ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Theoretical Population Biology 63 (2003) 91–103

Theoretical Population Biology

http://www.elsevier.com/locate/ytpbi

Maximum likelihood and Bayesian methods for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data

Carlos D. Bustamante,[a,*,1] Rasmus Nielsen,[b] and Daniel L. Hartl[c]

[a] *Mathematical Genetics Group, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, UK OX1 3TG*
[b] *Biological Statistics and Computational Biology, Cornell University, 434 Warren Hall, Ithaca, NY 14853-7801, USA*
[c] *Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA*

Received 16 February 2002; received in revised form 13 September 2002; accepted 23 October 2002

Abstract

Maximum likelihood and Bayesian approaches are presented for analyzing hierarchical statistical models of natural selection operating on DNA polymorphism within a panmictic population. For analyzing Bayesian models, we present Markov chain Monte-Carlo (MCMC) methods for sampling from the joint posterior distribution of parameters. For frequentist analysis, an Expectation–Maximization (EM) algorithm is presented for finding the maximum likelihood estimate of the genome wide mean and variance in selection intensity among classes of mutations. The framework presented here provides an ideal setting for modeling mutations dispersed through the genome and, in particular, for the analysis of how natural selection operates on different classes of single nucleotide polymorphisms (SNPs).

© 2003 Elsevier Science (USA). All rights reserved.

Keywords: SNPs; Bayesian statistics; Hierarchical models; Poisson random field; Population genetics; MCMC; EM algorithm; Gibbs sampling

1. Introduction

Population geneticists have long sought to quantify the importance of different evolutionary forces (mutation, selection, demographics, recombination, and epistasis) in patterning standing genetic variation within natural populations (Lewontin, 1974). This research program has relied on three components: population genetic theory, experimental observation of genetic variation, and statistical inference linking the observed variation to the theory. Population genetic theory has thus meant the mathematical description of how different evolutionary forces alone or in combination produce patterns of genetic variation within and between populations and species. By the "pattern of genetic variation" we mean either the stationary distributions of allele frequencies (in the broad sense) and their summary statistics (heterozygosity, number of alleles, and frequencies of mutations, etc.) under differing models of genetic evolution or the genealogical relationship among alleles.

Experimental population genetics has sought to quantify the amount of genetic variation within a chosen set of natural populations at a particular point in time and to interpret this variation in light of the theory. The chief problem in this endeavor is that the observed patterns of genetic variation can often be explained by several different combinations of evolutionary forces (i.e., in a statistical sense, there are too many degrees of freedom for the type of data). The standard solution to this problem has been to assume some strict "neutral" model (in the Kimuran sense) where the expected pattern of variation or some summary statistic of the pattern can be derived and to test whether the observed pattern of variation is consistent with the strict neutral model. For this reason,

Abbreviations: EB estimate, empirical Bayes estimate; EM algorithm, expectation–maximization algorithm; MAP, maximum a posteriori estimate; MLE, maximum likelihood estimate; MCMC, Markov chain Monte Carlo; SNPs, single nucleotide polymorphisms

*Corresponding author. Fax: +1-607-255-4698.

E-mail address: cdb28@cornell.edu (C.D. Bustamante).

[1] Present address: Biological Statistics and Computational Biology, Cornell University, 422 Warren Hall, Ithaca, NY 14853-7801, USA.

0040-5809/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0040-5809(02)00050-3

much of the statistical work in the field of population genetics has focused on developing so called "tests of selective neutrality". This paper describes an alternative approach, which is to assume a simple (statistically identifiable) selection model and to estimate the parameters of the selection model with the understanding that the magnitude of this estimate is only as good as the validity of the assumptions of the model.

Before describing our approach, it is important to understand why what we hope to accomplish cannot be done under the standard neutral frameworks used in population genetics. In the past 25 years dozens of tests of neutrality have been proposed for a myriad of types of population genetic data [for a current review see Nielsen (2001)]. For analyzing allelic variation, several standard tests of selective neutrality are used. For example, for testing whether allelic variation at several loci sampled from a set of populations is consistent with no selection operating at any of the loci, there is the Lewontin–Krakauer test (Lewontin and Krakauer, 1973), which uses as a test statistic the heterogeneity in the scaled allelic variances across loci. For analyzing allelic variation at a single locus, the standard test is the Ewens–Watterson $\hat{F}$ statistic (1975) which tests whether the sum of the squared allelic frequencies (i.e., the sample homozygosity at a particular locus) is consistent with a K allele neutral model (Wright, 1949, 1969; Ewens, 1972). An exact test based on the Ewens sampling distribution for selectively neutral alleles (1972) has also been proposed for analyzing allelic data (Slatkin, 1994, 1996).

