward Kolmogorov equation is defined (see Fig. 2.4). The path around the fourth triangular face, namely $A_1 \to A_2 \to A_3 \to A_1$, can be written as a sum of paths around the other three faces. The requirement that the off-diagonal elements of Q be positive implies that the following constraints must also hold:

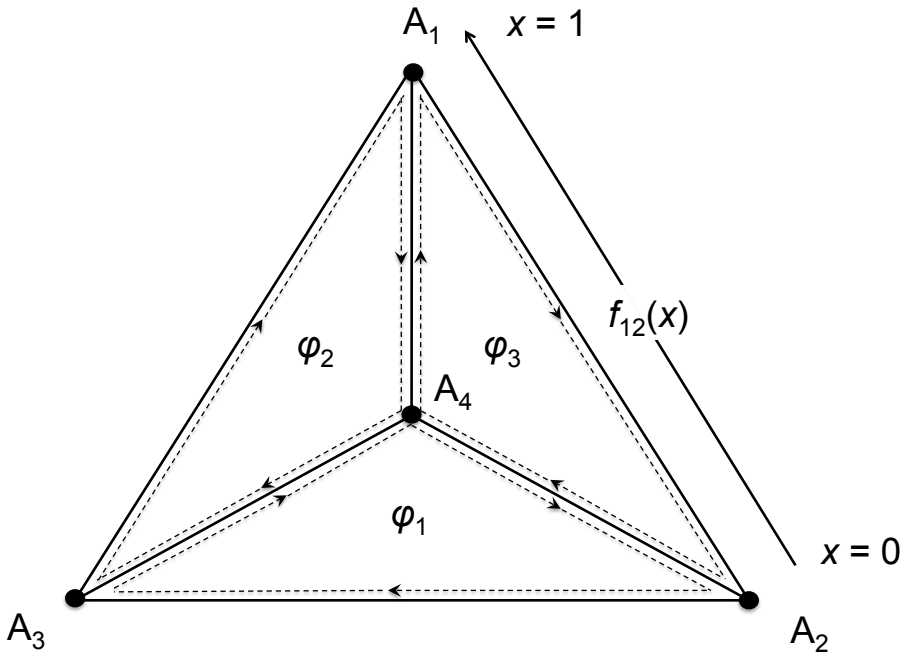


Figure 2.4: The simplex over which the stationary solution to the $K = 4$ forward Kolmogorov equation is defined. The solution is represented as a set of line densities $f_{ab}(x)$ along the edges, with $f_{12}(x)$ shown as an example. Φ_1, Φ_2 and Φ_3 are the net probability fluxes around three triangular paths shown.

$$
\begin{aligned}
|\Phi_1| &\leq 2\alpha_{23}\pi_2\pi_3 \\
|\Phi_2| &\leq 2\alpha_{13}\pi_1\pi_3 \\
|\Phi_3| &\leq 2\alpha_{12}\pi_1\pi_2 \\
|\Phi_1 - \Phi_2| &\leq 2\alpha_{34}\pi_3\pi_4 \\
|\Phi_3 - \Phi_1| &\leq 2\alpha_{24}\pi_2\pi_4 \\
|\Phi_2 - \Phi_3| &\leq 2\alpha_{14}\pi_1\pi_4
\end{aligned}
\tag{2.32}
$$

Under the assumption that $\max(\alpha_{ab}, |\Phi_a|) << 1$, we can again approximate solution of the forward Kolmogorov equation as a set of line densities on the edges of the simplex defined by Eq. (2.4), where the line density on each edge has the form of the function $f(x; C_{ab}, \Phi_{ab})$ defined by Eq. (2.10). Consider, for instance, the A_1-

A_2 edge of the simplex. By first partitioning the alleles into 3 effective alleles A_1, A_2, and $A_{34} = \{A_3, A_4\}$, we can first construct an effective 3-allele model using arguments analogous to those in Appendix B. Then, following the logic of Appendix A we see that the effective 3-allele model has a stationary distribution which can be approximated along the A_1-A_2 boundary by a line density of the appropriate form.

Thus we have a set of six functions

$$f_{ab}(x) = f(x; C_{ab}, \Phi_{ab}), \qquad 1 \leq a < b \leq 4 \tag{2.33}$$

defined using the convention that along the edge (ab), x is the relative proportion of type-a alleles and $1 - x$ is the relative proportion of type-b alleles, as illustrated in Fig. 2.4 for the edge $(ab) = (12)$. From the diagram we read off the flux parameters

$$\begin{aligned} \Phi_{12} = \Phi_3, \quad \Phi_{13} = -\Phi_2, \quad \Phi_{14} = \Phi_2 - \Phi_3, \\ \Phi_{23} = \Phi_1, \quad \Phi_{24} = \Phi_3 - \Phi_1, \quad \Phi_{34} = \Phi_1 - \Phi_2. \end{aligned} \tag{2.34}$$

Following the procedure used for the $K = 3$ case, the normalisations C_{ab} are determined by partitioning the alleles into two distinct subsets and equating the corresponding marginal distributions with the stationary distribution of the equivalent 2-allele model, as defined in Appendix B. There are 7 possible partitionings, namely 4 of the from $(a)(bcd)$ and 3 of the form $(ab)(cd)$, leading to 7 equations for the six unknowns C_{ab}. We have determined using Mathematica [Wolfram Research Inc., 2015] that, provided one uses the same lowest order approximations as those employed for the $K = 3$ case, the set of equations is not overdetermined, but yields the solution

$$C_{ab} = 2\alpha_{ab}\pi_a\pi_b, \tag{2.35}$$

irrespective of which subset of six equations is used. Note that the C_{ab} depend only on parameters defining $Q^{\rm GTR}$, and not on $Q^{\rm flux}$.

The solutions $P(A_a)$ at the corners of the simplex corresponding to the probabilities that a site is non-segregating and of allele type A_a in a population of size N are found by analogy with Eq. (2.24) to be

$$\begin{aligned} P(A_a) &\approx \pi_a N^{-2\sum_{b\neq a}\alpha_{ab}\pi_b} \\ &\approx \pi_a - \sum_{b\neq a} C_{ab}\log N. \end{aligned} \tag{2.36}$$

The accuracy of the approximate solution is illustrated in Figs. 2.5 and 2.6. These plots show simulations of the stationary distribution of the 4-allele neutral Wright-Fisher model defined by Eq. (2.1) for a population of $N = 30$. The mutation rates u_{ab} in Eqs. (2.1) and (2.2) correspond to an instantaneous rate matrix Q with parameters $(\alpha_{12}, \alpha_{13}, \alpha_{14}, \alpha_{23}, \alpha_{24}, \alpha_{34}) = (1, 2, 3, 4, 5, 6) \times \theta$, $(\pi_1, \pi_2, \pi_3, \pi_4) = (0.1, 0.2, 0.3, 0.4)$ and $(\Phi_1, \Phi_2, \Phi_3) = (0.4, 0.1, -0.03) \times \theta$ where $\theta = 0.001$ in Fig. 2.5 and 0.01 in Fig. 2.6. Superimposed in red are the approximate theoretical line densities, Eq. (2.10) with coefficients on the edge (ab) given by Eqs. (2.35) and (2.34), and the probabilities at the simplex corners, Eq. (2.36). As a rule of thumb we find that for these parameters and for a range of other parameter values that we have tried, the agreement between simulation and theory is very close provided the off-diagonal elements of Q are less than 10^{-2}, as in Fig. 2.5, but fails to be close for higher mutation rates, as in Fig. 2.6. More specifically, in Fig. 2.6 one sees that the theoretical solution under-estimates slightly at the corners as it fails to account for non-zero probability in the interior of the simplex near each corner, corresponding

to rare 3- or 4-allele SNPs. The normalisation of the complete distribution to unity then causes the approximate line densities to be correspondingly overestimated.

2.6 Arbitrary K

The analogous procedure can in principle be followed for a general $K \times K$ rate matrix Q with $K(K-1)$ parameters for any value of $K \geq 3$ by parameterising Q as a sum of a reversible part and a flux part, as in Eq. (2.28). For the reversible part an appropriate parameterisation for the elements of Q^{GTR} is [Tavaré, 1986]

$$Q_{ab}^{\mathrm{GTR}} = \begin{cases} \alpha_{ab}\pi_b, & \text{if } a \neq b, \\ -\sum_{c \neq a} \alpha_{ac}\pi_c & \text{if } a = b, \end{cases} \tag{2.37}$$

for parameters π_b and α_{ab}, $a \neq b = 1, \ldots, K$, subject to the constraints $\alpha_{ab} = \alpha_{ba} \geq 0$, $\pi_b \geq 0$, and $\sum_{b=1}^{K} \pi_b = 1$. The number of independent parameters required to define Q^{GTR} is

$$\tfrac{1}{2}K(K-1) + (K-1) = \tfrac{1}{2}(K-1)(K+2). \tag{2.38}$$

For the flux part, an appropriate parameterisation is

$$Q^{\mathrm{flux}} = \frac{1}{2} \sum_{1 \leq i < j \leq K-1} \Phi_{ij} F^{(ij)}, \tag{2.39}$$

where the $F^{(ij)}$ are a set of $K \times K$ matrices whose elements are

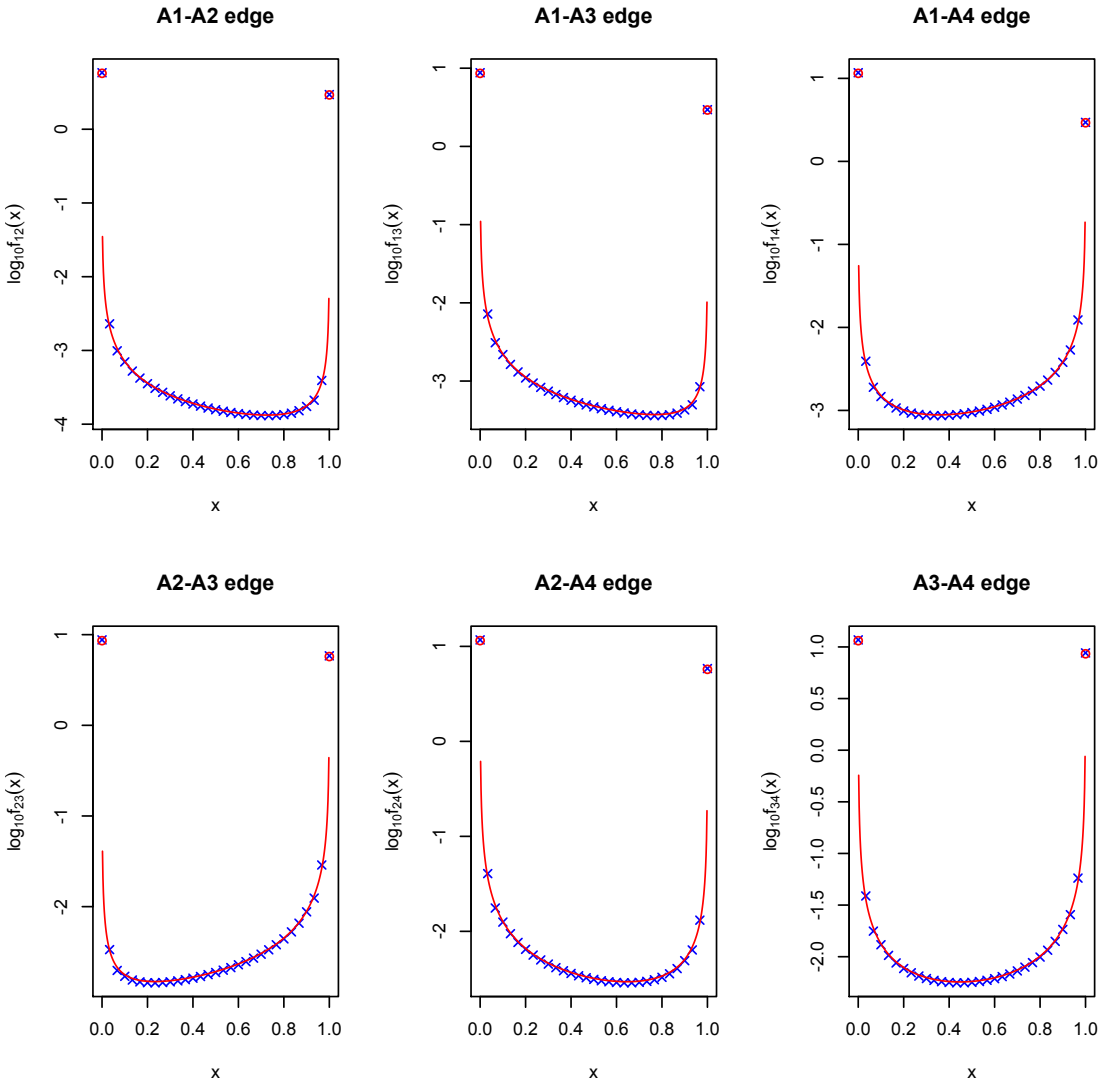


Figure 2.5: Simulation of the neutral Wright-Fisher model with $K = 4$ alleles and a population size $N = 30$. Blue crosses are the numerically determined stationary distribution along each edge of the simplex over which the site frequency distribution is defined. For any 2 alleles, A_a and A_b, the parameter x is the relative proportion of A_a alleles and $1 - x$ is the relative proportion of A_b alleles. Superimposed in red are the theoretical line densities $f_{ab}(x)$, Eq. (2.10) with coefficients on the edge (ab) given by Eqs. (2.35) and (2.34), plotted on a logarithmic scale. The red circles are the theoretical probabilities at the simplex corners, Eq. (2.36). Parameter values of the rate matrix Q are $(\alpha_{12}, \alpha_{13}, \alpha_{14}, \alpha_{23}, \alpha_{24}, \alpha_{34}) = 0.001 \times (1, 2, 3, 4, 5, 6)$, $(\pi_1, \pi_2, \pi_3, \pi_4) = (0.1, 0.2, 0.3, 0.4)$ and $(\Phi_1, \Phi_2, \Phi_3) = 0.001 \times (0.4, 0.1, -0.03)$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

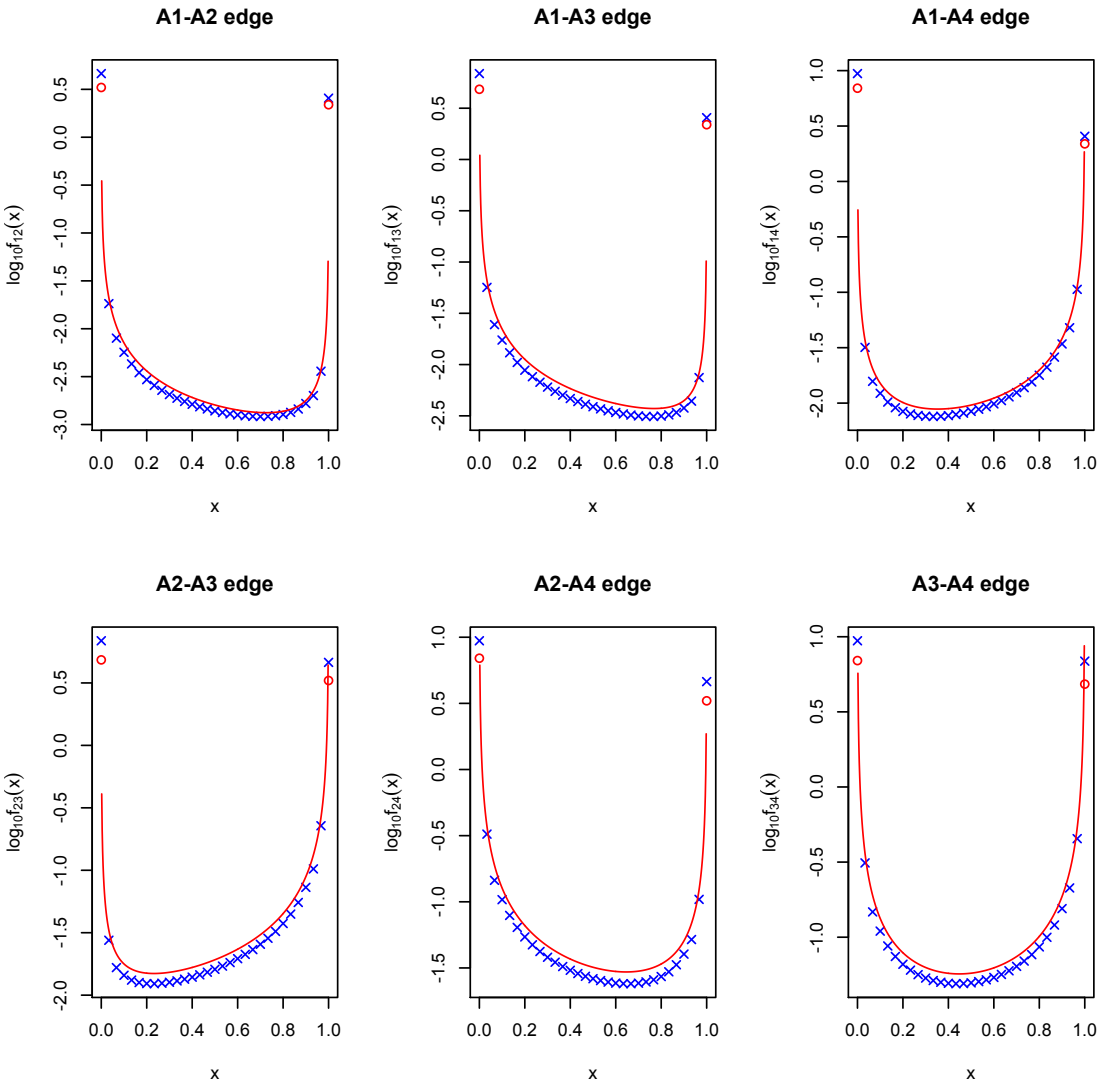


Figure 2.6: The same as Fig. 2.5, but with rate matrix parameters $(\alpha_{12}, \alpha_{13}, \alpha_{14}, \alpha_{23}, \alpha_{24}, \alpha_{34}) = 0.01 \times (1, 2, 3, 4, 5, 6)$, $(\pi_1, \pi_2, \pi_3, \pi_4) = (0.1, 0.2, 0.3, 0.4)$ and $(\Phi_1, \Phi_2, \Phi_3) = 0.01 \times (0.4, 0.1, -0.03)$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$F_{ab}^{(ij)} = \{(\delta_{ia}\delta_{jb} - \delta_{ib}\delta_{ja}) + \delta_{aK}(\delta_{ib} - \delta_{jb}) + \delta_{bK}(\delta_{ja} - \delta_{ia})\}/\pi_a, \tag{2.40}$$

for $1 \leq i < j \leq K-1$ and $a, b = 1, \ldots, K$. In the limit of small mutation rates Φ_{ij} represents a net flux of probability around the closed path $K \to i \to j \to K$ along the edges of the simplex over which the solution to Eq. (2.1) is defined. A flux around any other closed path $i \to j \to k \to i$ of length 3 can be written as a sum of these fluxes. The number of independent fluxes is $(K-1)(K-2)/2$, which, together with Eq. (2.38) implies the total number of parameters defining Q is

$$\tfrac{1}{2}(K-1)(K+2) + \tfrac{1}{2}(K-1)(K-2) = K(K-1),$$

as required for a general rate matrix. It is easy to check that $\pi^{\mathrm{T}} Q^{\mathrm{GTR}} = \pi^{\mathrm{T}} Q^{\mathrm{flux}} = 0$ and hence that $(\pi_1 \ldots \pi_K)$ is the stationary distribution of Q.

The notation for the flux part of the rate matrix can be reconciled with the notation used in Section 2.5 for the $K = 4$ case using the relationships

$$F^{(ij)} = \sum_{k=1}^{3} \epsilon_{ijk} F^{(k)}, \qquad \Phi_{ij} = \sum_{k=1}^{3} \epsilon_{ijk} \Phi_k, \tag{2.41}$$

where ϵ_{ijk} is the Levi-Civita symbol.

In the limit that the off-diagonal elements of Q are $<< 1$, the approximate stationary distribution of the neutral Wright-Fisher model is given as a line density $f_{ab}(x)$ along each edge (ab) of the simplex as

$$f_{ab}(x) = (C_{ab} - \Phi_{ab})\frac{1}{x} + (C_{ab} + \Phi_{ab})\frac{1}{1-x}, \tag{2.42}$$

where x is the relative proportion of A_a alleles, $1-x$ is the relative proportion of A_b alleles, and

$$\begin{aligned} C_{ab} &= 2\alpha_{ab}\pi_a\pi_b = \pi_a Q_{ab} + \pi_b Q_{ba}, \\ \Phi_{ab} &= 2\pi_a Q^{\text{flux}}_{ab} = \pi_a Q_{ab} - \pi_b Q_{ba}, \end{aligned} \tag{2.43}$$

where Q^{flux}_{ab} are the elements of the matrix Q^{flux}. A necessary requirement for the approximate line density to be accurate is that

$$\max_{a\neq b} Q_{ab} \times |\log x + \log(1-x)| << 1, \tag{2.44}$$

where Q_{ab} are the elements of the rate matrix Q. At the corners of the simplex, the probability that a site is non-segregating and of allele type A_a in a population of size N is

$$P(A_a) \approx \pi_a \left(1 - 2\sum_{b\neq a} \alpha_{ab}\pi_b \log N\right) = \pi_a - \sum_{b\neq a} C_{ab} \log N. \tag{2.45}$$

2.7 Strand symmetry

Most genomic sequences, when examined on a sufficiently large scale, are observed to be strand symmetric, that is, symmetric under simultaneous interchange of nucleotides A with T and C with G. This symmetry appears to result from a spectrum of causes, not only mutation rates [Baisnée et al., 2002]. Nevertheless it is interesting to explore the effect of a strand symmetric mutation rate matrix on the symmetries of the neutral-evolution site frequency spectrum.

