Theoretical Population Biology 98 (2014) 19–27

Contents lists available at ScienceDirect

Theoretical Population Biology

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

Estimating the scaled mutation rate and mutation bias with site frequency data

Claus Vogl

Institute of Animal Breeding and Genetics, Veterinärmedizinische Universität Wien, Veterinärplatz 1, A-1210 Vienna, Austria

ARTICLE INFO

Article history:
Received 7 May 2014
Available online 18 October 2014

Keywords:
Mutation–drift equilibrium
Beta–binomial
Stirling distribution
EM-algorithm
Markov chain Monte Carlo algorithm
Posterior

ABSTRACT

The distribution of allele frequencies of a large number of biallelic sites is known as "allele-frequency spectrum" or "site-frequency spectrum" (SFS). Without selection and in regions of relatively high recombination rates, sites may be assumed to be independently and identically distributed. With a beta equilibrium distribution of allelic proportions and binomial sampling, a beta–binomial compound likelihood for each site results. The likelihood of the data and the posterior distribution of two parameters, scaled mutation rate θ and mutation bias α, is investigated in the general case and for small scaled mutation rates θ. In the general case, an expectation–maximization (EM) algorithm is derived to obtain maximum likelihood estimates of both parameters. With an appropriate prior distribution, a Markov chain Monte Carlo sampler to integrate the posterior distribution is also derived. As far as I am aware, previous maximum likelihood or Bayesian estimators of θ, explicitly or implicitly assume small scaled mutation rates, i.e., $\theta \ll 1$. For $\theta \ll 1$, maximum-likelihood estimators are also derived for both parameters using a Taylor series expansion of the beta–binomial distribution. The estimator of θ is a variant of the Ewens–Watterson estimator and of the maximum likelihood estimator derived with the Poisson Random Field approach. With a conjugate prior distribution, marginal and conditional beta posterior distributions are also derived for both parameters.

1. Introduction

With relatively high recombination compared to mutation rates, each polymorphic nucleotide or site can be considered independently. The distribution of allele frequencies of a large number of such loci has been called "allele-frequency spectrum" or "site-frequency spectrum" (SFS). Without selection, identical mutation distributions and demographic history for all sites may be assumed, i.e., sites are independently and identically distributed and subject only to the population genetic forces of mutation and drift. With a biallelic model, the first to derive the beta equilibrium distribution of the allelic proportion in a population was apparently Wright (1931). Application to large scale data has usually focused on the parameter range of small scaled mutation rates, i.e., $\theta = \mu N \ll 1$, where μ is the total mutation rate (including unobservable mutations) per site and generation and N is the (effective) haploid population number or size. In this parameter range, maximum likelihood and Bayesian estimates of θ from SFS data of sample size L loci and M alleles have been derived using the Poisson Random Field (PRF) approach (Hartl et al., 1994; Bustamante et al., 2001, 2003; RoyChoudhury and Wakeley, 2010). In particular, the number of polymorphic sites in the sample was found to be a sufficient statistic. The other parameter of the beta, which can be parametrized as mutation bias, has been ignored so far.

In this article, an expectation–maximization (EM) algorithm is derived for the general case, i.e., arbitrary θ, to obtain the maximum likelihood estimates of both the scaled mutation rate θ and the mutation bias α. The EM algorithm requires the solution of a polynomial of order M during each step. A Metropolis–Hastings Markov chain Monte Carlo sampling algorithm that approximates the joint posterior is also presented. For small scaled mutation rates, i.e., $\theta \ll 1$ the maximum-likelihood estimates for both parameters are derived. Given appropriate conjugate prior distributions, the marginal distribution of θ then depends only on the frequency of polymorphic sites in the SFS and is a generalized beta; the conditional distribution of the mutation bias α depends on θ and the frequencies of sites with the two monomorphic types and is also a generalized beta. With these two distributions, the joint posterior distribution is also obtained.

2. Biallelic mutation drift

In this section, the biallelic mutation drift model and its equilibrium beta distribution is reviewed. Usually, the diploid

E-mail address: claus.vogl@vetmeduni.ac.at.

http://dx.doi.org/10.1016/j.tpb.2014.10.002
0040-5809/

Wright–Fisher model is taken as a starting point to derive the diffusion model. The haploid decoupled Moran model is, however, slightly more convenient.

Assume a population of N haploid individuals; each may assume the state of zero or one, corresponding to the two arbitrarily labeled alleles. With the decoupled Moran model (Baake and Bialowons, 2008; Etheridge and Griffiths, 2009; Vogl and Clemente, 2012), per step either (i) (*mutation*) at a rate of $\mu = \mu_0 + \mu_1$, a random individual i is picked to mutate to type one with probability μ_1/μ or to type zero with probability μ_0/μ; or (ii) (*genetic drift*) at a rate of one, a random individual i is replaced by another random individual j. Thus, the rate of change of the allelic proportion x per unit time of the mean is caused by mutation

$$\mathrm{M}_{\delta x} = \frac{1}{N^2}\theta(\alpha(1-x) - (1-\alpha)x)N = \frac{1}{N^2}\theta(\alpha - x)N, \tag{1}$$

and that of the variance by genetic drift

$$\mathrm{V}_{\delta x} = \frac{2}{N^2}x(1-x)N^2. \tag{2}$$

Scaling space with $1/N$ and time with $1/N^2$ and taking the appropriate limits, the Kolmogorov forward (or Fokker–Planck) diffusion equation

$$\frac{\partial}{\partial t}\phi(x,t) = \left(\frac{\partial^2}{\partial x^2}x(1-x) - \frac{\partial}{\partial x}\theta(\alpha - x)\right)\phi(x,t) \tag{3}$$

then describes the evolution of the probability of the allelic proportion x forward in time t. This is the same temporal direction as the transitions in the Wright–Fisher and Moran models.