The most general strand-symmetric rate matrix has 6 independent parameters and takes a form in which the third and fourth rows are permutations of the second and first rows respectively:

$$Q^{\mathrm{SS}} = \begin{pmatrix} \bullet & Q_{AC} & Q_{AG} & Q_{AT} \\ Q_{CA} & \bullet & Q_{CG} & Q_{CT} \\ Q_{CT} & Q_{CG} & \bullet & Q_{CA} \\ Q_{AT} & Q_{AG} & Q_{AC} & \bullet \end{pmatrix}, \tag{2.46}$$

where each diagonal element is minus the sum of the other three elements in its row. It is easy to check that stationary distribution of the continuous-time Markov model generated by this rate matrix is

$$\pi^{\mathrm{T}} = \tfrac{1}{2}(\eta, 1-\eta, 1-\eta, \eta), \tag{2.47}$$

where

$$\eta = \frac{Q_{CA} + Q_{CT}}{Q_{CA} + Q_{CT} + Q_{AC} + Q_{AG}}. \tag{2.48}$$

For the three fluxes Φ_{AC}, Φ_{CG} and Φ_{GA} defined by Eq. (2.41), substitution into Eq. (2.43) gives

$$\Phi_{AC} = \Phi_{GA} = \frac{Q_{CT}Q_{AC} - Q_{AG}Q_{CA}}{2(Q_{CA} + Q_{CT} + Q_{AC} + Q_{AG})}, \qquad \Phi_{CG} = 0. \tag{2.49}$$

Thus a non-reversible strand-symmetric rate matrix has only one independent flux corresponding to the closed path $A \to C \to T \to G \to A$ along the edges of the

tetrahedron in Fig. 2.1. It follows that, in a neutrally-evolving strand-symmetric Wright-Fisher model, the site-frequency spectrum is expected to be symmetric along the AT and CG edges, but may be asymmetric along the remaining edges if the rate matrix is non-reversible.

2.8 Discussion and conclusions

We have obtained approximate solutions for the diffusion limit of the stationary distribution of the K-allele neutral Wright-Fisher model in the realistic limit of small scaled mutation rates for arbitrary values of K. The solution is obtained in terms of a parameterisation in which the rate matrix Q is decomposed into the sum of a general time-reversible part Q^{GTR} and a non-reversible 'flux' part Q^{flux}. The solutions consist of the line densities Eq. (2.42) defined on the edges of the simplex over which the stationary distribution is defined, Eq. (2.4). The approximate line densities lose accuracy, and are not integrable, as the corners of the solution are approached. To overcome this, cutoffs can be imposed at both ends of the edge so that the density is defined on the interval $[1/N, 1 - 1/N]$, where N is an effective population size, and the solution can be represented as the point masses Eq. (2.45) at the corners of the simplex.

The ultimate aim of this work is to estimate rate matrices from allele frequency spectra obtained by genotyping moderate sized samples of a large population. Similar estimates of mutation and selection rate parameters have been obtained for an effective 2-allele model by Vogl and Bergman [2015] using as little as 10 haploid whole genome *Drosophila simulans* sequences. Note that in any genotyping data set the sample size is not the population size parameter N used in the above calculations. The parameter N should be regarded as essentially infinite in the diffusion limit

despite the fact that the high dimensionality of the state space has restricted our numerical simulations to small population sizes, and that, nevertheless, the diffusion limit was observed to be approached very rapidly in N. (For instance, Figs. 2.3 and 2.5 show that the finite population simulation for $N = 30$ is almost indistinguishable from the diffusion limit.)

In the case of $K = 2$ alleles, Vogl [2014] has solved the problem of estimating both parameters of the 2×2 rate matrix from an empirical allele frequency spectrum obtained from a finite sample of M haplotypes, where $M << N$. For higher values of K, his solution can be applied directly to each member of the set of effective 2-allele models defined by partitioning the alleles into distinct complementary subsets as described in Appendix B. This procedure is followed in Vogl and Bergman [2015] for the 4-letter genomic alphabet and the partitioning $(AT)(CG)$. From these estimates of effective 2×2 rate-matrix parameters an estimate of the reversible part Q^{GTR} of the full $K \times K$ can be constructed by considering all partitionings of the full set of K alleles into two exhaustive non-overlapping subsets to form two effective alleles. The non-reversible part Q^{flux} cannot be estimated from the this method, or, for that matter, from simple proportions of segregating and non-segregating sites. From the line-densities plotted in Figs. 2.3, 2.5 and 2.6 it is clear that information about Q^{flux} is contained instead in the asymmetries of the spectrum across the intermediate range of frequencies along each edge of the simplex on which the allele-frequency spectrum is defined. Development of estimators of the parameters of Q^{flux} following this line of attack is the subject of the authors' ongoing work [Burden and Tang, 2017].

Software

The R programs used to produce Figures 2.3, 2.5 and 2.6 and the Mathematica programs used to derive Eqs. (2.21) and (2.35) are available at

`https://github.com/cjb105/SimulateNeutralWF`.

Chapter 3

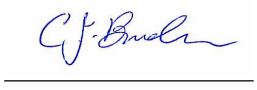
Rate matrix estimation from site frequency data

Conrad J. Burden, Yurong Tang

This chapter has been published in Theoretical Population Biology. The reference for the publication is:

Burden CJ, Tang Y. Rate matrix estimation from site frequency data. Theor Popul Biol. 113(2017):23-33.

My Contributions For this paper, my supervisor Associate Professor Conrad Burden devised and conducted the project. I took part in the designing and coding of parameter estimation method, and generating of synthetic data under the supervision of Associate Professor Conrad Burden.

Supervisor's signature: 

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 $<< 1$, 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.

3.1 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 the 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 3.2

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 specifying the non-reversible part, which sums to the required 12 parameters. Section 3.3 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 3.4. 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 3.5. Conclusions are summarised in Section 3.6.

3.2 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

$$Q_{ab} \geq 0, \quad \text{for } a \neq b, \text{ and } \quad \sum_{b=1}^{K} Q_{ab} = 0. \tag{3.1}$$

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,

$$Q = Q^{\mathrm{GTR}} + Q^{\mathrm{flux}}. \tag{3.2}$$

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 $\pi^{\mathrm{T}} = (\pi_1 \dots \pi_K)$ satisfying

$$\pi_a \geq 0, \quad \sum_{a=1}^{K} \pi_a = 1, \quad \sum_{a=1}^{K} \pi_a Q_{ab} = \pi_b. \tag{3.3}$$

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

$$C_{ab} = \pi_a Q_{ab} + \pi_b Q_{ba}, \qquad \Phi_{ab} = \pi_a Q_{ab} - \pi_b Q_{ba}. \tag{3.4}$$

It is easy to check that

$$Q_{ab} = \tfrac{1}{2}(C_{ab} + \Phi_{ab})/\pi_a. \tag{3.5}$$

Hence Q can be decomposed according to Eq. (3.2) where

$$Q_{ab}^{\mathrm{GTR}} = \tfrac{1}{2}C_{ab}/\pi_a, \tag{3.6}$$

satsifies the time-reversible condition $\pi_a Q_{ab}^{\mathrm{GTR}} = \pi_b Q_{ba}^{\mathrm{GTR}}$, and

$$Q_{ab}^{\mathrm{flux}} = \tfrac{1}{2}\Phi_{ab}/\pi_a. \tag{3.7}$$

It is clear from Eq. (3.4) that Φ_{ab} is the net flux of probability per unit time from allele-a to allele-b.

Note that there are certain dependencies between the parameters π_a, C_{ab} and Φ_{ab}. Firstly, the normalisation in Eq. (3.3) implies that only $K-1$ components of π_a are independent, i.e.

$$\pi_K = 1 - \sum_{i=1}^{K-1} \pi_i. \tag{3.8}$$

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

$$C_{aa} = -\sum_{b \neq a} C_{ab}, \qquad a, b = 1, \ldots, K. \tag{3.9}$$

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

$$\Phi_{iK} = -\Phi_{Ki} = -\sum_{j \neq i} \Phi_{ij}, \qquad i, j = 1, \ldots, K-1. \tag{3.10}$$

Equation (3.10) is at 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

$$Q = \frac{1}{2} \begin{pmatrix} \dfrac{-C_{12} - C_{13}}{\pi_1} & \dfrac{C_{12} + \Phi_{12}}{\pi_1} & \dfrac{C_{13} - \Phi_{12}}{\pi_1} \\ \dfrac{C_{12} - \Phi_{12}}{\pi_2} & \dfrac{-C_{12} - C_{23}}{\pi_2} & \dfrac{C_{23} + \Phi_{12}}{\pi_2} \\ \dfrac{C_{13} + \Phi_{12}}{1 - \pi_1 - \pi_2} & \dfrac{C_{23} - \Phi_{12}}{1 - \pi_1 - \pi_2} & \dfrac{-C_{13} - C_{23}}{1 - \pi_1 - \pi_2} \end{pmatrix}. \tag{3.11}$$

For $K = 4$ alleles here are three independent fluxes Φ_{12}, Φ_{23} and Φ_{31} as illustrated in Fig. 3.1.

To summarise, the general rate matrix Q can be parametrised via Eqs. (3.2), (3.6) and (3.7) using the following minimal set of parameters:

$$\begin{array}{lll} \pi_i, & i = 1, \ldots, K-1: & K-1 \text{ parameters;} \\ C_{ab} = C_{ba}, & 1 \leq a < b \leq K: & \frac{1}{2}K(K-1) \text{ parameters;} \\ \Phi_{ij} = -\Phi_{ji}, & 1 \leq i < j \leq K-1: & \frac{1}{2}(K-1)(K-2) \text{ parameters,} \end{array} \tag{3.12}$$

with the remaining, unspecified parameters given by Eqs. (3.8), (3.9) and (3.10). The total number of independent parameters listed in Eq. (3.12) is $K(K-1)$, as required. The requirement that the off-diagonal elements of Q be positive implies

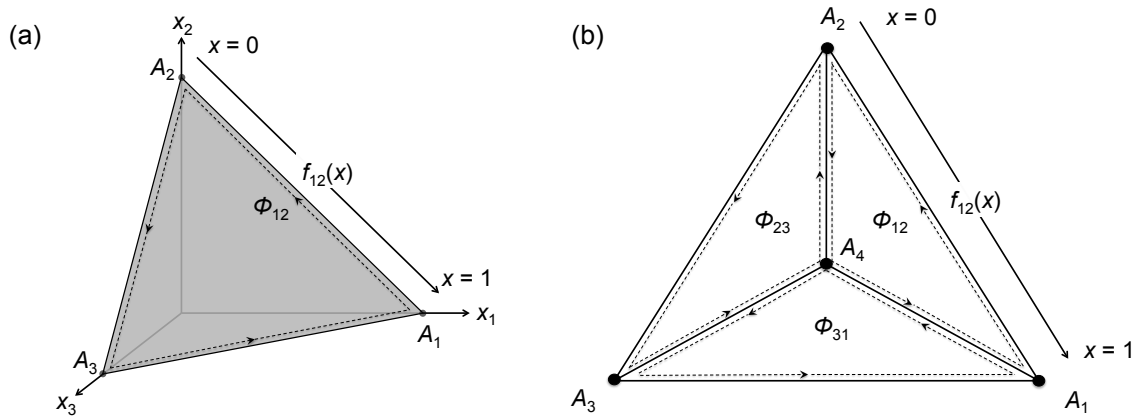


Figure 3.1: The simplex on which the solution to the forward Kolmogorov equation for the multi-allele Wright-Fisher model is defined for (a) $K = 3$ alleles and (b) $K = 4$ alleles. The corners labelled A_1, A_2, etc. indicate the co-ordinates corresponding to non-segregating sites at which the allele specified is prevalent throughout the population. The probability fluxes Φ_{ab} and line densities $f_{ab}(x)$ are explained in the text.

the further constraints on the parameter space that

$$\pi_a \geq 0, \quad C_{ab} \geq 0, \quad |\Phi_{ab}| \leq C_{ab}, \qquad 1 \leq a < b \leq K. \tag{3.13}$$

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.

3.3 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 ref. Etheridge [2011]). Given a haploid popula-

tion 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 from an allele frequency $\mathbf{i} = (i_1, \ldots, i_K)$ to an allele frequency $\mathbf{j} = (j_1, \ldots, j_K)$ in the population given by

$$\text{Prob}\left(\mathbf{Z}(\tau+1) = \mathbf{j} | \mathbf{Z}(\tau) = \mathbf{i}\right) = \frac{N!}{\prod_{a=1}^{K} j_a!} \prod_{a=1}^{K} \psi(\mathbf{i}, a)^{j_a}, \tag{3.14}$$

where $\sum_{a=1}^{K} i_a = \sum_{a=1}^{K} j_a = N$, and

$$\psi(\mathbf{i}, a) = \frac{i_a}{N}\left(1 - \sum_{b \neq a} u_{ab}\right) + \sum_{b \neq a} \frac{i_b}{N} u_{ba} = \sum_{b=1}^{K} \frac{i_b}{N} u_{ba}. \tag{3.15}$$

The usual diffusion limit is obtained by defining random variables $X_a(t) = Z_a(\tau)/N$ equal to the relative proportion of type-A_a alleles within the population at continuous time $t = \tau/N$. The limit $N \to \infty$ and $u_{ab} \to 0$ for $a \neq b$ is taken in such a way that the $K \times K$ instantaneous rate matrix Q, whose elements are defined by

$$Q_{ab} = N(u_{ab} - \delta_{ab}), \tag{3.16}$$

remains finite. This limit leads to the forward Kolmogorov equation

$$\frac{\partial f}{\partial t} = -\sum_{a=1}^{K-1} \frac{\partial}{\partial x_a} \sum_{b=1}^{K} x_b Q_{ba} f + \frac{1}{2} \sum_{a,b=1}^{K-1} \frac{\partial^2}{\partial x_a \partial x_b} \left\{(\delta_{ab} x_a - x_a x_b) f\right\}, \tag{3.17}$$

for the density function $f_{\mathbf{X}}(x_1, \ldots, x_{K-1}; t)$ of the vector of continuous random variables $X_1(t), \ldots, X_{K-1}(t)$. The function $f_{\mathbf{X}}$ is defined over the simplex (see Fig. 3.1)

$$\mathcal{S} = \left\{ (x_1, \ldots, x_K) : x_1, \ldots, x_K \geq 0, \sum_{a=1}^{K} x_a = 1 \right\}. \tag{3.18}$$

For notational convenience, we have defined $x_K = 1 - \sum_{a=1}^{K-1} x_a$ in Eq. (3.17).

Equation (3.17) can also be obtained from the multi-allelic decoupled Moran model [Baake and Bialowons, 2008, Etheridge and Griffiths, 2009] defined by the transition matrix whose off-diagonal elements are

$$\text{Prob}\left(\mathbf{Z}(\tau+1) = \mathbf{j} | \mathbf{Z}(\tau) = \mathbf{i}\right) = \begin{cases} \dfrac{i_a i_b}{N^2} + \dfrac{u_{ab} i_a}{N} & \text{if } \mathbf{j} = \mathbf{i} - \mathbf{e}_a + \mathbf{e}_b \text{ where } a \neq b, \\ \\ 0 & \text{otherwise,} \end{cases} \tag{3.19}$$

where $\mathbf{e}_a$ is the K-dimensional vector with 1 in the ath position and zeroes elsewhere. For this model the continuous time must be defined as $t = \tau/N^2$ and the instantaneous rate matrix defined as $Q_{ab} = 2N(u_{ab} - \delta_{ab})$ to obtain the same form for Equation (3.17).

Solution of the forward Kolmogorov equation for an arbitrary rate matrix Q and $K \geq 3$ alleles, even for the stationary distribution when $\partial f_{\mathbf{X}}/\partial t$ is set to zero, remains an unsolved problem. However, in Burden and Tang [2016] we derive an approximate stationary distribution in the biologically realistic limit of slow but otherwise arbitrary mutation rates. More specifically, we define an overall mutation

rate θ by

$$Q_{ab} = \theta q_{ab}, \qquad \text{where } \sum_{a \neq b} q_{ab} = 1. \tag{3.20}$$

Then in the limit $\theta \to 0$ the stationary probability distribution is concentrated close to the edges of the simplex $\mathcal{S}$ and can be represented accurately as a set of line densities defined along those edges. Suppose we label the corners of $\mathcal{S}$ by the allele prevalent within the population at that corner, so the corner A_1 corresponds to the co-ordinate $(x_1, \ldots, x_K) = (1, 0, \ldots, 0)$, and so on, as illustrated in Fig. 3.1. On the edge joining corner A_a to corner A_b define a line density $f_{ab}(x)$ for each pair of indices a and b. We will adopt the convention that the argument x is the relative proportion of type-a alleles, and $1 - x$ is the relative proportion of type-b alleles. The relative proportion of the remaining $K - 2$ alleles along this edge is zero.

The line densities are given in terms of the parametrisation introduced in Section 3.2 as (see Eq. (53) of Burden and Tang [2016])

$$f_{ab}(x) = C_{ab}\left(\frac{1}{x} + \frac{1}{1-x}\right) - \Phi_{ab}\left(\frac{1}{x} - \frac{1}{1-x}\right). \tag{3.21}$$

Note that $f_{ab}(x) = f_{ba}(1 - x)$. A necessary condition for the line density to be an accurate representation of the exact solution is that

$$\theta \left|\log(x) + \log(1-x)\right| << 1. \tag{3.22}$$

Thus the approximation loses accuracy as x approaches 0 or 1, that is, in the vicinity

of the corners of $\mathcal{S}$. However, for a population of size N, the value of the stationary distribution at the corners of the simplex corresponding to the discrete problem defined by Eq. (3.14) is, to a good approximation, (see Eq. (56) of Burden and Tang [2016])

$$P(A_a) = \text{Prob}\,(Z_a = N, Z_b = 0 \text{ for } b \neq a) = \pi_a - \sum_{b \neq a} C_{ab} \log N. \tag{3.23}$$

As a rule of thumb, we have observed in numerical simulations that Eqs. (3.21) and (3.23) provide a very good approximation to the stationary state of the discrete model provided the off-diagonal elements of Q are less than 10^{-2}. The approximate solution is normalised in the sense that, provided $N >> 1$ and $\theta \log N << 1$,

$$\sum_{1 \leq a < b \leq K} \int_{1/N}^{1-1/N} f_{ab}(x)\, dx + \sum_{a=1}^{K} P(A_a) \approx 1. \tag{3.24}$$

Below we give an outline of the derivation of the approximate solution presented above.

3.3.1 Outline of the derivation

Suppose we partition the set of allelle indices into n distinct nonoverlapping subsets

$$\{1, \ldots, K\} = \bigcup_{A=1}^{n} \mathcal{S}_A, \qquad \mathcal{S}_A \cap \mathcal{S}_B = \emptyset, \quad 1 \leq A, B, \leq n. \tag{3.25}$$

We postulate that the stationary distribution of the multiallele diffusion equation

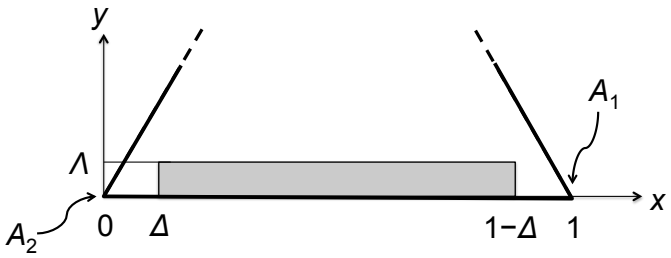


Figure 3.2: The region defined by Eq. (3.27) (shaded) and its relation to the simplex in Fig. 3.1(a)

corresponding to the partitioned instantaneous rate matrix

$$\tilde{Q}_{AB} = \frac{\sum_{B\in\mathcal{S}_B, A\in\mathcal{S}_A} \pi_a Q_{ab}}{\sum_{A\in\mathcal{S}_A} \pi_a}, \tag{3.26}$$

is the corresponding marginal distribution of the stationary distribution of the multiallele diffusion equation corresponding to the full rate matrix Q. This assumption is discussed in detail in Appendix B of Burden and Tang [2016] and can be confirmed for instance for the case of parent-independent mutations, but as far as we are aware it has not been rigorously proven in general. If we accept this postulate, then to derive the line density Eq. (3.21), it is sufficient to consider, without loss of generality, the edge $(ab) = (12)$, partition the set of alleles into three effective alleles A_1, A_2 and $\{A_3, \ldots, K\}$, and derive the result for $f_{12}(x)$ in the 3-allele case. The derivation of the above approximate solution for $K = 3$ is given in Appendix A of Burden and Tang [2016]. Following is a summary.

We will adapt Frobenius method to the full diffusion equation, Eq. (3.17) in the region shown in Fig. 3.2,

$$\Delta \le x \le 1 - \Delta, \qquad 0 \le y \le \Lambda, \tag{3.27}$$

where new coordinates (x, y) are defined by

$$(x_1, x_2, x_3) = (x - \tfrac{1}{2}y, 1 - x - \tfrac{1}{2}y, y). \tag{3.28}$$

The cutoffs Λ and Δ will be discussed below. On the assumption that the behaviour of the stationary solution $f(x, y)$ is analytic with the shaded region except for a power-law singularity as $y \to 0$, the most general form of the solution can be written as

$$f(x, y) = \theta^2 s(x) y^{\theta s(x)-1} \sum_{k=o}^{\infty} g_k(x) y^k, \tag{3.29}$$

where $s(x)$ and the $g_k(x)$ are analytic functions. We will see that the overall mutation rate θ has been included in such a way that $s(x)$ and $g_k(x)$ remain finite in the limit $\theta \to 0$. In Appendix A of Burden and Tang [2016] it is shown, by considering the leading order in y term of Eq. (3.17), that

$$s(x) = 2(q_{13}(x) + q_{23}(x)), \tag{3.30}$$

where q_{ab} are defined by Eq. (3.20).

Now define $f_{12}(x)\,dx$ to be the probability contained in the region $[x, x+dx] \times [0, \Lambda]$. Then

$$\begin{aligned} f_{12}(x) &= \int_0^{\Lambda} f(x,y)\,dy \\ &= \theta g_0(x)\Lambda^{\theta s(x)} + \theta^2 \sum_{k=1}^{\infty} \frac{\Lambda^{\theta s(x)+k}}{\theta s(x)+k} \\ &= \theta g_0(x) + O(\theta^2), \qquad \text{as } \theta \to 0. \end{aligned} \tag{3.31}$$

Importantly, we see that $f_{12}(x)$ is independent of Λ in the absolute limit $\theta \to 0$, and that for practical purposes the approximation to a line density is accurate to leading order in θ provided $\theta\,|\log\Lambda| << 1$, or equivalently, $\Lambda >> e^{-1/\theta}$. Furthermore, by solving the differential equation satisfied by $g_0(x)$, we find that $g_0(x)$ takes the form

$$g_0(x) = \frac{c}{x(1-x)} - \phi\left(\frac{1}{x} - \frac{1}{1-x}\right), \tag{3.32}$$

and that the flux of probability across a line joining $(x,0)$ to (x,Λ) in the x-y plane for any $\Delta \leq x \leq 1-\Delta$ is $\theta\phi + O(\theta^2)$. Equating this flux with the known flux from the exact $K = 2$ solution (see Section 3 of Burden and Tang [2016]), we obtain $\theta\phi = \Phi_{12}$ in the limit $\theta \to 0$.

To determine the scale c, we again appeal to the conjecture that the set of alleles can be partitioned into non-overlapping subsets to obtain the marginal distribution of the stationary distribution. Thus the $K = 3$ model can be partitioned in three different ways to give a $K = 2$ model, for which the exact solution is known. This gives three equations for the normalisations c on each of the three edges of the simplex, which can be solved to obtain $\theta c_{ab} = C_{ab}$ on the (ab) edge of the simplex, provided $\theta \log\Delta << 1$. Details are given in Section 4 of Burden and Tang [2016]. Finally, substituting Eq. (3.32) with these normalisation back into Eq. (3.31) gives Eq. (3.21).

To obtain Eq. (3.23), the partitioning $\{A_a\} \cap \{A_1, \ldots, A_{a-1}, A_{a+1}, \ldots, A_K\}$ is used to give an effective 2-allele model. Integrating the known exact stationary distribution for the $K = 2$ model over the interval $1 - 1/N < x_a < 1$ and assuming $\theta \log N << 1$ gives the required result. See Eq. (24) of Burden and Tang [2016] for details.

3.4 Parameter estimation

Assume we have a data set in the form of a site frequency spectrum (SFS) obtained by sampling L independent neutrally evolving sites within the genomes of M individuals from a population of size $N >> M$. Typically L might be at least 10^3, M in the range 10 to 100, and N is ideally essentially infinite in the sense that the diffusion limit forward-Kolmogorov equation is appropriate. L, M and N are known fixed parameters. At each genomic site l, define a vector of non-negative integer valued random variables

$$\mathbf{Y}^{(l)} = Y_1^{(l)}, \ldots, Y_K^{(l)}, \qquad l = 1, \ldots L, \tag{3.33}$$

where $Y_a^{(l)}$ equal to the number of times allele type-a occurs within the sampled individuals. Clearly $\sum_{a=1}^{K} Y_a^{(l)} = M$, so the data at any given genomic site can be specified as a point in a $(K-1)$-dimensional simplex lattice.

Given an observed data set $(\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(L)})$, the log-likelihood is

$$\mathcal{L}(\pi, C, \Phi | \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(L)}) = \sum_{l=1}^{L} \log \text{Prob}\left(\mathbf{Y}^{(l)} = \mathbf{y}^{(l)} \,|\, \pi, C, \Phi\right), \tag{3.34}$$

where the triplet (π, C, Φ) represents the $K(K-1)$ parameters of Eq. (3.12). The probabilities occurring in this sum are calculated under the assumption that the data is sampled randomly from a population with genomic sites distributed according to the approximate stationary solution of Section 3.3. Below we show that these probabilities are given by

$$\text{Prob}\,(\mathbf{Y}^{(l)} = \mathbf{y} \,|\, \pi, C, \Phi) = \begin{cases} \pi_a - H_M \sum_{b \neq a} C_{ab}, & \text{if } y_a = M \text{ and all other components of } \mathbf{y} \text{ are zero,} \\ C_{ab}\left(\frac{1}{y} + \frac{1}{M-y}\right) - \Phi_{ab}\left(\frac{1}{y} - \frac{1}{M-y}\right) & \text{if } y_a = y,\ y_b = M-y \text{ and all other components of } \mathbf{y} \text{ are zero,} \\ 0 & \text{otherwise,} \end{cases} \tag{3.35}$$

where

$$H_M = \sum_{y=1}^{M-1} \frac{1}{y}. \tag{3.36}$$

This distribution generalises the corresponding $K = 2$ distribution found previously by Vogl, namely Eq. (13) of Vogl [2014], for which Vogl and Bergman [2015] propose the name "generalised RoyChoudhury-Wakeley distribution". There is no Φ_{ab} contribution for $K = 2$ as any 2×2 rate matrix is automatically reversible. Furthermore RoyChoudhury and Wakeley's distribution, namely Eq. (10) of RoyChoudhury and Wakeley [2010], is actually the restriction of the second part of Eq. (3.35) to the

case of a K-allele parent-independent rate matrix, $Q_{ab} = \theta\pi_b/(K-1)$, for $a \neq b$, arbitrary $K \geq 2$ and overall scaling parameter $\theta << 1$ (see Eq. (3.20)). In terms of our parametrisation Eq. (3.12) this corresponds to setting $C_{ab} = 2\theta\pi_a\pi_b/(K-1)$ and $\Phi_{ij} = 0$. The parent-independent rate matrix is also reversible, and includes the most general 2×2 rate matrix when $K = 2$. We therefore propose that the name generalised RoyChoudhury-Wakeley-Vogl-Bergman distribution could be appropriately applied to Eq. (3.35). A proof of Eq. (3.35) follows.

Proof. Consider the three cases in turn.

Case I: Sites for which exactly 1 component of $\mathbf{y}^{(l)}$ is non-zero. Suppose site-l is non-segregating with allele A_a occurring in all individuals within the sample. In this case one can reduce the continuum diffusion limit to an effective 2-allele model in which all alleles except A_a are combined into a single allele, $A_{\bar{a}}$. As described in Eqs. (A.4) and (A.5) of Appendix A of Burden and Tang [2016], the effective 2-allele model has a rate matrix

$$\tilde{Q} = \begin{pmatrix} 1 - \tilde{Q}_{a\bar{a}} & \tilde{Q}_{a\bar{a}} \\ \tilde{Q}_{\bar{a}a} & 1 - \tilde{Q}_{\bar{a}a} \end{pmatrix}, \tag{3.37}$$

with,

$$\begin{aligned} \tilde{Q}_{a\bar{a}} &= \frac{\sum_{b\neq a} \pi_a Q_{ab}}{\pi_a} = \frac{\sum_{b\neq a} C_{ab}}{2\pi_a}, \\ \tilde{Q}_{\bar{a}a} &= \frac{\sum_{b\neq a} \pi_b Q_{ba}}{1-\pi_a} = \frac{\sum_{b\neq a} C_{ab}}{2(1-\pi_a)} \end{aligned} \tag{3.38}$$

where we have used Eq. (3.5) and the fact that Φ_{ab} is an anti-symmetric matrix whose rows sum to zero. In Section 4 of Vogl [2014] he provides an analysis of the 2-allele model in the small mutation rate limit. Vogl parametrises the 2×2 rate

matrix in terms of two parameters, ϑ and α, which are related to our parameters by $\tilde{Q}_{a\bar{a}} = \vartheta/(2\alpha)$, $\tilde{Q}_{\bar{a}a} = \vartheta/[2(1-\alpha)]$, or equivalently,

$$\vartheta = \sum_{b \neq a} C_{ab}, \qquad \alpha = \pi_a. \tag{3.39}$$

From Eq. (29) of Vogl [2014], using the above identification we can immediately read off the required probability

$$\text{Prob}\left\{\mathbf{Y}^{(l)} = (0, \ldots, Y_a^{(l)} = M, \ldots, 0)\right\} = \pi_a - H_M \sum_{b \neq a} C_{ab}. \tag{3.40}$$

Case II: Sites for which exactly 2 components of $\mathbf{y}^{(l)}$ are non-zero. Suppose site l is biallelic with alleles A_a and A_b occurring $y_a^{(l)} = y$ times and $y_b^{(l)} = M - y$ times respectively within the sample. Then, from Eq. (3.21),

$$\begin{aligned}
&\text{Prob}\left\{\mathbf{Y}^{(l)} = (0, \ldots, y, \ldots, M-y, \ldots, 0)\right\} \\
&= \int_{1/N}^{1-1/N} f_{ab}(x) \binom{M}{y} x^y (1-x)^{M-y} dx \\
&= \binom{M}{y} \left[(C_{ab} - \Phi_{ab}) \int_{1/N}^{1-1/N} x^{y-1}(1-x)^{M-y} dx \right. \\
&\quad \left. + (C_{ab} + \Phi_{ab}) \int_{1/N}^{1-1/N} x^y (1-x)^{M-y-1} dx \right] \\
&= \binom{M}{y} \left[(C_{ab} - \Phi_{ab}) B(y, M-y+1) \right. \\
&\quad \left. + (C_{ab} + \Phi_{ab}) B(y+1, M-y) \right] + O\left(\frac{1}{N}\right),
\end{aligned} \tag{3.41}$$

where $B(m,n) = \Gamma(m)\Gamma(n)/\Gamma(m+n)$ is the beta function. For positive integer

arguments $\Gamma(n) = (n-1)!$, which reduces the last line to

$$\begin{aligned} &\text{Prob}\left\{\mathbf{Y}^{(l)} = (0,\ldots,y,\ldots,M-y,\ldots,0)\right\} \\ &\quad\approx\ C_{ab}\left(\frac{1}{y}+\frac{1}{M-y}\right) - \Phi_{ab}\left(\frac{1}{y}-\frac{1}{M-y}\right), \end{aligned} \tag{3.42}$$

up to order $1/N$.

Case III: Sites for which 3 or more components of $\mathbf{y}^{(l)}$ are non-zero. The assumption that the solution to the forward Kolmogorov equation can be represented by a set of line densities on the edges plus point masses at the vertices implies that, in the entire the population, no more than two alleles can be represented at any given genomic site. In other words, the assumed model entails that this case will occur with probability zero. □

Regarding Case III, since 3-allelic and 4-allelic sites are rare in genomes [Cao et al., 2015, Phillips et al., 2015], we argue that removing such sites from the data will not do serious damage to a maximum likelihood estimator of the parameters. Alternatively, one could reassign such data points to the nearest point on an edge of the simplex lattice.

Now define the following variables:

$$\begin{aligned} L_a &= \sum_{i=1}^{L} I\left(Y_a^{(l)} = M, Y_b^{(l)} = 0 \text{ for } b \neq a\right), \quad a = 1,\ldots,K, \\ L_{ab}(y) &= \sum_{i=1}^{L} I\left(Y_a^{(l)} = y, Y_b^{(l)} = M-y, Y_c^{(l)} = 0 \text{ for } c \neq a,b\right), \\ &\qquad\qquad 1 \leq a < b \leq K;\ y = 1,\ldots,M-1, \\ L_{ab} &= \sum_{y=1}^{M-1} L_{ab}(y), \end{aligned} \tag{3.43}$$

where $I(\cdot)$ is the indicator random variable for the event specified. That is, L_a is a count of the number of non-segregating sites of allele type A_a, $L_{ab}(y)$ is a count of the number of biallelic polymorphisms with y occurrences of allele A_a and $M-y$ occurrences of allele A_b, and L_{ab} is the total number of biallelic polymorphisms of type A_a-A_b. We will assume the data is such that all sites observed are either non-segregating or biallelic, i.e., $\sum_a L_a + \sum_{a<b} L_{ab} = L$. Then

$$\begin{aligned}\text{Prob}\left(L_a = l_a, L_{ab}(y) = l_{ab}(y) \,|\, \pi, c, \Phi\right) = \\ \frac{L!}{(\prod_{a=1}^{K} l_a!)(\prod_{a<b}\prod_{y=1}^{M-1} l_{ab}(y)!)} \left[\prod_{a=1}^{K}\left(\pi_a - H_M \sum_{b\neq a} C_{ab}\right)^{l_a}\right] \times \\ \left[\prod_{a<b}\prod_{y=1}^{M-1}\left(C_{ab}\left\{\frac{1}{y}+\frac{1}{M-y}\right\} - \Phi_{ab}\left\{\frac{1}{y}-\frac{1}{M-y}\right\}\right)^{l_{ab}(y)}\right],\end{aligned} \tag{3.44}$$

and

$$\begin{aligned}\text{Prob}\left(L_a = l_a, L_{ab} = l_{ab} \,|\, \pi, c, \Phi\right) = \\ \frac{L!}{(\prod_{a=1}^{K} l_a!)(\prod_{a<b} l_{ab}!)} \left[\prod_{a=1}^{K}\left(\pi_a - H_M \sum_{b\neq a} C_{ab}\right)^{l_a}\right] \left[\prod_{a<b}\left(2H_M C_{ab}\right)^{l_{ab}}\right].\end{aligned} \tag{3.45}$$

Since Eq. (3.45) does not depend on Φ, the observations L_a and L_{ab} are sufficient statistics for estimating π_i and C_{ab} [Ewens, 1974, RoyChoudhury and Wakeley, 2010]. Moreover, for any index $a = 1, \ldots, K$, one can again define an effective 2-allele model by partitioning the set of alleles into A_a and an effective allele $A_{\bar{a}}$ consisting of the remaining $K-1$ alleles, and then use Vogl's unbiased maximum-

likelihood estimator for $\hat{\alpha}$ (see Eq. (37) of Vogl [2014]) together with Eq. (3.39) to obtain the estimators

$$\hat{\pi}_a = \frac{1}{L}\left(L_a + \frac{1}{2}\sum_{b\neq a} L_{ab}\right). \tag{3.46}$$

Similarly, one can consider a broader set of partitionings of alleles into two exhaustive disjoint subsets and the corresponding effective 2-allele models, together with Vogl's unbiased maximum-likelihood estimator for $\hat{\vartheta}$ (see Eq. (36) of Vogl [2014]) to obtain the estimators

$$\hat{C}_{ab} = \frac{L_{ab}}{2LH_M}. \tag{3.47}$$

It is a straightforward exercise to confirm using Eqs.(3.35) and (3.43) that these are unbiased estimators.

It remains to estimate the flux parameters Φ_{ij}. In the following we carry out a numerical maximisation of the log-likelihood formed from Eq. (3.44) by plugging in the above estimates of $\hat{C}_{ab}$. Up to an additive constant this gives

$$\begin{aligned}&\mathcal{L}(\Phi_{ij}|l_a, l_{ab}(1), \ldots, l_{ab}(M-1)) \\ &\quad = \sum_{a<b}\sum_{y=1}^{M-1} l_{ab}(y)\log\left(\hat{C}_{ab}\left\{\frac{1}{y}+\frac{1}{M-y}\right\} - \Phi_{ab}\left\{\frac{1}{y}-\frac{1}{M-y}\right\}\right).\end{aligned} \tag{3.48}$$

The right hand side of this formula is a function of $\frac{1}{2}(K-1)(K-2)$ parameters Φ_{ij} for $1 \leq i < j \leq K-1$, with Eq. (3.10) used to interpret the terms in the sum for which $b = K$.

3.5 Results

We have constructed a number of synthetic datasets to test the efficacy of the above theory. Since the exact solution of the Forward Kolmogorov equation is unknown, starting from an assumed scaled rate matrix Q we first generate numerically a stationary site frequency spectrum from the neutral Wright-Fisher model for as large a population N as is practicable. For this step the full transition matrix, Eq. (3.14), of size $\binom{N+K-1}{K-1} \times \binom{N+K-1}{K-1}$ is used. Each dataset is then created corresponding to a multiple alignment at a large number of L independent genomic sites of the genomes of M individuals sampled randomly from the population. Here we have aimed to satisfy the ideal limit $M << N$ within the constraints of the simulation. Each genomic site is assumed to be the result of the Markov process of the full transition matrix evolving to its stationary state. The sampling process is described in detail below. Finally the parameters of Q for each of 1000 such independently generated datasets are estimated using the theory of Section 3.4, and the results presented as histograms.

3.5.1 Synthetic data: $K = 3$ alleles

For $K = 3$ alleles, and for each rate matrix tested, a numerical stationary solution to the neutral Wright-Fisher model with transition matrix Eq. (3.14) was first created for a population size $N = 100$. Three rate matrices were considered. Each matrix had the same reversible part, Q^{GTR}, corresponding to the parameters (see Eq. (3.11))

$$(\pi_1, \pi_2) = (0.5, 0.3), \qquad (C_{12}, C_{23}, C_{13}) = (0.0003, 0.0006, 0.0004). \tag{3.49}$$

For the single flux parameter which determines Q^{flux}, namely $\Phi \equiv \Phi_{12}$, three cases were considered:

$$\Phi = 0, \quad \Phi = 0.0001, \quad \text{and} \quad \Phi = 0.0002. \tag{3.50}$$

For each value of Φ total of 1000 synthetic datasets were constructed, each assuming a sample of $M = 10$ individuals sequenced at $L = 10^5$ independent genomic sites. At each site $l \in \{1, \ldots, L\}$ and within each dataset the stationary distribution was sampled to establish the relative frequencies of the 3 alleles in the population at that site. These relative frequencies were used as parameters of a multinomial distribution from which we sampled the observed counts $y_a^{(l)} \in \{0, \ldots, M\}$ of the number of times allele A_a was observed at site l. Any genomic site displaying more than 2 alleles in the sample was discarded, though generally this amounted to no more than 1 or 2 tri-allelic SNPs observed per dataset. Maximum likelihood estimates $\hat{\pi}_a$, $\hat{C}_{ab}$ and $\hat{\Phi}$ of the parameters defining Q were obtained for each dataset using the theory developed in Section 3.4 (see Eqs. (3.46) to (3.48)).

Histograms of the estimates $\hat{\pi}_a$ and $\hat{C}_{ab}$ are shown in Fig. 3.3. The estimates $\hat{\pi}_a$ are centred about their true values, while the estimates $\hat{C}_{ab}$ tend to be slightly low, perhaps because of the approximate nature of the solution to the forward Kolmogorov equation. The distributions of these parameters are independent of Φ, as expected.