The equilibrium density of this process is beta (Wright, 1931),

$$\Pr(x \mid \alpha, \theta) = \frac{\Gamma(\theta)}{\Gamma(\alpha\theta)\Gamma(\beta\theta)}\, x^{\alpha\theta-1}(1-x)^{\beta\theta-1}, \tag{4}$$

as can be shown by substituting into the forward diffusion equation (3). The corresponding equilibrium density for the diploid Wright–Fisher model is obtained from (4) by replacing $\theta = \mu N$ with $\theta^* = 4\mu N$.

3. General model

The following is assumed:

Assumption 1. Allele frequency data of biallelic loci (sites), i.e., SFS data, are available from $L < \infty$ loci, indexed by $1 \le l \le L$.

Assumption 2. The allelic proportions x_l at each of the L sites are independently and identically beta distributed.

Assumption 3. The likelihood of the allelic frequencies y_l in the sample are binomial with parameters x_l and identical sample size M.

Note that generalization to multiallelic loci is possible, as long as mutations are parent independent. Furthermore, dropping the assumption of constant M for all L only increases the complexity of notation.

3.1. The equilibrium distribution

The frequency of the, arbitrarily chosen, first allelic type in the sample is y; that of the other allelic type is $M-y$. Set L_y to the number of samples with y alleles of the first type; obviously: $\sum_y L_y = L$.

As shown in the previous section, the equilibrium distribution of the allelic proportion x of the first allele is beta (4). Given a small sample of size M the joint distribution of the number of alleles y of the first type, i.e., the likelihood given x, is assumed to be binomial. The joint distribution of y and x is:

$$\Pr(y, x \mid \theta, M) = \binom{M}{y}\frac{\Gamma(\theta)}{\Gamma(\alpha\theta)\Gamma(\beta\theta)}\, x^{y+\alpha\theta-1} \times (1-x)^{M-y+\beta\theta-1}. \tag{5}$$

Integrating out x gives the beta–binomial (compound) distribution:

$$\begin{aligned}\Pr(y \mid \alpha, \theta, M) &= \binom{M}{y}\frac{\Gamma(\theta)}{\Gamma(\alpha\theta)\Gamma(\beta\theta)}\int_0^1 x^{y+\alpha\theta-1}(1-x)^{M-y+\beta\theta-1}\,dx \\ &= \binom{M}{y}\frac{\Gamma(\theta)}{\Gamma(\alpha\theta)\Gamma(\beta\theta)}\frac{\Gamma(y+\alpha\theta)\Gamma(M-y+\beta\theta)}{\Gamma(M+\theta)}.\end{aligned} \tag{6}$$

The likelihood is a product of beta–binomials:

$$\Pr(L_0, \ldots, L_M \mid \alpha, \theta, M) = \frac{L!}{\prod_{i=0}^{M} L_y!}\prod_{y=0}^{M}\left(\binom{M}{y}\frac{\Gamma(\theta)}{\Gamma(\alpha\theta)\Gamma(\beta\theta)} \times \frac{\Gamma(y+\alpha\theta)\Gamma(M-y+\beta\theta)}{\Gamma(M+\theta)}\right)^{L_y}. \tag{7}$$

Interest is centered on obtaining (maximum-likelihood) estimates of θ and α given the vector of allelic counts $(L_0, \ldots, L_M)$ or, in a Bayesian context, their posterior distribution given a suitable prior. As a function of α, the distribution is a weighted sum of beta distributions; as a function of θ, the distribution is a rational function. A rational function can be integrated by partial fraction decomposition. Introduction of auxiliary variables that count the number of mutations in each allelic class conditional on θ, α, y and M simplifies the task of finding estimators and posterior distributions. The following theorem provides their distribution.

Theorem 1. *With parent independent mutation, the number j of mutations in an allelic class, conditional on the scaled mutation rate towards this class, i.e., $\alpha\theta$ (or $\beta\theta$ respectively), and the number of samples of this type y (or $M-y$ respectively) is distributed according to the weighted Stirling distribution of the first kind (Ewens, 1972) and conditionally independent of that in other allelic classes.*

Proof. Only the case for the first allelic type is provided; that of the other allelic type follows analogously.—It is well known that the probability of coalescence within a sample of size $i = y$ with $y \ge 1$ is proportional to $i-1$ while the probability of a mutation is proportional to $\alpha\theta$. Thus the number of mutations in this sample is one with probability $\alpha\theta/(\alpha\theta + i - 1)$, and zero otherwise. The moment generating function is obviously

$$mgf_j(t \mid \alpha\theta, i) = \frac{\alpha\theta e^{tj} + i - 1}{\alpha\theta + i - 1}. \tag{8}$$

With a coalescence event, the number of samples is reduced by one. With parent independent mutation, no information is gained by knowing the allelic type from which the mutation occurred. Thus the number in the sample is reduced by one to $i-1$ both with a mutation and a coalescence. The process then repeats until $i = 1$, where a mutation event occurs with certainty. Since the events are obviously independent for each level, the moment generating functions multiply:

$$mgf_j(t \mid \alpha\theta, y) = \prod_{i=0}^{y-1}\frac{\alpha\theta e^{tj} + i}{\alpha\theta + i} = \frac{\Gamma(\theta e^{tj} + y)\Gamma(\alpha\theta)}{\Gamma(\alpha\theta e^{tj})\Gamma(\alpha\theta + y)}. \tag{9}$$