Histograms of the estimates $\hat{\Phi}$ are shown in Fig. 3.4. These estimates are well centred about their true values. Under the null hypothesis that $\Phi = 0$ the likelihood

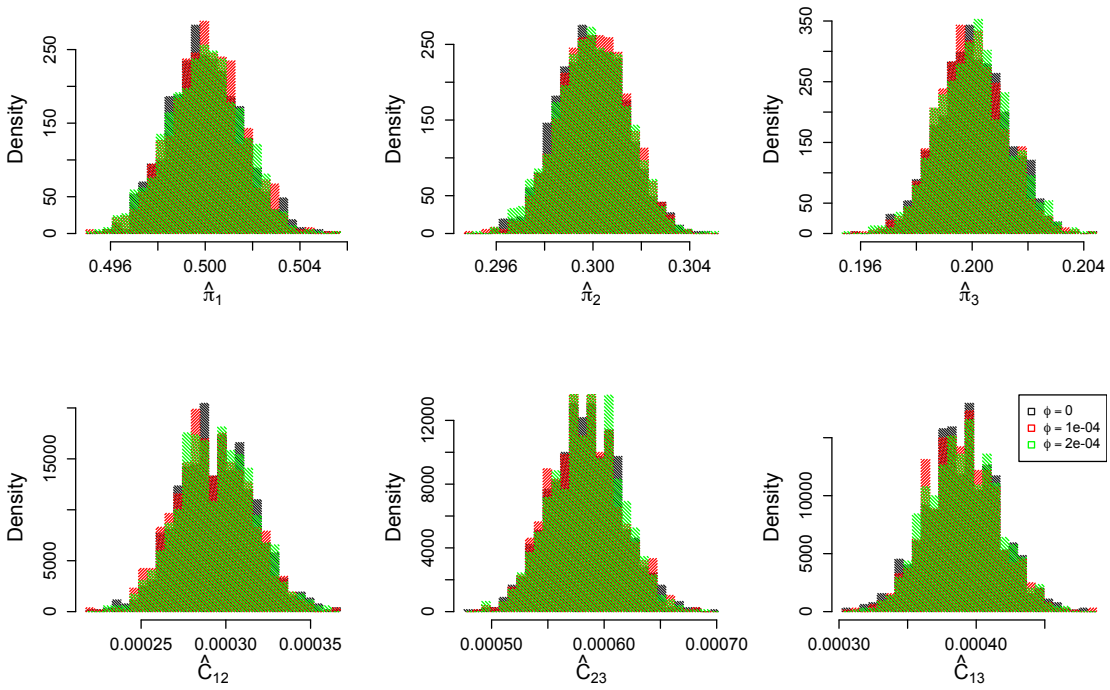


Figure 3.3: Histograms of estimates $\hat{\pi}_a$ and $\hat{C}_{ab}$ of the parameters defining the reversible part Q^{GTR} of the rate matrix for $K = 3$ alleles. True values of the parameters are $(\pi_1, \pi_2, \pi_3) = (0.5, 0.3, 0.2)$ and $(C_{12}, C_{23}, C_{13}) = (0.0003, 0.0006, 0.0004)$. 1000 independent datasets were generated for each of 3 values of the flux parameter, $\Phi = 0$, 0.0001 and 0.0002.

ratio test statistic

$$-2\left[\mathcal{L}(0|l_a, l_{ab}(1), \ldots, l_{ab}(M-1)) - \mathcal{L}(\hat{\Phi}_{ij}|l_a, l_{ab}(1), \ldots, l_{ab}(M-1))\right], \quad (3.51)$$

where $\mathcal{L}(\cdot)$ is given by Eq. (3.48), should have a chi-squared distribution with 1 degree of freedom. For each true value of Φ and each dataset a one-sided p-value was calculated from this statistic using the upper tail of the chi-squared distribution. Shown in the right-hand panel of Fig. 3.4 for each Φ-value are ordered p-values from the 1000 datasets, plotted against quantiles of a uniform distribution. For the $\Phi = 0$

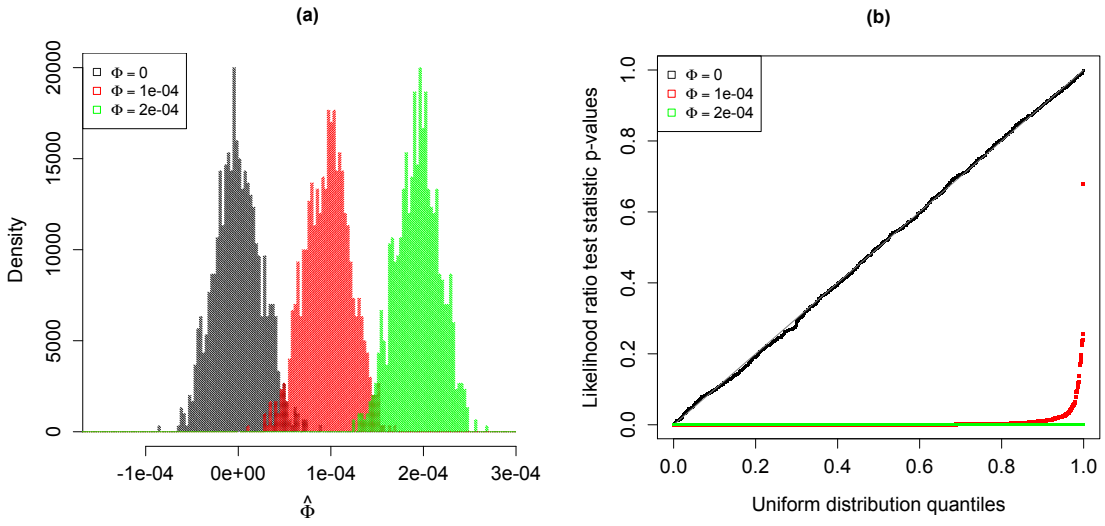


Figure 3.4: (a) Histograms of estimates $\hat{\Phi}$ of the flux parameter contributing to the non-reversible part Q^{flux} of the rate matrix for $K = 3$ alleles. True values of the parameters are as in the caption to Fig. 3.3. (b) Ordered 1-sided p-values calculated from the likelihood ratio test statistic Eq. (3.51) plotted against uniform quantiles. For each true value of Φ p-values are calculated from 1000 synthetic datasets.

datasets the p-values have a uniform distribution as required, while the p-values for the remaining two values of Φ are suitably small.

3.5.2 Synthetic data: $K = 4$ alleles

A numerical stationary solution to the neutral Wright-Fisher model with transition matrix Eq. (3.14) was created for a population size $N = 30$ for each of the following 4 rate matrices, defined by the parameters in Table 3.1:

GTR. The 4×4 general time reversible matrix has 9 independent parameters, which can be specified as the π_i and C_{ab} defined in Eq. (3.12). The first 9 parameters in the first column of Table 3.1 are chosen to reproduce very approximately the time-reversible matrix in Table 5 of Lanave et al. [1984] estimated from the rat-mouse

phylogeny. The resulting scaled GTR matrix is

$$Q^{\mathrm{GTR}} = \begin{pmatrix} -5.375 & 1.875 & 2.000 & 1.5000 \\ 2.500 & -17.500 & 0.333 & 14.667 \\ 16.000 & 2.000 & -21.000 & 3.000 \\ 2.400 & 17.600 & 0.600 & -20.600 \end{pmatrix} \times 10^{-4}. \tag{3.52}$$

Note that Lavane et al's rate matrix is a per-generation rate and differs from our scaled matrix by a factor of 10^5, which, by Eq. (3.16), corresponds to assuming the effective population size of Lanave et al's data set to be 10^5.

GRM. The most general 4×4 rate matrix has 12 parameters, which can be specified as the parameters π_i, C_{ab} and Φ_{ij} defined in Eq. (3.12). The general rate matrix Q^{GRM} in our simulations uses the same reversible part as Q^{GTR}, plus a non-reversible part using the Φ_{ij} specified in the first column of Table 3.1, which are chosen arbitrarily subject to the constraint that they should lie within the allowable range specified by Eq. (3.13). This gives

$$Q^{\mathrm{GRM}} = \begin{pmatrix} -5.375 & 3.125 & 2.125 & 0.1250 \\ 0.833 & -17.500 & 0.583 & 16.083 \\ 15.000 & 0.500 & -21.000 & 5.500 \\ 4.600 & 15.900 & 0.100 & -20.600 \end{pmatrix} \times 10^{-4}. \tag{3.53}$$

SS. A strand-symmetric rate matrix is one which is symmetric under simultaneous interchange of nucleotides A with T and C with G. Strand symmetry imposes certain constraints on the parameters listed in Eq. (3.12), namely,

$$\begin{aligned} &\pi_C = \pi_G = 0.5 - \pi_A, \\ &C_{AC} = C_{GT}, \qquad C_{AG} = C_{CT}, \\ &\Phi_{AC} = \Phi_{GA}, \qquad \Phi_{CG} = 0, \end{aligned} \tag{3.54}$$

which reduces the number of independent parameters to six, as required of a strand-symmetric matrix [Sueoka, 1995]. Most genomic sequences, when examined on a sufficiently large scale are observed to be strand-symmetric. The parameter choices π_a and C_{ab} in the right-hand column of Table 3.1 are obtained by averaging pairs of parameters constrained to be equal by Eq. (3.54). The one independent flux was then chosen arbitrarily within the range allowed by the constraint Eq. (3.13). The resulting rate matrix is

$$Q^{\mathrm{SS}} = \begin{pmatrix} -11.231 & 2.154 & 7.231 & 1.846 \\ 1.143 & -18.000 & 0.571 & 16.286 \\ 16.286 & 0.571 & -18.000 & 1.143 \\ 1.846 & 7.231 & 2.154 & -11.231 \end{pmatrix} \times 10^{-4}. \tag{3.55}$$

SSR. If the one remaining flux, namely $\Phi_{AC} = \Phi_{GA}$ which corresponds to a closed path $A \to C \to T \to G \to A$ in Fig. 3.1(b), is constrained to be zero, the resulting rate matrix is strand-symmetric and reversible. Such a matrix has five free parameters. Setting $\Phi_{AC} = \Phi_{GA} = 0$ while retaining the remaining parameters in the second column of Table 3.1 yields the strand-symmetric, reversible matrix

$$Q^{\mathrm{SSR}} = \begin{pmatrix} -11.231 & 1.385 & 8.000 & 1.846 \\ 2.571 & -18.000 & 0.571 & 14.857 \\ 14.857 & 0.571 & -18.000 & 2.571 \\ 1.846 & 8.000 & 1.385 & -11.231 \end{pmatrix} \times 10^{-4}. \tag{3.56}$$

Table 3.1: Rate matrix parameters used in $K = 4$ numerical simulations. The conventions of Eq. (3.12) are used with the indices 1 to 4 representing the nucleotides A, C, G and T respectively. The values of Φ_{ij} in the table refer to the non-reversible matrices Q^{GRM} and Q^{SS} only. For the reversible matrices Q^{GTR} and Q^{SSR} all Φ_{ij} are zero.

	GRM and GTR	SS and SSR
π_A	0.400	0.325
π_C	0.300	0.175
π_G	0.050	0.175
C_{AC}	1.5×10^{-4}	0.9×10^{-4}
C_{AG}	1.6×10^{-4}	5.2×10^{-4}
C_{AT}	1.2×10^{-4}	1.2×10^{-4}
C_{CG}	0.2×10^{-4}	0.2×10^{-4}
C_{CT}	8.8×10^{-4}	5.2×10^{-4}
C_{GT}	0.3×10^{-4}	0.9×10^{-4}
Φ_{CG}	0.15×10^{-4}	0
Φ_{GA}	-0.10×10^{-4}	0.50×10^{-4}
Φ_{AC}	1.00×10^{-4}	0.50×10^{-4}

For each of the above four rate matrices a total of 1000 synthetic datasets each sampled at 10^5 independent genomic sites were constructed using the procedure outlined for the $K = 3$ case in the previous section, except that the numerical experiment was carried out assuming a sample of $M = 8$ individuals. Maximum likelihood estimates of the parameters in the rate matrices Q^{GRM} and Q^{GTR} were calculated for each dataset using the theory of Section 3.4. Maximum likelihood estimates of the parameters of the rate matrices Q^{SS} and Q^{SSR} were calculated using analogous likelihood functions constrained by Eqs. (3.54). Histograms of the estimated parameters are plotted in Figs. 3.5 and 3.6.

In common with the $K = 3$ case, estimates of the parameters π_i and C_{ab} defining the reversible part are independent of Φ_{ij}, as expected. Also in common with the $K = 3$ case we observe that the $\hat{\pi}_i$ estimates are centred about their true value,

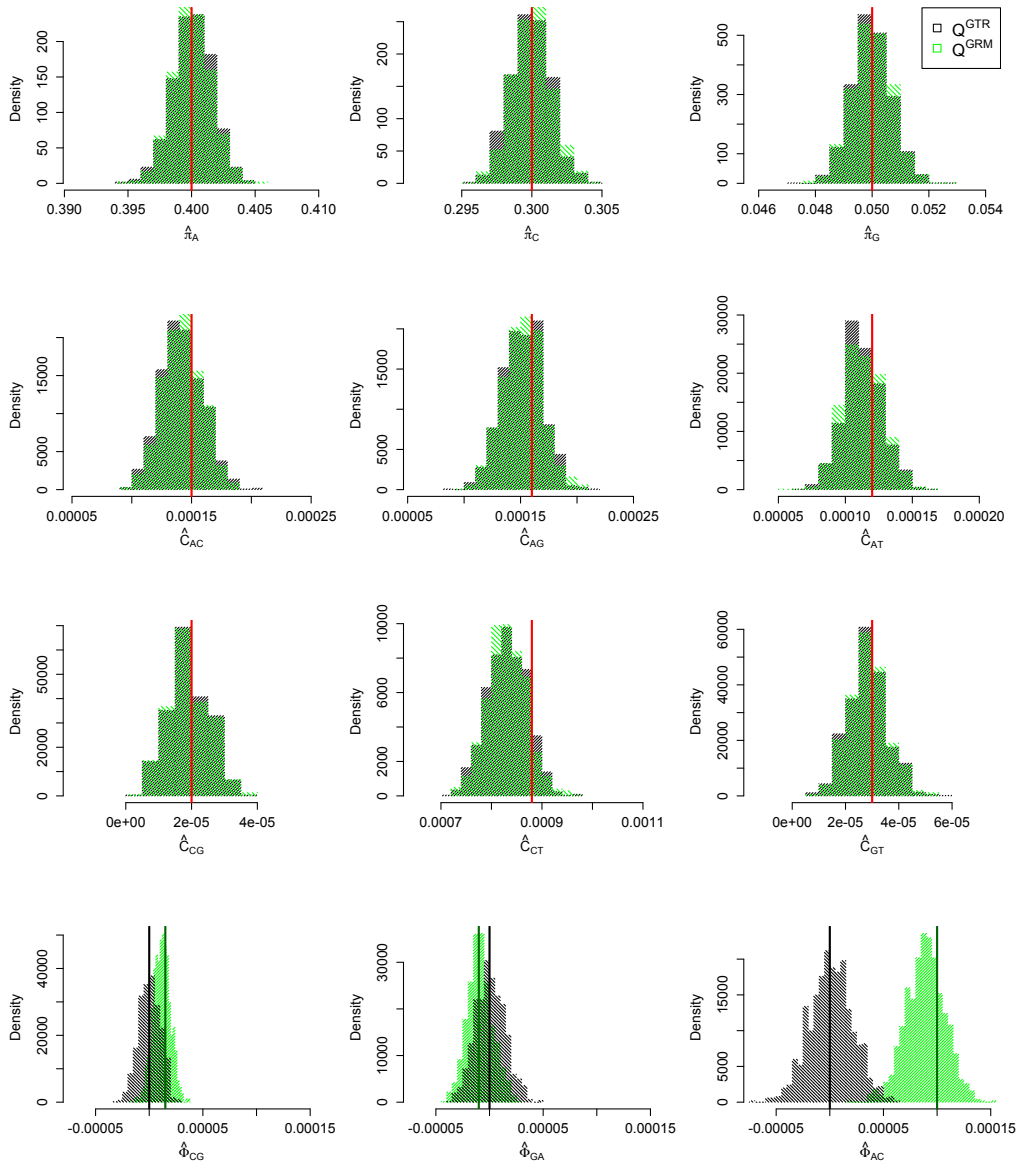


Figure 3.5: Histograms of estimates of the parameters π_i, C_{ab} and Φ_{ij} defining the rate matrices Q^{GTR} (black histograms) and Q^{GRM} (green histograms) for $K = 4$ alleles. True values of the parameters are given in the first column of Table 3.1 and are indicated by the thick vertical lines. 1000 independent datasets were generated for each rate matrix assuming the data to be sampled from $M = 8$ independently chosen individuals from a defining population of $N = 30$ individuals. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

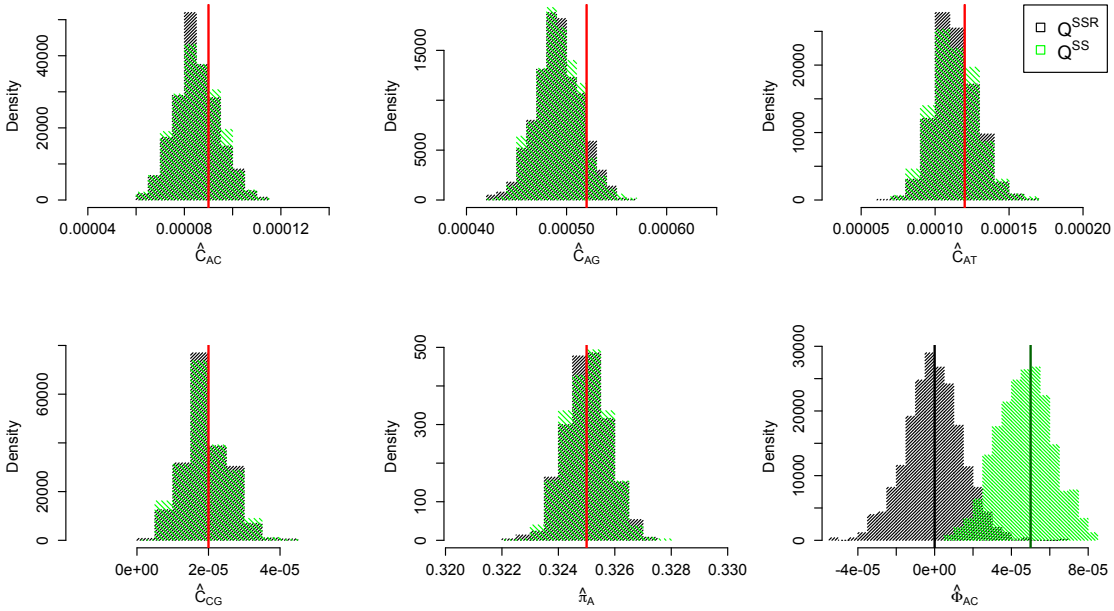


Figure 3.6: Histograms of estimates of the parameters π_A, C_{AC}, C_{AG}, C_{AT}, C_{CG} and Φ_{AC} defining the strand-symmetric rate matrices Q^{SSR} (black histograms) and Q^{SS} (green histograms) for $K = 4$ alleles. True values of the parameters are given in the second column of Table 3.1 and are indicated by the thick vertical lines. 1000 independent datasets were generated for each rate matrix assuming the data to be sampled from $M = 8$ independently chosen individuals from a defining population of $N = 30$ individuals. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

but the estimates $\hat{C}_{ab}$ tend to be low. A similar pattern was observed by Vogl and Bergman for $K = 2$ simulations with selection (see Fig. 3 of Vogl and Bergman [2015]). Estimates $\hat{\phi}_{ij}$ of the parameters determining the extent of non-reversibility are centred about their true values, or slightly skewed towards zero. Figure 3.7 shows qq-plots of 1-sided p-values calculated from likelihood ratio test statistics analogous to Eq. (3.51) against a uniform distribution, assuming the null hypothesis $\Phi_{ij} = 0$. For the otherwise unconstrained reversible and non-reversible rate matrices Q^{GTR} and Q^{GRM}, the p-values were calculated assuming a chi-squared distribution with 3 degrees of freedom for the likelihood ratio test statistic. For the strand-symmetric rate matrices Q^{SSR} and Q^{SS} a chi-squared distribution with 1 degree of freedom

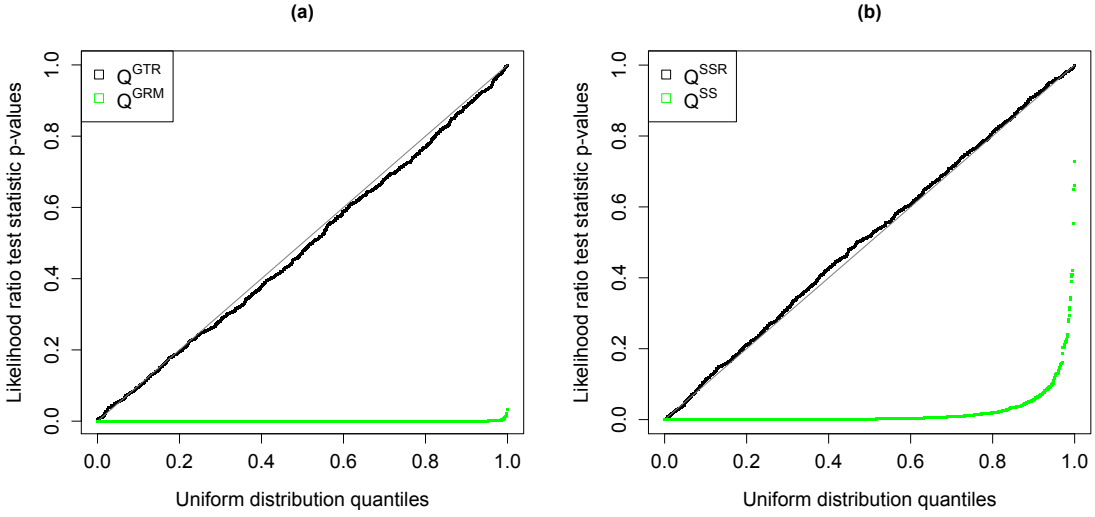


Figure 3.7: QQ plots of 1-sided p-values calculated from the likelihood ratio statistic assuming a null hypothesis $\Phi_{ij} = 0$. (a) For the matrices Q^{GTR} and Q^{GRM} likelihoods are maximised without constraints on any parameters, and p-values are calculated from the likelihood ratio statistic assuming a chi-squared distribution with 3 degrees of freedom. (b) For the matrices Q^{SSR} and Q^{SS} likelihoods are maximised under the constraints of Eq. (3.54) and p-values are calculated assuming a chi-squared distribution with 1 degree of freedom. The p-values are expected to have a uniform distribution for the reversible matrices Q^{GTR} and Q^{SSR}, which satisfy the null hypothesis, but not for the non-reversible matrices Q^{GRM} and Q^{SS}.

was assumed. For the reversible matrices the p-values have a uniform distribution as required, while the p-values for the non-reversible rate matrices are suitably small.

Finally, in order to illustrate the limits of applicability of the method, Fig. 3.8 shows parameter estimates for a rate matrix corresponding to mutation rates 10 times that of Fig. 3.6. These parameters correspond to a total mutation rate $\theta = \sum_{a \neq b} Q_{ab} \approx 0.06$. We see that mutation rates are underestimated as the approximated fails to deal properly with tri-allelic sites, but is still able to distinguish between the reversible and non-reversible cases.

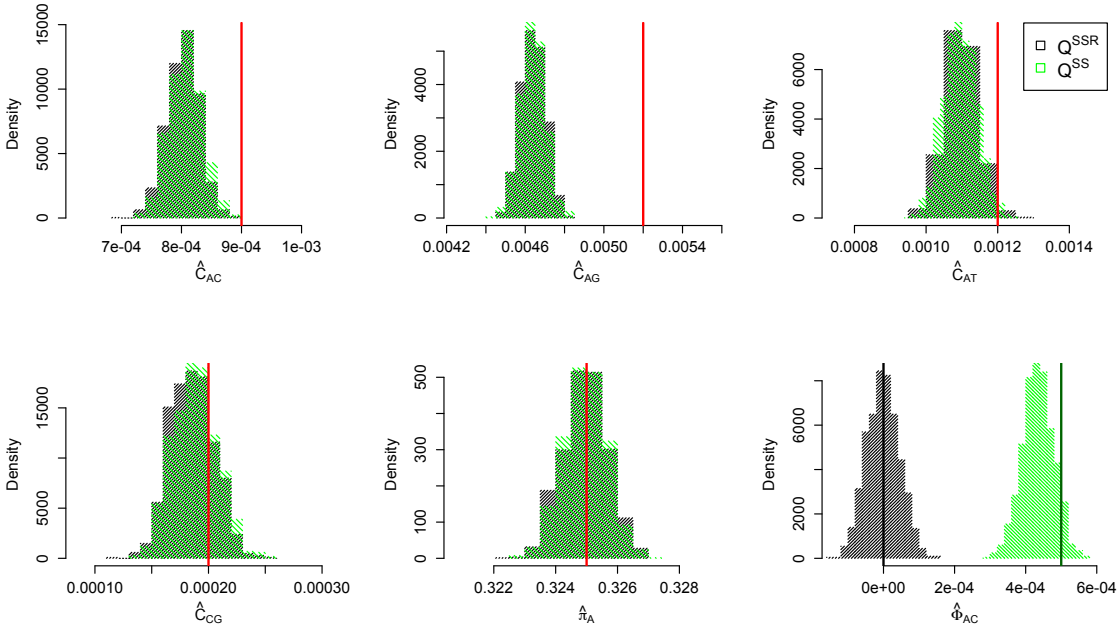


Figure 3.8: The same as Fig. 3.6, except that the rate matrix parameter values C_{ab} and Φ_{ab} are 10 times those listed in the right hand column of Table 3.1, while the π_a remain unchanged.

3.6 Discussion and conclusions

We have demonstrated that it is possible, in principle, to estimate an evolutionary rate matrix from the site frequency spectrum of an alignment of genomes sampled from a population, provided certain conditions are met. The procedure consists of a maximum likelihood estimate of all 12 parameters of the 4×4 nucleotide mutation rate matrix based on the distribution of observed single nucleotide polymorphisms at neutral sites in a multiple alignment. Given the site frequency spectrum constructed from the alignment of a large number of neutral sites obtained from sequencing a moderate number of individuals, the calculation is computationally straightforward and also provides a likelihood ratio test of the significance of the non-reversible part of the rate matrix.

We feel it is important to visit in turn each of the conditions required of our procedure to highlight the remaining challenges inherent in the approach.

Firstly we have assumed the population to have a constant size and to evolve according to the dynamics of a model whose diffusion limit is Eq. (3.17), for instance a Wright-Fisher or decoupled Moran model. Except in papers specifically addressing the point, this assumption is almost universally made implicitly throughout both the population dynamics and phylogenetics literature. In practice, though, the assumption is often violated for real datasets. It is important to recognise that a non-constant population size can alter the dynamics of genetic drift in ways which can effect the site frequency spectrum. For instance the recent explosive growth in the human population is believed to have skewed the site frequency spectrum towards rare numbers of derived alleles [Keinan and Clark, 2012]. On the other hand, if a rapidly growing population is modelled with Galton-Watson branching process, alleles present in the population at the beginning of the growth period can remain endemic in the population indefinitely without dissipating through genetic drift, even in the absence of new mutations at the spectrum boundary [Burden and Simon, 2016]. We therefore caution that, as a general principle, a constant population model may be inappropriate for estimating mutation rates from populations of rapidly varying size.

Secondly we have assumed that a large number L of sites within a genome can be identified which have evolved independently and neutrally. Promising candidates are short intron sites [Parsch et al., 2010] and 4-fold degenerate sites within codons, which are assumed to be relatively free of selective pressures. Vogl and Bergman [2015] have estimated selection effects in datasets consisting the short intron sites and 4-fold degenerate sites of a multiple alignment of 10 whole genome *Drosophila*

simulans genomes. By partitioning the set of nucleotides into two effective alleles, $\{A, T\}$ and $\{C, G\}$, they are able to reduce the problem of estimating the parameters of a $K = 2$ Wright-Fisher model with both mutation and selection included, and find clear evidence of directional selection favouring the $\{C, G\}$ state. Therefore our analysis is not suitable for this particular *D. simulans* dataset, for instance. For consistency, any analysis of a real dataset to determine the full set of 12 parameters of a 4×4 mutation matrix would first have to pass Vogl and Bergman's test of neutrality.

Thirdly we have assumed that the genomic sites considered evolve with a common rate matrix. There is clear evidence however that biochemical effects can render mutation rates context dependent [Zhao and Boerwinkle, 2002, Siepel and Haussler, 2004, Hwang and Green, 2004], that is, rates may depend on the identity of neighbouring nucleotides. A strong example of this is the effect on $C \to T$ transitions of the CpG context. Assuming the context of a neutral site to be subject to selection pressures and therefore relatively stable, one could possibly accommodate context dependence by restricting the dataset, and hence the estimated rate matrix, to genomic sites with a particular context.

Finally we mention the caveat that the estimation procedure relies on an approximate solution to the forward Kolmogorov equation valid in the limit of small scaled mutation rates, by which we essentially mean the Ewen-Watterson 'θ-parameter' of order $u_{ab}N$, where N is the population size and u_{ab} the per-generation mutation rate from allele A_a to allele A_b. As a rule of thumb, the approximate solution agrees well with numerical simulations of the neutral Wright-Fisher model provided $\theta < O(10^{-2})$ [Burden and Tang, 2016].

Chapter 4

The Inherited Rate Matrix algorithm for phylogenetic model selection for non-stationary Markov processes

Yurong Tang, Benjamin Kaehler, Hua Ying, Gavin Huttley

This chapter has been in preparation for PeerJ. The reference for the publication is:

Tang Y, Kaehler B, Ying H, Huttley G. The Inherited Rate Matrix algorithm for phylogenetic model selection for non-stationary Markov processes. PeerJ

My Contributions For this paper, my supervisor Professor Gavin Huttley and I conceived and designed the algorithms and the experiments. Dr Benjamin Kaehler and Dr Hua Ying reviewed the manuscript and gave some modification suggestions for the manuscript. During the project, I conducted the research, designed the algorithms, wrote the software and analysis scripts, and drafted the manuscript under the supervision of Professor Gavin Huttley.

Acknowledgments We thank Teresa Neeman for suggesting the McNemar test and Helmut Simon for comments on an earlier version of the manuscript.

Supervisor's signature: 

ABSTRACT In phylogenetic reconstruction, substitution models that assume nucleotide frequencies do not change through time are in very widespread use. DNA sequences can exhibit marked compositional difference. Such compositional heterogeneity implies sequence divergence has arisen from substitution processes that differ between the lineages. Accordingly, there has been considerable interest in exploring the non-stationary model class, which is capable of generating divergent sequence composition. These models are typically parameter rich and this richness is further compounded by the use of time-heterogeneous variants in order to capture compositional heterogeneity. This illuminates two barriers limiting the widespread application of the non-stationary model class: both the compute-time required to fit a model and the risk of model over-fitting increase with the number of parameters and the extent of time-heterogeneity. In this study, we address these issues with a novel model selection algorithm we term the Inherited Rate Matrix algorithm (hereafter IRM). This approach is based on the notion that a species inherits the substitution tendencies of its ancestor. We further present the non-stationary heterogeneous across lineages model (hereafter ns-HAL algorithm) which extends the HAL algorithm of [Jayaswal et al., 2014] to the general nucleotide Markov process. The IRM algorithm substantially reduces the complexity of identifying a sufficient solution to the problem of time-heterogeneous substitution processes across lineages. We also address the issue of reducing the computing time with development of a constrained-optimisation approach for the IRM algorithm (fast-IRM). Our algorithms are implemented in Python 3. From a simulation study based on 2nd codon position genome sequences of yeast, we establish that IRM is significantly more accurate than both ns-HAL and heterogeneity in the substitution process across lineages (hereafter HAL) for close and dispersed sequences. IRM is up to 70-times faster than ns-HAL while fast-IRM is around 140× faster. fast-IRM also showed a

marked speed improvement over a C++ implementation of HAL. Our comparison of the accuracy of IRM with fast-IRM showed no difference with identical inferences made for all data sets. These two algorithms greatly improve the compute time for model selection of a non-stationary process, increasing the suite of problems to which this important substitution model class can be applied.

4.1 Introduction

Phylogenetic reconstruction methods are widely employed and the results are widely debated. A multitude of studies have shown that the results are sensitive to the taxa included [e.g. Rokas et al., 2003b,a, Reddy et al., 2017, Pisani et al., 2015] and that the tendency to generate conflicting trees is affected by choice of substitution model [e.g. Huttley, 2009]. Such discrepant results are likely to arise from model misspecification (incongruence between the assumptions made by the methods and the properties of the biological data to which they are applied). Biological sequence data arises by complex processes that act on species at the molecular and population level. The methodological component responsible for capturing that complexity is the substitution model and thus choice of suitable substitution model is crucial to the robustness of phylogenetic inference. Non-stationary substitution models are able to represent divergence in sequence composition. This class of models have been demonstrated to fit sequence data extremely well in comparison to the widely employed time-reversible substitution models [Kaehler et al., 2015]. The superior ability to describe the data of non-stationary models may cause over-fitting. Kaehler et al. [2015] implemented bootstrap tests to efficiently identify over-fitting. In this work, we develop a new method aimed at reducing the risk of over-fitting.

At present, stationary substitution models are employed nearly exclusively. These

models have the advantage of being relatively simple with a small number of parameters. However, they incorrectly assume that the composition of nucleotides, or amino acids, remains roughly the same across a tree [Felsenstein, 2003]. It has been demonstrated that violation of this assumption results in phylogenetic errors [e.g. Foster, 2004].

It has been known for decades that homologous DNA sequences can exhibit marked compositional difference in the relative abundance of the four nucleotides [Macaya et al., 1976, Jukes and Bhushan, 1986, Karlin and Mrázek, 1997]. Such observations indicate that the evolutionary process producing these sequences is not stationary; the sequence composition changes over time. This property has been the motivation for development of substitution models that can accommodate this non-stationarity. The mechanistic basis of this divergence in sequence composition can include contributions from both natural selection and changes in mutagenesis. For the sake of simplifying the arguments here, we consider only the influence of changes in mutation. It has been demonstrated that these models can give a markedly better description of sequence evolution than stationary ones [Kaehler et al., 2015]. However, a disadvantage of non-stationary models is they are very parameter rich compared to their time reversible counterparts.

Increasing the number of parameters can improve model fit. However, such increases in the number of parameters also makes a model prone to over-fitting [Burnham and Anderson, 2002] where the extra parameters fit noise in the data, inflating bias and causing errors in inference. In phylogenetic reconstruction, the assumption of heterogeneity in the substitution process across lineages is supported by evidence of compositional heterogeneity between the sequences [Jayaswal et al., 2014]. But an unrestricted accommodation of heterogeneity across lineages may contribute to over-

parameterising the model [Jayaswal et al., 2014]. A sensible reduction in parameter number can be achieved by restricting the number of rate matrices. However, the total number of possible ways of reducing heterogeneity among lineages is enormous for even a modest number of taxa [Jayaswal et al., 2011]. Therefore, an efficient strategy for model selection is required to identify an "optimal" time-heterogeneous model for a data set.

Multiple approaches for representing time-heterogeneous processes have been proposed. The most extreme possible choice is the most general model which allows a distinct non-stationary process for each edge of a phylogenetic tree [Jayaswal et al., 2011, Zou et al., 2012]. Another strategy is a distinct vector of equilibrium nucleotide frequencies on each edge of the tree [Yang and Roberts, 1995, Foster, 2004, Zou et al., 2012]. The former naturally represents non-stationary models but is more prone to over-fitting due to the larger number of free parameters. The latter is a composite of a time-reversible parameterisations with explicitly imposed shifts in the stationary distribution nodes of the tree, producing a non-stationary model (hereafter we term such models as pseudo non-stationary models). These models will also be susceptible to over-fitting.

One model complexity reducing approach employs a top-down algorithm to identify optimal and near-optimal time-heterogeneous models [Jayaswal et al., 2011]. The algorithm starts with the most general model (a distinct rate matrix on every edge) and then gradually decreases the model complexity until a stop condition is satisfied, producing a near-optimal model. This approach is only suitable for a small number of taxa. Subsequently, a bottom-up algorithm accounting for heterogeneous across lineages model (HAL) was developed [Jayaswal et al., 2014]. This algorithm starts with the simplest model and sequentially increases the model complexity until a

stop condition is satisfied. In their approach, Jayaswal et al. [2014] employed a non-stationary model based on General Time Reversible rate matrices [Lanave et al., 1984]. While the algorithm is comprehensive in terms of model reduction, it still has a massive solution space that makes its application impractical.

One notable attribute of the HAL algorithm is that disjoint edges can share the same rate matrix, which is distinct from the rate matrices operating on the intervening edges. This corresponds to a biological case of convergent evolution, whereby mutation and natural selection operate identically on edges that are not adjacent on the phylogenetic tree. While convergent evolution is plausible, it is a "special case" and not representative of the general tendencies of divergence between biological lineages.

Based on the limitations of the above approaches, it is apparent that new model selection strategies are required in order for non-stationary Markov processes to become a practical addition to the molecular evolutionary and phylogenetic toolkits. Systematic changes to mutation are important drivers for sequence substitution processes and changes to sequence composition [e.g. Karlin and Burge, 1995, Knight et al., 2001]. Mutation tendencies are determined, to a substantial degree, by DNA modification and repair systems which are transmitted from parent to offspring. Accordingly, it seems reasonable to conjecture that lineages inherent their mutation tendencies (and thus substitution processes) from their immediate ancestor. Thus, lineages whose immediate ancestor is the same are most likely to exhibit the same substitution process. For instance, gene loss(es) in some lineages have led to complete absence of the hypermutable base 5-methyl-cytosine [Albalat et al., 2012]. Descendants of those lineages thus exhibit a significant shift in the relative abundance of the different point mutations. Motivated by this conjecture, we

propose a bottom-up Inherited Rate Matrix (IRM) algorithm for selection of time-heterogeneous non-stationary models.

Here we report our evaluation of the question of whether a model selection strategy based on inheritance of rate matrices can prove sufficient in model selection of non-stationary models. IRM algorithm coupled with an efficient tree traversal mechanism are described. We further present a non-stationary HAL algorithm (ns-HAL) based on the non-stationary general Markov nucleotide substitution model [Kaehler et al., 2015]. We show via simulation that IRM is fast and less prone to model misspecification than ns-HAL. The source codes of the methods presented in this paper are made available at `http://doi.org/10.5281/zenodo.3754681`. The raw data, results and scripts used for analyses are available at `http://doi.org/10.5281/zenodo.4361748`.

4.2 Materials and methods

4.2.1 Data used in this study

We used the same data set as that employed by Jayaswal et al. [2014]. This data set consists of the concatenation of second codon positions from 106 multiple sequence alignments of orthologs from eight yeast species. The resulting concatenated alignment is 42,337bp long. The species, and their corresponding abbreviation are: *Saccharomyces cerevisiae* (Scer), *Saccharomyces paradoxus* (Spar), *Saccharomyces mikatae* (Smik), *Saccharomyces kudriavzevii* (Skud), *Saccharomyces castellii* (Scas), *Saccharomyces kluyveri* (Sklu), *Saccharomyces bayanus* (Sbay), and *Candida albicans* (Calb). A phylogenetic tree for these species reported by Rokas et al. [2003b]

is shown in Figure 4.1. The alignment, a concatenation of alignments of one-to-one orthologs were determined by Rokas et al. [2003b], was employed by Jayaswal et al. [2014] in development of the HAL algorithm.

In order to decrease running time of ns-HAL, we selected two 5-species subsets from the data. The "5-close" subset consists of the five most closely related lineages Scer, Spar, Smik, Skud and Sbay. The "5-dispersed" subset consists of five more dispersed lineages Scer, Skud, Scas, Sklu, Calb. The topologies of 5-close data set and 5-dispersed data set are subgraphs from the topology of Figure 4.1. We expected that the extent of divergence between a collection of taxa would influence power to resolve different levels of time-heterogeneity. The 5-close and 5-dispersed subsets were chosen to facilitate the examination of this hypothesis. How this was done is explained in the section on statistical evaluation.

To assess the scalability of IRM and fast-IRM for real data we applied both to mitochondrial DNA sequences from 50 anthozoans. The mitochondrial genomes of Anthozoa have an extremely slow rate of evolution [Kitahara et al., 2014, Huang et al., 2008]. The data set was chosen due to evidence that these species differ in the level of mitochondrial DNA repair, attributable in part due to the presence of the MutS gene in the octocorals [Pont-Kingdon et al., 1995, Culligan et al., 2000] and inference of reduced DNA repair in the scleractinia [Kitahara et al., 2014]. The 50 taxon included 31 scleractinia, 12 corallimorpharia, 2 sea anemones, single antipatharia and zoanthid species, and 3 octocorals. We use a tree reported by Kitahara et al. [2014] that built using codonphyml [Gil et al., 2013, Yap et al., 2010]. Here we analysed the second codon position of the mitochondrial protein coding DNA sequences as a concatenated alignment with 3934 sites.

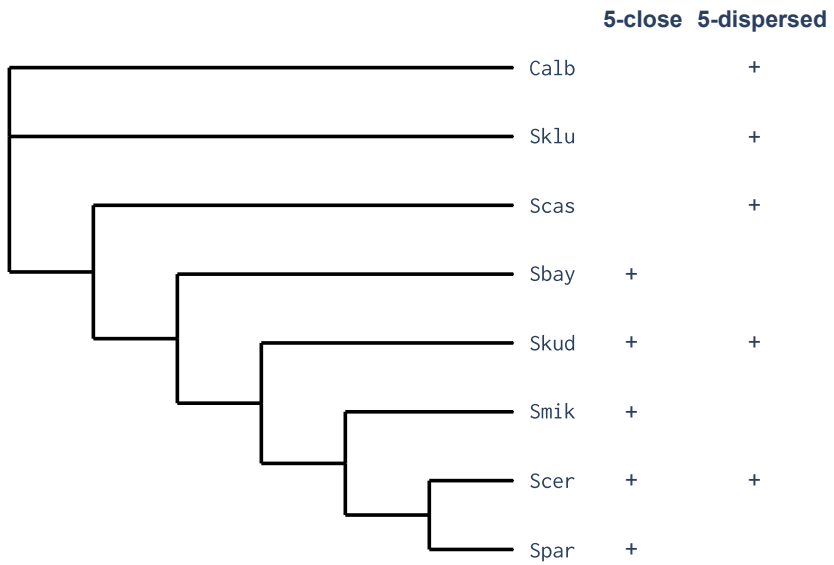


Figure 4.1: The phylogeny of the 8 Yeast species from Rokas et al. [2003b]. The eight species are *S. cerevisiae* (Scer), *S. paradoxus* (Spar), *S. mikatae* (Smik), *S. kudriavzevii* (Skud), *S. castellii* (Scas), *S. kluyveri* (Sklu), *S. bayanus* (Sbay), and *C. albicans* (Calb). Membership of the 5-close species subset {Scer, Spar, Smik, Skud, Sbay} and 5-dispersed species subset {Scer, Skud, Scas, Sklu, Calb} are shown in the corresponding columns using the '+' symbol.

4.2.2 General Markov nucleotide substitution model

The first non-stationary general Markov nucleotide substitution model to be studied was that of Barry and Hartigan [1987] (hereafter referred to as the BH model). This is a discrete-time, non-stationary Markov process for nucleotides that has 12 different parameters. The most general continuous-time Markov process on nucleotides [hereafter GN Kaehler et al., 2015] is equivalent to BH if the Markov process is embeddable [Verbyla et al., 2013]. In GN, all possible exchanges receive their own

rate parameter, thus there are 12. In PyCogent3 [Huttley, 2019], the rate matrix is calibrated such that $-\Sigma_i \pi_i Q_{ii} = 1$. For a stationary process, this is the expected number of substitutions in a unit time interval and, as a result, the scalars applied per branch are exactly the expected number of substitutions per site for that branch. For a non-stationary process, π changes through time and thus that relationship no longer holds. This choice of calibration is a convenience for the purpose of most model fitting and the correct expected number of substitutions for each branch can be derived by post fitting [Kaehler et al., 2015]. For GN model developed by Kaehler et al. [2015], the substitution number, genetic distance and ENS distance are the same. Therefore, the above three terms described in this paper are the same.

Sufficient conditions under which the BH model is identifiable were established by Chang [1996] and we refer the reader to that article for a complete exposition of these properties. One sufficient condition is that rate matrices exhibit the property of diagonal largest in column (hereafter DLC). Identifiability of the continuous-time general Markov nucleotide counterpart to BH [Kaehler et al., 2015] requires in addition that there is a unique mapping between $P(t) = \exp^{Qt}$ [Kaehler et al., 2015]. Hence, we employed two identifiability checks: (i) the estimated processes exhibit DLC, and (ii) a unique mapping exists [Kaehler et al., 2015]. Solutions from numerical optimisation that failed these checks were discarded.

4.2.3 Identifying the direction of "time"

As shown by Chang [1996], the general Markov process cannot identify the direction of time. Stated another way, such a process is not identifiable for an unrooted tree of three taxa. This makes it problematic to identify descendants from an edge, a key prerequisite for IRM and ns-HAL. Of the possible approaches to solve this, we

used the midpoint method as it is computationally efficient (needs to be done only once) and was an existing PyCogent3 tree capability. In essence, we orient the tree such that the "root" node is as close to being the chronological root as possible. Once this is done, the assumption that children of a node are its actual descendants seems reasonable.

4.2.4 ns-HAL: HAL algorithm for non-stationary Markov processes

In our ns-HAL algorithm, we made the following modifications to the HAL algorithm of Jayaswal et al. [2014]:

- *Use the GN model* Jayaswal et al. [2014] employed a non-stationary model based on the time-reversible GTR process. It achieved non-stationarity by associating distinct GTR matrices with different stationary frequency vectors (π). We do not need to do this as the general Markov nucleotide (GN) process is naturally non-stationary [Kaehler et al., 2015].

- *Use an unrooted tree* Jayaswal et al. [2014] specified HAL for a strictly binary tree, hence its root node has two branches. As a GN model with a rooted tree is not identifiable, we used an unrooted tree.

- *Decrease model complexity using Frobenius norm [Golub and Van Loan, 2013, p. 70]* In order to avoid overparameterization, Jayaswal et al. [2014] applied a method of decreasing model complexity [the CORE algorithm, Jayaswal et al., 2011] to determine whether the first model in set $\mathbb{S}$ produced by HAL is overparameterized. In their method, clustering of π was used to guide

model complexity reduction. As GN does not require distinct π for a time-heterogeneous non-stationary process, we used the Frobenius norm between the two closest rate matrices, combining them into one rate matrix so that a simpler model is obtained.

- *Different model selection criteria* Jayaswal et al. [2014] used Akaike information criteria (AIC) and Bayesian information criterion (BIC) for selecting models. In addition to the above two information criteria, we also evaluated using AIC correction ($AICc$) for model selection as it has a stronger penalty for the number of parameters than AIC.

- *Different iteration stop condition* In Jayaswal et al. [2014], the iteration stop condition is controlled by the size of a set S. In each iteration, a fixed number of candidate models are put into this set. In ns-HAL, candidate models can be selected through two methods: (i) use a fixed number of candidate models as done by Jayaswal et al. [2014], (ii) choose candidate models according to $AICc$ difference.

- *Identifiability check* Tests of the two identifiability conditions, DLC and Unique mapping [Kaehler et al., 2015] are performed to check if the rate matrix is identifiable. Jayaswal et al. [2014] do not perform this step.

4.2.5 IRM: Inherited Rate Matrix algorithm for non-stationary Markov processes

Motivated by the greater likelihood of lineages to inherent the substitution properties of their immediate ancestor, we impose the restriction that only adjacent edges can share a rate matrix. A phylogenetic tree is an acyclic connected graph in which any

two nodes are connected by exactly one path constituted by some joint edges. In order to guarantee that all edge set partitions consist of adjacent edges, we adapted a graph cut algorithm.

In graph cut theory, a cut is a partition of the nodes (vertices) of a graph into two disjoint subsets. An important and relevant partition algorithm is that of "minimum cut". The minimum cut algorithm partitions nodes of graph into two sets according to the weight of edges, choosing the edge with the smallest weight. Our application here is different, as we seek to obtain a partition of the edges of a graph instead of a partition of the nodes. Therefore, we need to cut edges into two groups from one node (we denote this node as the cutting node). The notion of a weight is used to determine the order of cutting nodes.

We developed a variant algorithm of minimum cut, which is illustrated in Figure 4.2. In it, we choose cutting nodes according to weights of branches and then partition edges into two groups. We define the weight as $W_{ij} = \frac{1}{d_{ij}}$, where d_{ij} is the genetic distance from the node i to the node j. We choose the cutting node by a genetic distance based weight for the following reasons. The genetic distance is a quantity derived explicitly from the rate matrix. It is related to the amount of information in the data regarding a specific edge. Finally, the probability of a change in substitution process, and hence the rate matrix, increases with increasing chronological time. Our choice of edge weight should therefore identify the edge that is most likely to have evolved in a manner described by a distinct rate matrix.

The implementation of our variant minimum algorithm includes two steps. First, with the edges ranked in ascending order by their weight (For example edge ID 1 and 2 are ranked as the first and second ones respectively), an iterative process is then applied. Parent node of the edges having the smallest weight is then chosen

as the cutting node. The cutting node defines two sets of edges. Edges within a set share the same rate matrix. (Note that all edges retain an independent scalar parameter that is monotonic to genetic distance [Kaehler et al., 2015].) In cases where there are ties in weight, all edges with the equivalent weight are evaluated. Second, an iteration stop condition is evaluated for every new choice of cutting node. An information theoretic measure, such as AIC, is employed. Simplified Python implementations of the core IRM algorithm and associated utility functions are presented in Algorithms S1-S3 of Supporting Information.

4.2.6 fast-IRM: IRM with constrained optimisation

The computational speed of algorithms for model selection depends on whether full numerical optimisation is used for evaluating each candidate model. In a full numerical optimisation, all free parameters are subject to modification during the optimisation. For a general nucleotide Markov model, this corresponds to 11 free parameters per rate matrix and a unique branch scalar for each edge that is monotonic to the expected number of substitutions [Kaehler et al., 2015]. In both the IRM and ns-HAL cases, maximum likelihood estimates from the parent (simpler) model were used as initial values for each derived model. Here, a derived model was generated from a parent model according to the model selection algorithms. In order to further speed the algorithm, we developed a constrained-optimisation variant of IRM that we denote fast-IRM.

In fast-IRM reduces compute time by decreasing the number of parameters optimised at each step. When a clade is split into two at the cutting node, the two resulting clades are represented by 2 rate matrices. In IRM, the entire model is

numerically optimised. Hence, the number of free parameters to be numerically optimised increases with each iteration. In fast-IRM, only parameters relating to the 2 new rate matrices are optimised whilst all other parameters are fixed as constant at the parent model optimised values. For instance, the cutting node that defines Model 3 in Figure 4.2 splits the blue (dot) clade of Model 2. The black (dash) edges in Model 3 lie outside of this and thus their parameters are fixed constant at the values from the red clade of Model 2 (black edges in Figure 4.2). The number of free parameters at each iteration of fast-IRM is therefore fixed as 2× the number per rate matrix plus the total number of edges in the clade being split (6 in Model 3, Figure 4.2).

For the GN model, where each additional rate matrix introduces 11 new rate terms, the numerical optimisation performed at each iteration through fast-IRM is limited to 22 rate terms plus a branch scalar per edge within the clade. Thus, for Model 3 in Figure 4.2, there are 28 free parameters. The fast-IRM algorithm is an approximate likelihood method and thus the convergence property of maximum likelihood is not guaranteed.

4.2.7 Size of the solution space

We contrast the solution space size (number of possible models) for 5-taxa for ns-HAL, IRM and the theoretical maximum. For the two algorithms, we determined the maximum number by exhaustive evaluation using the implemented algorithms. The theoretical maximum is a Bell number [Jayaswal et al., 2014]. The solution space size for differing values of k is shown in Table 4.1. A $k = 1$ corresponds to a model in which there is a single rate matrix for which there is only one possible model. When $k = 7$, every edge has a separate rate matrix for which there is

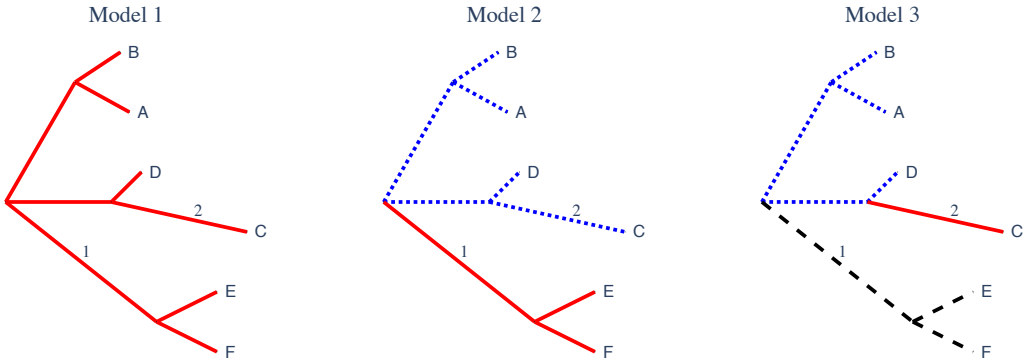


Figure 4.2: An example of the IRM and fast-IRM algorithms. Model 1 is the parent model of Model 2 which is the parent of Model 3. In each panel, edges with the same colour share a rate matrix. There is one rate matrix for Model 1, 2 rate matrices for Model 2 and 3 rate matrices for Model 3. The algorithm compares the fit of Model 2 to that of Model 1 using a nominated statistic like $AICc$. It accepts a model if that fit exceeds a minimum threshold and proceeds to consider the next model. fast-IRM is distinguished from IRM by making the parameters from the parent model constant to reduce numerical optimisation time. For instance, in Model 3, the edges coloured black (dash) are constrained to be constant at the estimated values of the parent (Model 2). All other parameters are free to be optimised.

only one possible model. In these two extreme cases, all algorithms have the same solution space size. For $k \in \{2, \cdots, 6\}$ the number of possible solutions for ns-HAL grows substantially due to possible ways of combining edges. In contrast, because the IRM algorithm uses mincut only a single additional edge can be considered at each iteration through possible values of k. The exception occurs if there are tied branch length values, in which case the solution space for IRM will equal the number of tied values.

k	Maximum	ns-HAL	IRM
1	1	1	1
2	63	9	1
3	301	38	1
4	350	75	1
5	140	50	1
6	21	13	1
7	1	1	1

Table 4.1: Comparison of the number of possible models to be evaluated for a 5-taxa tree. k is the number of rate matrices, Maximum is the maximum size of the solution space (a Bell number) for k, numbers under the IRM and ns-HAL columns correspond to the solution size determined by exhaustive search using our implementation of the algorithms.

4.2.8 Statistical evaluation of IRM, ns-HAL and HAL

We calculated the accuracy with which the IRM, ns-HAL and HAL algorithms correctly identified the generating model. We indicate the model used to generate the simulated alignments as the *true model*. Specifically, the true model is characterised by the set of k distinct rate matrices and the k edge sets, denoted $\mathbb{E}_k$, to which they map. The model with n rate matrices and the edges to which they map (i.e. $\mathbb{E}_n$) identified by an algorithm on a simulated alignment is referred to as the *inferred model*, or *estimated model*. We consider an inferred model correct when $\mathbb{E}_k = \mathbb{E}_n$. In words, the inferred model was "correct" if it found the same partition of the tree into edge sets as that used to simulate the alignment. The relative performance of each algorithm was established by comparing the total count, obtained across all model conditions, of correctly inferred models.

Simulated alignments were generated based on a fit of the GN model to the yeast data. The specification of time-heterogeneity was done in a manner consistent with the assumption of rate matrix inheritance. The model with a single rate matrix is

the first fitted model. Then the second fitted model was generated where the edges descending from one edge with the biggest genetic distance has one rate matrix where all the other edges have a different rate matrix. The process, choosing an edge according to genetic distance order and a distinct rate matrix is assigned to the edges descending to the chosen edge, continue until the last model with the maximum number rate matrices (k) was generated. A 5-species tree has 7 edges so the maximum number of k is 7. For each value of k, all possible time-heterogeneous models consistent with IRM were fit to both the 5-close and 5-dispersed yeast data sets. From each such fitted model, 10 simulated alignments with the same length with fitted model were generated. The IRM, ns-HAL and HAL algorithms were then applied to each generated model.

As each algorithm was applied to the same simulated alignments, we have a paired study design. We illustrate how algorithms were compared using IRM and ns-HAL. The paired design allows producing a 2×2 contingency table with columns corresponding to whether the IRM algorithm correctly recovered the generating model, or not. The rows correspond to the equivalent outcome for the ns-HAL algorithm.

4.2.9 Software implementation

The algorithms described above were implemented in Python 3.7 or greater and Numpy 1.18.1. Core functions for likelihood calculation and numerical optimisation are provided by PyCogent3. Implementation accuracy was validated using unit tests and these tests covered ∼86% of the code [Batchelder, 2019]. Version 1.09 of HAS-HAL, written in C++ [Jayaswal et al., 2014, Wong, 2017] was used. The source codes of the methods presented in this paper are made available at `http://doi.org/10.5281/zenodo.3754681`. The raw data, results and scripts used for

analyses are available at `http://doi.org/10.5281/zenodo.4361748`.

4.3 Results

We briefly reiterate the experimental design here. We denote the number of rate matrices in a phylogenetic model as k. For a phylogenetic tree of 5 taxa, and 7 edges, k can be at minimum 1, and at maximum 7. We generated synthetic data for each possible value of k for the 5-close and 5-dispersed yeast data sets. These data were used to contrast the accuracy of the algorithms, their computational speed and the consistency between the IRM and fast-IRM variants.

4.3.1 IRM is more accurate than ns-HAL and HAL

For a specific simulated alignment, we considered an algorithm accurate if it correctly recovered the configuration of rate matrices corresponding to the generating model (i.e. $\mathbb{E}_k = \mathbb{E}_n$). For assessing the accuracy calculation of HAL, we used AIC information criteria based on the same simulated alignments used by ns-HAL and IRM. Table 4.2 indicates the accuracy on the 5-close and 5-dispersed species data, respectively. The accuracy of HAL is is better than ns-HAL when $k = 2$, $k = 5$ and $k = 6$ for 5-close. However, the accuracy of ns-HAL is better than HAL when $k = 1$ and $k = 3$. In total, ns-HAL has nearly same accuracy with HAL for 5-close. For 5-dispersed, ns-HAL is better than HAL for all the k. IRM exhibited higher accuracy than both HAL and ns-HAL when k is 1, 2, 3, 4 and 7 on 5-close species. For k values from 5 to 6, IRM was at least as good as ns-HAL. The three algorithms had low accuracy for k vales from 4 to 7 for 5-close. There was a tendency on the 5-close data set (which has more closely related sequences) for the algorithms to

identify optimal models with few rate matrices. For 5-dispersed species, IRM had systematically higher accuracy than both HAL and ns-HAL.

	5-close			**5-dispersed**		
k	**HAL**	**ns-HAL**	**IRM**	**HAL**	**ns-HAL**	**IRM**
1	6	8	9	8	10	10
2	6	5	9	5	8	10
3	2	3	7	7	10	10
4	0	0	1	8	10	10
5	2	0	0	5	9	10
6	2	0	0	3	8	10
7	0	0	1	1	4	10

Table 4.2: Accuracy comparison between the IRM, ns-HAL and HAL algorithms. 5-close and 5-dispersed are the different data sets. k is the number of rate matrices used to generate the 10 replicate data sets. Numbers correspond to the counts of when the inferred models returned respectively by IRM, ns-HAL and HAL were correct.

In order to verify if IRM is significantly different to ns-HAL, we applied the continuity corrected version of the McNemar test. Two 2×2 contingency tables were generated on the 5-close data set and 5-dispersed data set (Table 4.3). These contrast the ability of the two algorithms to recover the correct model. For both the 5-close and 5-dispersed data sets, we rejected the null hypothesis that the two algorithms were equivalent with p-values of 0.0026 and 0.0026 respectively. Given its systematically higher accuracy (Table 4.2), we conclude that IRM is significantly more accurate than ns-HAL. Our comparison of ns-HAL and HAL (Table 4.4) indicated that ns-HAL is as accurate as HAL for 5-close and is significantly more accurate than HAL for 5-dispersed.

We evaluated the potential contribution of differing statistical power to the difference of accuracy between 5-close and 5-dispersed. The relationship among the 5-close taxa indicates there will be less genetic distance separating these sequences. As the model is being fit to substitution events, a smaller number of events may impact

		IRM		
		True	**False**	**Total**
ns-HAL	**True**	16	0	16
	False	11	43	54
	Total	27	43	70

(a) 5-close

		IRM		
		True	**False**	**Total**
ns-HAL	**True**	59	0	59
	False	11	0	11
	Total	70	0	70

(b) 5-dispersed

Table 4.3: Comparison of algorithm accuracy for IRM and ns-HAL. As each algorithm was applied to exactly the same simulated data we have a paired study design. From the 70 simulated alignments we record in the 2×2 table how often the algorithms were in agreement. Here, "True" means the algorithm recovered the correct model, "False" indicates it did not. Using the McNemar test with correction, we reject the null hypothesis of no difference between the algorithms. The p-values for both 5-close and 5-dispersed were ~ 0.0026.

		ns-HAL		
		True	**False**	**Total**
HAL	**True**	12	6	18
	False	4	48	52
	Total	16	54	70

(a) 5-close

		ns-HAL		
		True	**False**	**Total**
HAL	**True**	36	1	37
	False	23	10	33
	Total	59	11	70

(b) 5-dispersed

Table 4.4: Comparison of algorithm accuracy for ns-HAL and HAL. As each algorithm was applied to exactly the same simulated data we have a paired study design. From the 70 simulated alignments we record in the 2×2 table how often the algorithms were in agreement. Here, "True" means the algorithm recovered the correct model, "False" indicates it did not. Using the McNemar test with correction, we reject the null hypothesis of no difference between the algorithms. p-values were ~ 0.75 and $\sim 1.81e-05$ for 5-close and 5-dispersed, respectively.

on resolution. We determined the difference between the two data sets for the two algorithms in the estimated total genetic distance. As shown in Table C.2 of Supporting Information, there is roughly an order of magnitude more events in the 5-dispersed data compared with 5-close. The likely impact of such a difference on statistical power seems the most plausible explanation for the difference accuracy between the data sets.

4.3.2 Accuracy of fast-IRM

In order to assess the efficacy of the fast-IRM constrained-optimisation approach we contrasted its accuracy with that of IRM. In our implementation, it is in fact full optimisation when k is 1 and 2. Therefore, we only compared the accuracy when $k \in \{3, 4, \ldots 7\}$. Table 4.5 shows that constrained-optimisation and full-optimisation have the same accuracy for both 5-close and 5-dispersed species when k ranges from 3 to 7. In our tests, the constrained-optimisation did not changed the overall accuracy. The 2×2 contingency tables (Table 4.6) from the 5-close data set and 5-dispersed data set, revealed the algorithms recovered identical edge sets on each alignment.

	5-close		**5-dispersed**	
k	IRM	fast-IRM	IRM	fast-IRM
3	7	7	10	10
4	1	1	10	10
5	0	0	10	10
6	0	1	10	10
7	1	1	10	10

Table 4.5: Accuracy comparison between algorithm IRM and algorithm fast-IRM. Here, k is the rate matrix number. The numbers of the column "IRM" and the column "fast-IRM" are correct number of inferred models returned respectively by IRM and fast-IRM based on 10 simulated alignments.

		fast-IRM		
		True	**False**	**Total**
IRM	**True**	9	0	9
	False	1	41	41
	Total	10	40	50

(a) 5-close

		fast-IRM		
		True	**False**	**Total**
IRM	**True**	50	0	50
	False	0	0	0
	Total	50	0	50

(b) 5-dispersed

Table 4.6: Inferences from IRM and fast-IRM were perfectly congruent. From the 50 simulated alignments where $k \in \{3, 4, \ldots 7\}$, we record in the 2×2 table how often the algorithms were in agreement.

We assessed whether the fast-IRM algorithm resulted in log-likelihoods that were different to those from IRM. The differences in log-likelihood for each value of k

are presented in Figure 4.3. The absolute difference in the log-likelihood is tiny when $k \in \{3, 4, \ldots 7\}$ for 5-close species. While there were some outlier points, the absolute value of those difference was big for $k = 6$. For 5-dispersed species, however, the differences were larger with an Interquartile Range (IQR) ranges of around 4. As k increased, the difference increased also. Despite these discrepancies, the choice of optimal model was perfectly congruent between the two algorithms.

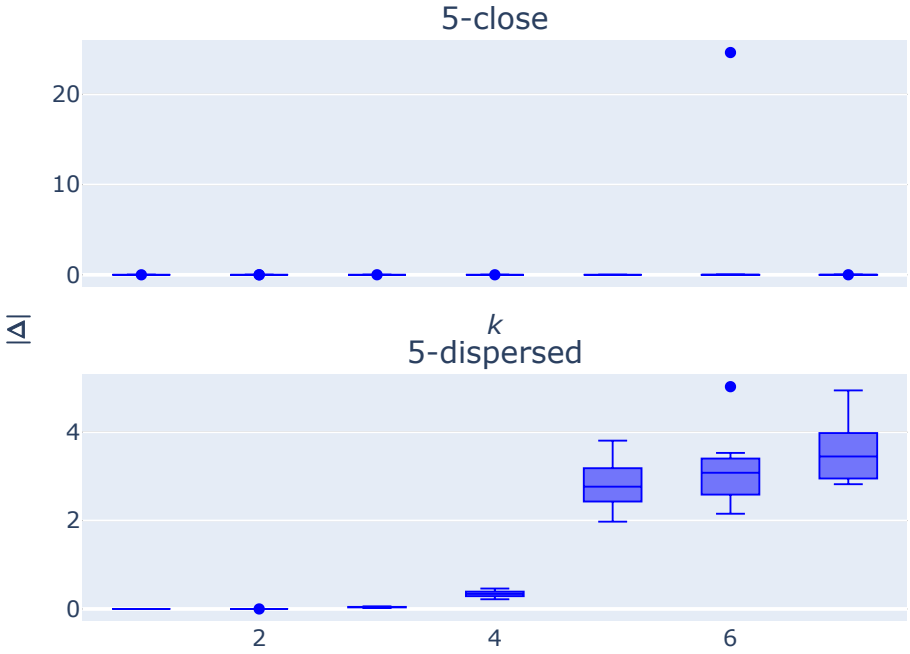


Figure 4.3: Consistency in log-likelihood between the IRM and fast-IRM algorithms. The y-axis is $|\Lambda| = |\log L(\text{IRM}) - \log L(\text{fast-IRM})|$. The x-axis is k, which is the number of distinct rate matrices.

4.3.3 Computational performance

We compared computational time required to identify the optimal model between ns-HAL, IRM and fast-IRM. The comparison was fair in so much as performance was measured from evaluation of the same simulated data sets as those used for

assessing statistical performance. The three algorithms were run on a shared high-performance computing environment. Each simulated alignment of each k was assigned to one compute node of this environment. Each node had 2× 8 core Intel Xeon E5-2670 (Sandy Bridge) 2.6GHz. Then the median and quartiles of running times from 10 simulated alignments for each k are displayed (Figure 4.4).

Our results established that both variants of IRM were markedly faster than ns-HAL. The compute time of ns-HAL ranged from $\sim 8\times$ ($k = 1$) to $\sim 30\times$ ($k = 2$) slower than IRM for 5-close species. For the 5-dispersed data set, the relative speed ranged from $\sim 5\times$ ($k = 1$) to $\sim 75\times$ ($k = 2$) slower than IRM. The speed relative to fast-IRM was considerably worse, with ns-HAL speeds slower by as much as $\sim 200\times$ ($k = 2$). The slowdown from using IRM compared to fast-IRM ranged from $\sim 1.6\times$ ($k = 1$, 5-close) to $\sim 3.4\times$ ($k = 7$, 5-dispersed). fast-IRM was systematically faster than HAL for both 5-close and 5-dispersed data sets (with the exception of $k = 1$). The speed of IRM was faster than HAL except for $k = 1$ and $k = 7$ for 5-close and except $k = 1$ for 5-dispersed.

4.3.4 Consistency between IRM and fast-IRM

On 50-taxon coral data with the second codon sites, on NCI, we compared the optimal models obtained by IRM and fast-IRM and their running time shown in the Table 4.7. We found that their optimal models are exactly same and their rate matrix number is 4.

	IRM	**fast-IRM**
k	4	4
t(sec)	175.0903	80.7143

Table 4.7: Optimal model number and running time on naked coral data. k is the number of rate matrices and t is the running time

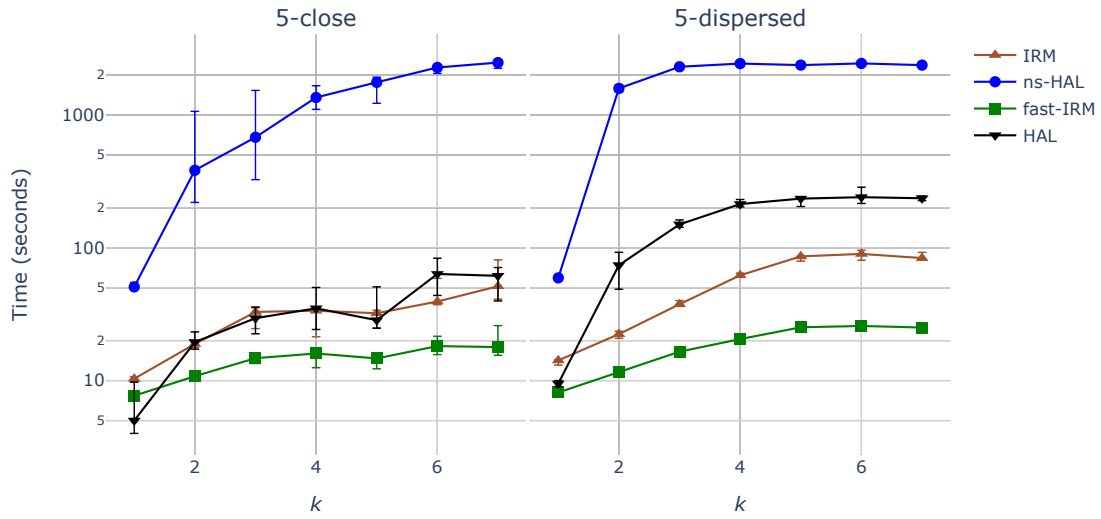


Figure 4.4: Median running time (log_{10}-scale) of HAL, ns-HAL, IRM and fast-IRM. The data set is indicated by the subplot title. Rate matrix number, k, is shown on the x-axis. The y-axis is the median of $\log_{10}(seconds)$ across the 10 replicates and the error bars being the lower and upper quartiles.

4.4 Discussion

We developed two algorithms, IRM and ns-HAL, for selection of the optimal non-stationary model for a given tree topology. In essence, both algorithms are bottom-up approaches; they start with the simplest model (one rate matrix) and gradually increase model complexity (number of rate matrices) of the model until a stop condition is satisfied. For a tree with large taxa number, bottom-up approaches should be faster than top-down approaches because the latter start with the most complex (parameter rich) model. When the number of taxa is large, the total solution space dramatically increases. Our results indicated that IRM or fast-IRM should be preferred over ns-HAL or HAL because of greater accuracy and markedly less compute-time.

Our comparison of algorithm accuracy established IRM was better than ns-HAL (Ta-

ble 4.2). For 5-close species, IRM had high accuracy for $k \in \{1, 2, 3\}$ but ns-HAL had high accuracy only for $k = 1$. In contrast, both algorithms showed systematically greater accuracy for 5-dispersed species compared to their 5-close results. IRM produced perfect accuracy for all values of k while ns-HAL showed a sharp reduction in accuracy for larger k on 5-dispersed data and it generated same or higher accuracy for each k than ns-HAL on 5-close (Table 4.2). The differences between the two data sets likely derives from differences in statistical power. The total number of substitutions in the two data sets differed by almost an order of magnitude (Table C.2). Support for a rate matrix depends on genetic distance in the edge set. Better accuracy was associated with greater genetic distance. The lower accuracy of all algorithms on the 5-close data set for larger k is a plausible consequence of this effect. At the same time, the greater accuracy of IRM shows that it is better suited for optimal models selection than either HAL or ns-HAL. This lower accuracy of the HAL algorithms is expected given they are considering a larger possible solution space and thus has a greater chance for false positives.

The IRM was motivated by empirical evidence that tendencies for mutation are heritable. Such evidence manifests within a species where, for example, defects in mismatch DNA repair genes are associated with familial cases of colon cancer [Peltomäki, 2001]. This heritability of mutation tendencies also manifest between species. For instance, numerous species are known to have lost completely, or substantially, the capacity to methylate their DNA and these properties are shared among closely related lineages [Capuano et al., 2014]. IRM explicitly captures this tendency. The mincut algorithm improves the efficiency of tree traversal by evaluating longer edges first. If, as seems plausible, the likelihood of evolving distinctive mutation processes is proportional to chronological time then this approach should prove efficient. While it remains a possibility that lineages may independently evolve

the exact same mutation tendencies, a prospect captured by HAL and not IRM, such convergent evolution seems less likely. Therefore, IRM is very important and very valuable as a null for examining whether lineages exhibiting comparable gene loss exhibit comparable substitution matrices. For example gene loss affecting 5mC and skeleton loss.

Given its consistency with biological drivers underpinning mutation, we suggest fast-IRM is well suited as a model selection strategy for non-stationary process models. On 50-taxon naked coral alignments with the second codon sites, we found that IRM and fast-IRM are consistent. In terms of compute speed, both fast-IRM and IRM were superior to ns-HAL for all values of k. The $\sim 2\times$ speed-up of fast-IRM over IRM along with the identical inferences (Table 4.6) shows fast-IRM is to be preferred. We expect this performance difference will increase with increasing number of taxa.

Supporting Information

Algorithms and tables are described in more detail in the Appendix C.

Chapter 5
Conclusion

In this thesis, we have addressed the methods of estimating nucleotide mutation and substitution rates. The results chapters of this thesis are composed of three projects about mutation rates and substitution rates. Chapter 2 obtained approximate solutions for the diffusion limit of the stationary distribution of the K-allele neutral Wright-Fisher model in the realistic limit of small scaled mutation rates for arbitrary values of K. The performance of this approximate solution was demonstrated by simulation study for 3-alleles and 4-alleles. In this chapter, a stationary solution with an arbitrary mutation rate matrix has been obtained to first order in θ. Chapter 3 is an application of the approximate distribution developed in Chapter 2. In chapter 3, we applied the approximate solution developed in Chapter 2 and maximum likelihood to estimate mutation rate matrix of site frequency data from multiple alignments. A number of synthetic datasets were constructed to test the efficiency of our methods. 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. At the same time, one gained physical insight into the role of the reversible and non-reversible parts of the rate matrix. In this chapter, the corresponding sampling distribution has been determined.

Chapter 4 studied heterogeneity in nucleotide substitution rate matrices across lineages. We developed algorithms IRM, fast-IRM and ns-HAL to select the optimal non-stationary model for a given tree topology. The three algorithms were all developed based on non-stationary GN model. IRM or fast-IRM are preferred over ns-HAL or HAL because of greater accuracy and markedly less compute-time. These two algorithms greatly improved the compute time for model selection of a non-stationary process, increasing the suite of problems to which this important substitution model class can be applied. In this study, we did not consider heterogeneity in the substitution process across sites which assume that different sites on the sequences have different rate matrices. However, compute time problem should be the first consideration for the implementation of site-heterogeneity.

In conclusion, the approximate solutions found for the stationary distribution and corresponding sampling distribution of a neutral multi-allele Wright-Fisher model are accurate for small scaled mutation rates. We have demonstrated that it is possible, in principle, to estimate a rate matrix from the site frequency spectrum of an alignment of genomes sampled from a population, provided certain conditions are met. IRM, fast-IRM and ns-HAL algorithms are very important and very valuable optimal model selection methods which can be effectively applied to the problem of estimating substitute rates which are heterogeneous across lineages.

Appendix A

Stationary solution to the $K = 3$ forward Kolmogorov equation near the simplex boundary

In this Appendix we consider the asymptotic behaviour of the stationary solution to the full forward Kolmogorov equation in the vicinity of an edge of the simplex in Fig. 2.2, but not immediately close to the corners. For $K = 3$ the stationary solution satisfies, from Eq. (2.3),

$$\begin{aligned} 0 = & \frac{1}{2}\frac{\partial^2}{\partial {x_1}^2}\left[x_1(1-x_1)f\right] - \frac{\partial^2}{\partial x_1 \partial x_2}(x_1 x_2 f) + \frac{1}{2}\frac{\partial^2}{\partial {x_2}^2}\left[x_2(1-x_2)f\right] \\ & - \frac{\partial}{\partial x_1}\left(\sum_{b=1}^{3} x_b Q_{b1} f\right) - \frac{\partial}{\partial x_2}\left(\sum_{b=1}^{3} x_b Q_{b2} f\right). \end{aligned} \tag{A.1}$$

Since we are interested in slow mutation rates, we introduce an overall mutation rate $\theta << 1$ defined by

$$Q_{ab} = \theta q_{ab}, \quad \sum_{a \neq b} q_{ab} = 1. \tag{A.2}$$

Without loss of generality, consider the A_1-A_2 edge of the simplex. Define new coordinates (x, y) via

$$(x_1, x_2, x_3) = (x - \tfrac{1}{2}y, 1 - x - \tfrac{1}{2}y, y). \tag{A.3}$$

We are interested in determining the form of the stationary solution in a region

$$|\log x + \log(1 - x)| << \theta^{-1}, \qquad 0 < y \leq \Lambda, \tag{A.4}$$

which is close to the A_1-A_2 edge, but sufficiently far from the corners that we can expect to be able to recover the line density Eq.(2.10). The cutoff Λ is not necessarily small. In the new coordinates the gradient operators are

$$\frac{\partial}{\partial x_1} = \frac{1}{2}\frac{\partial}{\partial x} - \frac{\partial}{\partial y}, \qquad \frac{\partial}{\partial x_2} = -\frac{1}{2}\frac{\partial}{\partial x} - \frac{\partial}{\partial y}. \tag{A.5}$$

A straightforward but lengthy calculation gives Eq. (A.1) in terms of the new coordinates:

$$
\begin{aligned}
0 &= \tfrac{1}{2}\frac{\partial^2}{\partial x^2}\left[x(1-x)f\right] \\
&\quad + \tfrac{1}{2}\frac{\partial^2}{\partial y^2}\left[y(1-y)f\right] \\
&\quad - \tfrac{1}{8}\frac{\partial^2}{\partial x^2}\left[yf\right] \\
&\quad + \tfrac{1}{2}\frac{\partial}{\partial x}\frac{\partial}{\partial y}\left[(1-2x)yf\right] \\
&\quad + \theta\frac{\partial}{\partial x}\left\{\left[(q_{12}+\tfrac{1}{2}q_{13})x-(q_{21}+\tfrac{1}{2}q_{23})(1-x)\right]f\right\} \\
&\quad + \tfrac{1}{2}\theta\frac{\partial}{\partial x}\left\{(-q_{12}+q_{21}-\tfrac{1}{2}q_{13}+\tfrac{1}{2}q_{23}-q_{31}+q_{32})yf\right\} \\
&\quad - \theta\frac{\partial}{\partial y}\left\{\left[q_{13}x+q_{23}(1-x)\right]f\right\} \\
&\quad + \theta\frac{\partial}{\partial y}\left\{(\tfrac{1}{2}q_{13}+\tfrac{1}{2}q_{23}+q_{31}+q_{32})yf\right\} \\
&= \text{Term 1} + \text{Term 2} + \ldots + \text{Term 8}. \qquad \text{(A.6)}
\end{aligned}
$$

Guided by the form of the Dirichlet solution relevant to the special 'parent-independent' case [Wright, 1969, Tier and Keller, 1978, Griffiths, 1979], consider the Ansatz

$$f(x,y) = \theta^2 s(x) y^{\theta s(x)-1} \sum_{k=0}^{\infty} g_k(x) y^k. \qquad \text{(A.7)}$$

Here $s(x)$ and each $g_k(x)$ are finite, analytic functions in the region defined by Eq. (A.4). The reason for the normalising factor θ^2 will become apparent later.

For fixed x and θ the asymptotic behaviour of each term in Eq. (A.6) as $y \to 0$ is as listed in the first column of Table A.1. The dominant terms are Term 2 and Term 7. Keeping only the dominant $O(y^{\theta s(x)-2})$ parts of these two terms, Eq. (A.6) entails that

Table A.1: Asymptotic behaviour of each term in Eq. (A.6).

	$\lim_{y\to 0}(\text{Term } i)$	$\lim_{\theta\to 0}\int_0^\Lambda(\text{Term } i)dy$
Term 1:	$O(y^{\theta s(x)-1}(\log y)^2)$	$\frac{1}{2}\theta\frac{d^2}{dx^2}[x(1-x)g_0(x)] + O(\theta^2)$
Term 2:	$O(y^{\theta s(x)-2})$	divergent
Term 3:	$O(y^{\theta s(x)}(\log y)^2)$	$O(\theta^2)$
Term 4:	$O(y^{\theta s(x)-1}\log y)$	$O(\theta^2)$
Term 5:	$O(y^{\theta s(x)-1}\log y)$	$O(\theta^2)$
Term 6:	$O(y^{\theta s(x)}\log y)$	$O(\theta^3)$
Term 7:	$O(y^{\theta s(x)-2})$	divergent
Term 8:	$O(y^{\theta s(x)-1})$	$O(\theta^3)$
Term 2 + Term 7:		$O(\theta^2)$

$$\begin{aligned} 0 &= \tfrac{1}{2}s(x)\frac{\partial^2}{\partial y^2}\left[y^{\theta s(x)}\right] - \theta\frac{\partial}{\partial y}\left[(q_{13}x + q_{23}(1-x))y^{\theta s(x)-1}\right] \\ &= \frac{\partial}{\partial y}\left\{\tfrac{1}{2}\theta\left[s(x) - 2(q_{13}x + q_{23}(1-x))\right]y^{\theta s(x)-1}\right\}. \end{aligned} \tag{A.8}$$

The term inside the curly brackets is the flux of probability across the edge $y = 0$ at a point x. This flux must be zero, so

$$s(x) = 2(q_{13}x + q_{23}(1-x)). \tag{A.9}$$

Now define $f_{12}(x)\,dx$ to be the probability contained in the region $[x, x+dx]\times[0,\Lambda]$. Then

$$\begin{aligned} f_{12}(x) &= \int_0^\Lambda f(x,y)\,dy \\ &= \theta^2 s(x) \sum_{k=0}^{\infty} g_k(x) \int_0^\Lambda y^{\theta s(x)-1+k} dy \\ &= \theta g_0(x) \Lambda^{\theta s(x)} + \theta^2 \sum_{k=1}^{\infty} \frac{\Lambda^{\theta s(x)+k}}{\theta s(x)+k} \\ &= \theta g_0(x) + O(\theta^2), \qquad \text{as } \theta \to 0. \end{aligned} \tag{A.10}$$

Similarly we have that

$$\int_0^\Lambda y f(x,y)\,dy = O(\theta^2), \quad \int_0^\Lambda \partial_y [y f(x,y)]\,dy = O(\theta^2), \qquad \text{as } \theta \to 0. \tag{A.11}$$

Thus the asymptotic behaviour of the integral of each term in Eq. (A.6) is as listed in the second column of Table A.1. Note that, while the integrals of Term 2 and Term 7 are separately divergent, Eq. (A.9) ensures that there is no g_0 contribution to the sum of these terms, which, by a similar calculation to that leading to Eq. (A.11), integrates to a contribution of order θ^2. Integrating Eq. (A.6) term-by-term, dividing by θ, and taking the limit $\theta \to 0$ then gives

$$\frac{d^2}{dx^2}\left[x(1-x)g_0(x)\right] = 0, \tag{A.12}$$

whose general solution is

$$g_0(x) = \frac{c}{x(1-x)} - \phi\left(\frac{1}{x} - \frac{1}{1-x}\right), \tag{A.13}$$

where c and ϕ are arbitrary constants. Finally, Eq. (A.10) implies that $f_{12}(x)$ takes the form of Eq. (2.10) with constants

$$C = \theta c, \qquad \Phi = \theta\phi. \tag{A.14}$$

In Section 2.4, C is identified by matching the sum of solutions along neighbouring edges with the normalisation of the effective 2-allele problem corresponding to partitioning of alleles, as in Appendix B, and Φ is identified as the flux across a line running from $(x, 0)$ to (x, Λ). Note from Eq. (A.10) that $f_{12}(x)$ is independent of Λ provided $\theta\,|\log\Lambda| << 1$, or equivalently, $\Lambda >> e^{-(1/\theta)}$. Note also from Eq. (2.22) that C and Φ are proportional to the rate matrix elements, and therefore of order θ. This justifies the normalisation factor θ^2 in our Ansatz Eq. (A.7).

Figure A.1 is a comparison between the asymptotic analytic solution $f(x, y)$ and a numerical simulation of the stationary distribution of the discrete Wright-Fisher model for a population size $N = 100$. The population size is chosen to be large enough to demonstrate the comparison over the first few lattice spacings away from the A_1-A_2 edge of the simplicial lattice.

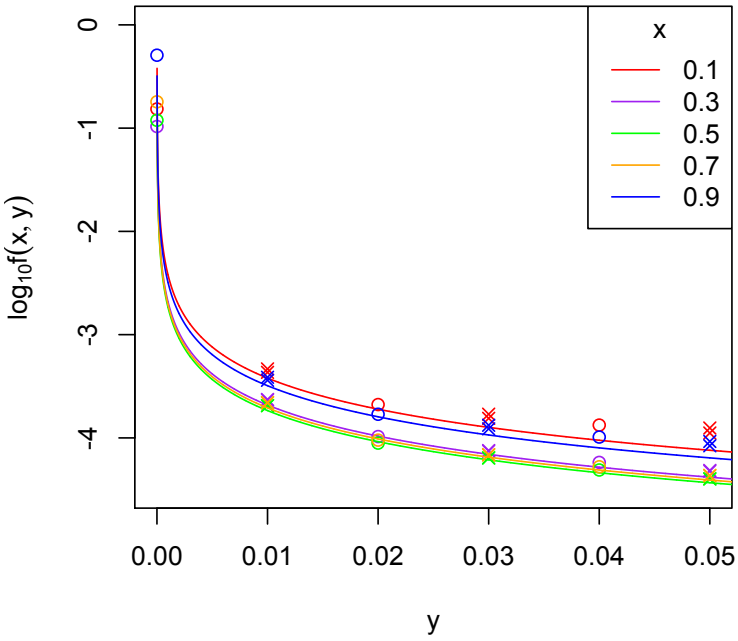


Figure A.1: Comparison between the Ansatz $f(x, y)$ and numerical simulation. The solid lines are the function $f(x, y)$ defined by Eqs. (A.7), (A.9), (A.13), (A.14) and (2.23) corresponding to the same rate matrix as Fig. 2.3. The circles and crosses are a numerical determination of the stationary distribution of the corresponding discrete Wright-Fisher model, Eq. (2.1), for a population size $N = 100$. Probabilities of the discrete distribution have been multiplied by N^2 for comparison with the continuum probability density. Circles are simplicial lattice points corresponding to continuum coordinates (x, y), crosses correspond to continuum coordinates $(x \pm (2N)^{-1}, y)$.

Appendix B

Equivalent 2-allele model

Suppose we divide the set of allele indices $\mathcal{I} = \{1, \ldots, K\}$ into two distinct complementary subsets $\mathcal{S}_1$ and $\mathcal{S}_2$, such that

$$\mathcal{S}_1 \cup \mathcal{S}_2 = \mathcal{I}, \qquad \mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset. \tag{B.1}$$

Given an instantaneous $K \times K$ rate matrix Q, we wish to construct an effective 2×2 rate matrix $\tilde{Q}$ which is equivalent in the sense that it defines a 2-state Markov chain whose stationary state is the corresponding marginal distribution of the stationary state of Q. More to the point, the transition frequencies between the two states of the effective Markov chain should match the transition frequencies between the two subsets of $\mathcal{I}$ in the original Markov chain.

Suppose Q has elements Q_{ab} for $a, b \in \mathcal{I}$; $\tilde{Q}$ has elements $\tilde{Q}_{AB}$, for $A, B \in \{1, 2\}$; and $L(t)$ is a random variable taking values in $\mathcal{I}$ representing the index of the allele occupied by the original Markov chain at time t. Then the probability of a transition

from state A at time t to state B at time $t + \delta t$ is

$$\begin{aligned}
\delta_{AB} + \tilde{Q}_{AB}\delta t &= \text{Prob}\,(L(t+\delta t) \in S_B \,|L(t) \in S_A) \\
&= \frac{\sum_{b\in S_B, a\in S_A} \text{Prob}\,\{L(t+\delta t) = b, L(t) = a\}}{\sum_{a\in S_A} \text{Prob}\,\{L(t) = a\}} \\
&= \frac{\sum_{b\in S_B, a\in S_A} (\delta_{ab} + Q_{ab}\delta t)\text{Prob}\,\{L(t) = a\}}{\sum_{a\in S_A} \text{Prob}\,\{L(t) = a\}}. \\
&= \delta_{AB} + \frac{\sum_{b\in S_B, a\in S_A} Q_{ab}\text{Prob}\,\{L(t) = a\}}{\sum_{a\in S_A} \text{Prob}\,\{L(t) = a\}}\delta t. \qquad \text{(B.2)}
\end{aligned}$$

It is clear that there can be no dynamically equivalent 2-state Markov chain in general as the right hand side depends on t. However, as we are only interested in equivalence of the stationary behaviour, we can set $t = \infty$ and replace $\text{Prob}\,\{L(t) = a\}$ with the stationary distribution π_a. We can then immediately read off the elements of the equivalent 2-state Markov chain:

$$\tilde{Q}_{AB} = \frac{\sum_{b\in S_B, a\in S_A} \pi_a Q_{ab}}{\sum_{a\in S_A} \pi_a}. \qquad \text{(B.3)}$$

The corresponding stationary distribution is

$$\tilde{\pi}_A = \sum_{a\in S_A} \pi_a. \qquad \text{(B.4)}$$

In this paper we have assumed that the marginal distributions of the stationary solution to the forward Kolmogorov equation for a K-allele neutral Wright-Fisher

model corresponding to partitionings of alleles are equal to solutions to the equivalent 2-allele model with a rate matrix $\tilde{Q}$ whose elements are given by Eq. (B.3). Note, however, the subtle point that we have not actually derived a rigorous proof of this claim, which, nevertheless is used implicitly in a number of papers including the current paper.

However, there are certain situations in which the right hand side of Eq. (B.2) is time-independent, and therefore the partitioned 2-state system is genuinely Markovian and the required result follows. For instance, if $Q_{ab} = Q_{a'b}$ whenever alleles a and a' belong to the same partitioning subset, then $\tilde{Q}_{AB} = \sum_{b \in \mathcal{S}_b} Q_{ab}$, where a is any element of $\mathcal{S}_A$. An example of this is the partitioning $\{A, T\}$, $\{C, G\}$ within the context of a strand-symmetric model [Vogl and Bergman, 2015]. Another example is the parent-independent model with $Q_{ab} = Q_b$, whose full stationary distribution is known to be a Dirichlet distribution [Wright, 1969, Tier and Keller, 1978, Griffiths, 1979]. One can confirm using the aggregation property of the Dirichlet distribution [Frigyik et al., 2010] that in this case the marginal distribution of the full K-allele model is indeed a beta distribution with the correct parameters.

Appendix C

Supporting information for Chapter 4

We present implementations of the core algorithms central to IRM below. These are presented as Python3 code. The code is not exactly identical to that in the distributed source code due to simplifications made for readability. Table C.1 identifies the counterpart in `phylotree.irm`.

Below	`phylotree.irm`
`variant_min_cut()`	`variant_min_cut()`
`get_edges()`	part of `get_edgeset()`
`cut_cluster_graph()`	part of `generate_cluster_graph()` and `cut_off_child()`
`cut_model_graphv`	`generate_cluster_graph()`
`is_cuttable()`	`is_cuttable()`

Table C.1: Mapping between functions below and the implementation in `phylotree.irm`

Algorithm S1 variant minimum cut algorithm

```
# mincut IRM model space generator
def variant_min_cut(tree):
    # Sort in descending order of edge/branch ENS distance
    ordered_edges = get_edges(tree, sort=True)
    num_edges = len(ordered_edges)
    # Graph stores all the parent models
    # Only a graph from the whole tree initially
    graphs = [[tree.deepcopy()]]
    start = 0
    while start < num_edges:
        # next_generation_graphs stores new cluster graphs
        # (all the children models) in the current
        # generation
        next_generation_graphs = []
        stop = start + 1
        start_edge = ordered_edges[start]
        # Finding nodes with the same edge/branch length
        # as one of start edge
        for i in range(stop, num_edges):
            if start_edge.length != ordered_edges[i].length:
                break
            stop = i + 1
        # Cut each graph to generate a new graph with
        # more subgraphs
        for cluster_graph in graphs:
            # The corresponding 2-dimension edge set of
            # parent model
            parent_model_edge_set = \
                get_model_edgeset(cluster_graph)
            # The corresponding 3-dimension edge set of
            # children models generated by parent model
            children_models_edge_set = []
            for cut_id in range(start, stop):
                # Find the subgraph with cutting edge and cut
                # it into two smaller subgraphs to generate a
                # new model
                new_model_graph = cut_model_graph( \
                    cluster_graph, ordered_edges[cut_id].name)
```

```
                    if new_model_graph:
                        new_edges = \
                        get_model_edgeset(new_model_graph)
                        children_models_edge_set.append(new_edges)
                        next_generation_graphs. \
                        append(new_model_graph)
                if children_models_edge_set:
                    yield parent_model_edge_set, \
                    children_models_edge_set
                else:   # Not cut, pass to next generation
                    next_generation_graphs.append(cluster_graph)

            graphs = next_generation_graphs
            start = start + 1
```

Algorithm S2 Edge and graph selection utility functions

```
# Construct edge array from a tree
def get_edges(tree, sort=True):
    edges = [(node.length, node) for node in tree.preorder()
             if not node.is_root()]
    if sort:  # Make descending by edge/branch length
        edges.sort(reverse=True)
    return [e for _, e in edges]

# Construct model edgeset (a 2-dimension array)
def get_model_edgeset(cluster_graph, sort=True):
    edge_names_set = []
    for subgraph in cluster_graph:  # Traverse subgraph
        edges = get_edges(subgraph)
        edge_names = [edge.name for edge in edges]
        if not subgraph.is_root():
            edge_names.append(subgraph.name)
        if sort:
            edge_names.sort() # alphabetical sort
        edge_names_set.append(edge_names)
    if sort:
        edge_names_set.sort() # alphabetical sort
    return edge_names_set
```

Algorithm S3 Graph and model cutting utility functions

```
def is_cuttable(node):
    if node.is_root():
        return False
    parent = node.parent
    return not parent.is_root() or len(parent.children) > 1

# graph is modified, so user must make sure they copy it
def cut_cluster_graph(graph, node_name):
    node = graph.get_node_matching_name(node_name)
    detached_child = node.parent.pop(node.index_in_parent())
    node.parent = None
    return node, graph

# cut a model graph
def cut_model_graph(model_graph, node_name):
    new_model_graph = []
    for subgraph in model_graph:
        new_subgraph = None
        # Check whether the subgraph includes the node
        # If yes, split the matched source subgraph into 2
        for node in subgraph.preorder():
            if node.name == node_name:
                if not is_cuttable(node):
                    return None # Unnecessary to cut

                # Construct a subgraph from the parent node
                copied = subgraph.copy_topology()
                new_subgraph, cut_subgraph = \
                    cut_cluster_graph(copied, node_name)
                # Add new subgraph at the beginning
                new_model_graph.insert(0, new_subgraph)
                # Add source subgraph which is cut
                new_model_graph.append(cut_subgraph)
                break
        if new_subgraph is None:
            new_model_graph.append(subgraph)
    return new_model_graph
```

	5-close		5-dispersed	
k	**ns-HAL**	**IRM**	**ns-HAL**	**IRM**
1	2081	2081	17923	17923
2	2099	2099	17840	17840
3	2091	2091	17780	17780
4	2091	2091	17700	17700
5	2081	2081	17784	17783
6	2079	2079	17751	17751
7	2097	2098	17750	17749

Table C.2: Substitution number comparison between algorithm ns-HAL and algorithm IRM on 5-close and 5-dispersed species. Here, the values in table are round number of average ENS of 10 simulated alignments

References

R. Albalat, J. Martí-Solans, and C. Cañestro. DNA methylation in amphioxus: from ancestral functions to new roles in vertebrates. *Briefings in functional genomics*, 11(2):142–155, 2012.

E. Baake and R. Bialowons. Ancestral processes with selection: Branching and Moran models. In J. Miekisz, editor, *Stochastic Models in Biological Sciences.*, volume 80 of *Banach Center Publications*, pages 33–52. Institute of Mathematics, Polish Academy of Sciences, 2008.

P.-F. Baisnée, S. Hampson, and P. Baldi. Why are complementary DNA strands symmetric? *Bioinformatics*, 18(8):1021–1033, 2002.

D. Barry and J. Hartigan. Asynchronous distance between homologous DNA sequences. *Biometrics*, 43(2):261, 1987.

N. Batchelder. Coverage.py: a tool for measuring test coverage. https://coverage.readthedocs.io, 2019.

C.J. Burden and H. Simon. Genetic drift in populations governed by a galton–watson branching process. *Theoretical population biology*, 109:63–74, 2016.

C.J. Burden and Y. Tang. An approximate stationary solution for multi-allele neutral diffusion with low mutation rates. *Theoretical population biology*, 112:22–32, 2016.

C.J. Burden and Y. Tang. Rate matrix estimation from site frequency data. *Theoretical population biology*, 113:23–33, 2017.

K.P. Burnham and D.R. Anderson. *Model selection and multimodel inference-a practical information theoretic approach*. Springer-Verlag, second edition, 2002.

M. Cao, j. Shi, J. Wang, J. Hong, B. Cui, and G. Ning. Analysis of human triallelic SNPs by next-generation sequencing. *Annals of human genetics*, 79(4):275–281, 2015.

F. Capuano, M. Mülleder, R. Kok, H.J. Blom, and M. Ralser. Cytosine DNA methylation is found in Drosophila melanogaster but absent in Saccharomyces cerevisiae, Schizosaccharomyces pombe, and other yeast species. *Analytical chemistry*, 86(8): 3697–3702, 2014.

J.T. Chang. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Mathematical biosciences*, 137(1):51–73, 1996.

A. Couce, L.V. Caudwell, C. Feinauer, T. Hindré, J.-P. Feugeas, M. Weigt, R.E. Lenski, D. Schneider, and O. Tenaillon. Mutator genomes decay, despite sustained fitness gains, in a long-term experiment with bacteria. *Proceedings of the national academy of sciences*, 114(43):E9026–E9035, 2017.

J.F. Crow. The high spontaneous mutation rate: is it a health risk? *Proceedings of the national academy of sciences*, 94(16):8380–8386, 1997.

G. Culligan, K.M.and Meyer-Gauen, J. Lyons-Weiler, and J.B. Hays. Evolutionary origin, diversification and specialization of eukaryotic MutS homolog mismatch repair proteins. *Nucleic acids research*, 28(2):463–471, 2000.

A.L. Edwards. Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. *Psychometrika*, 13(3):185–187, 1948.

A.M. Etheridge. *Some Mathematical Models from Population Genetics: École D'Été de Probabilités de Saint-Flour XXXIX-2009*, volume 2012 of *Lecture Notes in Mathematics*. Springer, Berlin Heidelberg, 2011.

A.M. Etheridge and R.C. Griffiths. A coalescent dual process in a Moran model with genic selection. *Theoretical population biology*, 75(4):320–330, 2009.

W.J. Ewens. A note on the sampling theory for infinite alleles and infinite sites models. *Theoretical population biology*, 6(2):143–148, 1974.

W.J. Ewens. *Mathematical population genetics*, volume v. 27. Springer, New York, 2nd ed edition, 2004.

J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981.

J. Felsenstein. *Inferring phylogenies*. Sinauer, 2003.

P.G. Foster. Modeling compositional heterogeneity. *Systematic biology*, 53(3):485–495, 2004.

P.G. Foster and D.A. Hickey. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *Journal of molecular evolution*, 48(3): 284–290, 1999.

B.A. Frigyik, A. Kapila, and M.R. Gupta. Introduction to the Dirichlet distribution and related processes. *Department of Electrical Engineering, University of Washignton*, 2010.

Y. Fu and H. Huai. Estimating mutation rate: how to count mutations? *Genetics*, 164(2):797–805, 2003.

M. Gil, M.S. Zanetti, S. Zoller, and M. Anisimova. CodonPhyML: fast maximum likelihood phylogeny estimation under codon substitution models. *Molecular biology and evolution*, 30(6):267–281, 2013.

G.H. Golub and C.F. Van Loan. *Matrix computations*. The Johns Hopkins University Press, Baltimore, fourth edition, 2013.

R.C. Griffiths. A transition density expansion for a multi-allele diffusion model. *Advances in applied probability*, 11(2):310–325, 1979.

M. Hasegawa, H. Kishino, and T.-a. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution*, 22(2): 160–174, 1985.

A. Hodgkinson and A. Eyre-Walker. Human triallelic sites: evidence for a new mutational mechanism? *Genetics*, 184(1):233–241, 2010.

B.R. Holland, P.D. Jarvis, and J.G. Sumner. Low-parameter phylogenetic inference under the general Markov model. *Systematic biology*, 62(1):78–92, 2013.

D. Huang, R. Meier, P.A. Todd, and L.M. Chou. Slow mitochondrial COI sequence evolution at the base of the metazoan tree and its implications for DNA barcoding. *Journal of molecular evolution*, 66(2):167–174, 2008.

G. Huttley. Do genomic datasets resolve the correct relationship among the placental, marsupial and monotreme lineages? *Australian journal of zoology*, 57(4): 167–174, 2009.

G. Huttley. cogent3: A toolkit for statistical analysis of biological sequences. https://github.com/cogent3/cogent3, 2019.

D.G. Hwang and P. Green. Bayesian markov chain monte carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the national academy of sciences*, 101(39):13994–14001, 2004.

V. Jayaswal, L.S. Jermiin, and J. Robinson. Estimation of phylogeny using a general markov model. *Evolutionary bioinformatics*, 1:62–80, 2005.

V. Jayaswal, L.S. Jermiin, L. Poladian, and J. Robinson. Two stationary nonhomogeneous markov models of nucleotide sequence evolution. *Systematic biology*, 60 (1):74–86, 2011.

V. Jayaswal, T. Wong, J. Robinson, L. Poladian, and L.S. Jermiin. Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. *Systematic biology*, 63(5):726–742, 2014.

T.H. Jukes and V. Bhushan. Silent nucleotide substitutions and G+C content of some mitochondrial and bacterial genes. *Journal of molecular evolution*, 24(1-2): 39–44, 1986.

B.D. Kaehler, V.B Yap, R. Zhang, and G.A. Huttley. Genetic distance for a general non-stationary markov substitution process. *Systematic biology*, 64(2):281–293, 2015.

B.D. Kaehler, V.B. Yap, and G.A. Huttley. Standard codon substitution models overestimate purifying selection for nonstationary data. *Genome biology and evolution*, 9(1):134–149, 2017.

S. Karlin and C. Burge. Dinucleotide relative abundance extremes: a genomic signature. *Trends in genetics*, 11(7):283 – 290, 1995.

S. Karlin and J. Mrázek. Compositional differences within and between eukaryotic genomes. *Proceedings of the national academy of sciences*, 94(19):10227–10232, 1997.

A. Keinan and A.G. Clark. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, 336(6082):740–743, 2012.

M.V. Kitahara, M. Lin, S. Forêt, G. Huttley, D.J. Miller, and C.A. Chen. The "naked coral" hypothesis revisited – evidence for and against scleractinian monophyly. *Plos one*, 9:1–13, 2014.

R.D. Knight, S.J. Reeland, and L.F. Landweber. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome biology*, 2(4):RESEARCH0010, 2001.

C. Lanave, G. Preparata, C. Sacone, and G. Serio. A new method for calculating evolutionary substitution rates. *Journal of molecular evolution*, 20(1):86–93, 1984.

S.E. Luria and M. Delbruck. Mutation of bacteria from virus sensitivity to virus resistance. *Genetics*, 28(6):491–511, 1943.

G. Macaya, J.-P. Thiery, and G. Bernardi. An approach to the organization of eukaryotic genomes at a macromolecular level. *Journal of molecular biology*, 108 (1):237–254, 1976.

K.T. Nishant, N.D. Singh, and E. Alani. Genomic mutation rates: what high-throughput methods can tell us. *Bioessays*, 31(9):912–920, 2009.

J. Parsch, S. Novozhilov, S.S. Saminadin-Peter, K.M. Wong, and P. Andolfatto. On the utility of short intron sequences as a reference for the detection of positive and negative selection in drosophila. *Molecular biology and evolution*, 27(6):1226–1234, 2010.

P. Peltomäki. Deficient DNA mismatch repair: a common etiologic factor for colon cancer. *Human molecular genetics*, 10(7):735–40, 2001.

C. Phillips, J. Amigo, A. Carracedo, and M. Lareu. Tetra-allelic SNPs: informative forensic markers compiled from public whole-genome sequence data. *Forensic science international: genetics*, 19:100–106, 2015.

D. Pisani, W. Pett, M. Dohrmann, R. Feuda, R. Omar, H. Philippe, N. Lartillot, and G. Wörheide. Genomic data do not support comb jellies as the sister group to all other animals. *Proceedings of the national academy of sciences*, 112(50): 15402–15407, 2015.

G.A. Pont-Kingdon, N.A. Okada, J.L. Macfarlane, C.T. Beagley, D.R. Wolstenholme, T. Cavalier-Smith, and G.D. Clark-Walker. A coral mitochondrial mutS gene. *Nature*, 375(6527):109–111, 1995.

J.E. Pool, I. Hellmann, J.D. Jensen, and R. Nielsen. Population genetic inference from genomic sequence variation. *Genome research*, 20(3):291–300, 2010.

Z. Qi. New algorithms for luria-delbrück fluctuation analysis. *Mathematical biosciences*, 196(2):198–214, 2005.

S. Reddy, R.T. Kimball, A. Pandey, P.A. Hosner, M.J. Braun, S.J. Hackett, K. Han, J. Harshman, C.J. Huddleston, S. Kingston, B.D. Marks, K.J. Miglia, W.S. Moore, F.H. Sheldon, C.C. Witt, T. Yuri, and E.L. Braun. Why do phylogenomic data sets yield conflicting trees? data type influences the avian tree of life more than taxon sampling. *Systematic biology*, 66(5):857–879, 2017.

Y. Roggo, L. Duponchel, and J.-P. Huvenne. Comparison of supervised pattern recognition methods with McNemar's statistical test application to qualitative analysis of sugar beet by near-infrared spectroscopy. *Analytica chimica acta*, 477 (2):187–200, 2003.

A. Rokas, N. King, J. Finnerty, and S.B. Carroll. Conflicting phylogenetic signals at the base of the metazoan tree. *Evolution & development*, 5(4):346–359, 2003a.

A. Rokas, B.L. Williams, N. King, and S.B. Carroll. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425(6960):798–804, 2003b.

A. RoyChoudhury and J. Wakeley. Sufficiency of the number of segregating sites in the limit under finite-sites mutation. *Theoretical population biology*, 78(2): 118–122, 2010.

A. Siepel and D. Haussler. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Molecular biology and evolution*, 21(3):468–488, 2004.

F.M. Stewart, D.M. Gordon, and B.R. Levin. Fluctuation analysis: the probability distribution of the number of mutants under different conditions. *Genetics*, 124 (1):175–185, 1990.

N. Sueoka. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *Journal of molecular evolution*, 40(3):318–325, 1995.

S. Tavaré. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences*, 17:57–86, 1986.

C. Tier and J.B. Keller. A tri-allelic diffusion model with selection. *SIAM journal on applied mathematics*, 35(3):521–535, 1978.

K.L. Verbyla, V.B. Yap, A. Pahwa, Y. Shao, and G.A. Huttley. The embedding problem for markov models of nucleotide substitution. *Plos one*, 8(7):e69187, 2013.

C. Vogl. Estimating the scaled mutation rate and mutation bias with site frequency data. *Theoretical population biology*, 98:19–27, 2014.

C. Vogl and J. Bergman. Inference of directional selection and mutation parameters assuming equilibrium. *Theoretical population biology*, 106:71–82, 2015.

G.A. Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical population biology*, 7(2):256–276, 1975.

S. Wielgoss, J.E. Barrick, O. Tenaillon, S. Cruveiller, B. Chane-Woon-Ming, C. Médigue, R.E. Lenski, and D. Schneider. Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with escherichia coli. *G3: genes, genomes, genetics*, 1(3):183–186, 2011.

Wolfram Research Inc. *Mathematica Version 10.3*. Wolfram Research, Inc., Champaign, Illinois, 2015.

T. Wong. Hal-Has. https://github.com/thomaskf/Hal-Has, 2017.

S. Wright. Evolution in mendelian populations. *Genetics*, 16(2):97–159, 1931.

S. Wright. Evolution and the genetics of populations: the theory of gene frequencies. vol 2: The theory of gene frequencies, 1969.

X. Xiong, J.M. Boyett, R.G. Webster, and J. Stech. A stochastic model for estimation of mutation rates in multiple-replication proliferation processes. *Journal of mathematical biology*, 59(2):175–191, 2009.

Z. Yang. Estimating the pattern of nucleotide substitution. *Journal of molecular evolution*, 39(1):105–111, 1994.

Z. Yang and D. Roberts. On the use of nucleic acid sequences to infer early branchings in the tree of life. *Molecular biology and evolution*, 12(3):451–458, 1995.

V.B. Yap, H. Lindsay, S. Easteal, and G. Huttley. Estimates of the effect of natural selection on protein-coding content. *Molecular biology and evolution*, 27(3):726–734, 2010.

S. Yi. Neutrality and molecular clocks. *Nature education knowledge*, 4(2):3, 2013.

A. Zemach, I.E. McDaniel, P. Silva, and D. Zilberman. Genome-wide evolutionary analysis of eukaryotic dna methylation. *Science*, 328(5980):916–919, 2010.

K. Zeng. A simple multiallele model and its application to identifying preferred–unpreferred codons using polymorphism data. *Molecular biology and evolution*, 27(6):1327–1337, 2010.

Z. Zhao and E. Boerwinkle. Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome research*, 12(11):1679–1686, 2002.

L. Zou, E. Susko, C. Field, and A.J. Roger. Fitting nonstationary general-time-reversible models to obtain edge-lengths and frequencies for the barry–hartigan model. *Systematic biology*, 61(6):927–940, 2012.