Several tests of selection have also been proposed for analyzing DNA sequences sampled from a single population. In one way or another, almost all the tests focus on comparing some summary statistic of the "site-frequency spectrum" to the expected distribution of the statistic under an infinite-sites neutral model. The site-frequency spectrum is the observed distribution of frequencies of mutations or the number of sites that are at a frequency 1 out of n, 2 out of $n, \ldots, n-1$ out of n in the sample where n is the sample size. The underlying rationale behind using these tests to detect selection is that different parts of the site-frequency spectrum will respond differently to selection. Under neutrality, the site-frequency spectrum has a backwards "J" shaped distribution with expected number of sites at frequency i given by θ/i for $1 \leqslant i \leqslant n-1$ where θ is the scaled neutral mutation rate. Under negative selection there will be an increase in the proportion of rare variants in the sample relative to what would be expected under neutrality. Likewise, under balancing selection, an increase in middle frequency variants is expected. Lastly, under directional selection, the site-frequency spectrum becomes "U" shaped with an excess of high- and low-frequency variants relative to mid-frequency variants. An example of how to calculate the site-frequency spectrum will be discussed in the next section.

The most popular tests in the field compare estimators of θ that depend on different parts of the site-frequency spectrum, which accounts for the differences between tests (e.g., some tests compare rare variants to mid-frequency variants while other tests compare mid-frequency variants to high-frequency variants). The most commonly used test of this form is Tajima's D (1989) which compares, θ_W, Watterson's (1975) estimate of θ, to π, the average number of pair-wise differences among sequences in the sample. θ_W is a measure heavily influenced by singletons and high-frequency variants whereas π is influenced by mid-frequency variants. The sign and magnitude of the test statistic as well as the power of the test to detect deviations from neutrality under a wide range of demographic and selective models have been explored through computer simulation (Braverman et al., 1995; Simonsen et al., 1995; Akashi, 1999). Fu and Li (1993) have also proposed several test statistics of this type that compare the number of singletons in the sample to θ_W or π in the case where the ancestral states of mutations is known as well as in the case where the ancestral states are not known. Fu (1994) has also proposed several other estimates of θ based on linear combinations of different parts of the site-frequency spectrum that lead to related tests of selective neutrality (Fu, 1996). Likewise, Fay and Wu (2000) have proposed a statistic that compares mid- and high-frequency variants in a similar type of test. One problem with all these statistics is that each statistic uses some part of the site-frequency spectrum and none of the statistics take account of all of the information in the site-frequency spectrum. We will return to this point shortly.

Another class of tests of neutrality for polymorphism data are tests of homogeneity that compare different parts of the site-frequency spectrum for two mutational classes (e.g., "silent" to "replacement" changes or "preferred" to "unpreferred" codon classes). For the specifics of these tests as well as their power see Akashi (1999). The rationale behind this class of tests is that if the two mutational classes are *neutral* and one conditions on the observed number of sites in each class, then the proportion of sites that fall into each part of the site-frequency spectrum should be the same for the two classes of mutations regardless of the underlying demographics of the species being considered.

One drawback that all of the tests mentioned above share is that only parameters in a neutral model are ever estimated, if any population genetic parameters are estimated at all. That is, one never estimates the parameter of interest, namely, the strength of selection on new mutations. The reason for this is that in the infinite-sites, no recombination coalescent-based models that are used to derive all of the statistics, the effect that selection will have on the sample statistic depends on specifying assumptions about interaction between sites, the frequency of selected mutations, and the level of recombination among sites. While several coalescent models of neutral variation linked to singly selected sites have been explored in detail for several types of selection (Kaplan et al., 1988; Hudson and Kaplan, 1988; Kaplan et al., 1989; Kaplan et al., 1991), little statistical work has been done on fitting these models to data. Likewise, recent work (Neuhauser and Krone, 1997; Krone and Neuhauser, 1997; Slade, 2000a,b, 2001) on the "ancestral selection graph" (i.e., the selection analog of the coalescent) has not provided a direct method for parameter estimation or hypothesis testing. One approach that seems quite promising is the use of simulation methods for likelihood inference in certain classes of non-neutral population genetic models where the type of the mutant allele does not depend on the type of the ancestral allele (Donnelly et al., 2001). For the infinite-sites model, though, this approach does not yet apply.

An alternative to the coalescent which does take account of all of the information in the site-frequency spectrum and therefore provide a clear likelihood framework for estimating the mutation and selection parameters in various population genetic settings is the Poisson Random Field (PRF) framework (Sawyer et al., 1987; Sawyer and Hartl, 1992; Hartl et al., 1994). In PRF models, each aligned DNA site is treated as conditionally independent given selection and mutation parameters for an infinite-sites model with haploid selection (Sawyer and Hartl, 1992). While the PRF models may not be applicable to regions that have tight linkage, they are appropriate for the analysis of genome-wide variation such as the growing data on single nucleotide polymorphisms (SNPs) from different parts of the genome. We have previously shown that the likelihood ratio test of no selection in the PRF framework has very good power to detect both negative and positive selection when the ancestral states of all mutations in the sample are known, and that confidence intervals for the selection parameter in the model based on the log-likelihood function have the desired coverage (Bustamante et al., 2001). In this paper we address the issue of "meta-analysis" of polymorphism data in the PRF framework (i.e., the issue of comparing the strength of selection among a set of loci or mutational classes of interest) and in particular, the question of how to estimate the genomic mean and variance in the strength of selection using this type of data.

Our approach is to develop a "hierarchical" PRF model of DNA polymorphism under weak selection for a set of classes of mutations where the selection parameter for each class is a random quantity sampled from some specified probability distribution. In particular, we focus on a class of hierarchical models where the mutation rates for different classes of sites are jointly independent given the vector of selection coefficients for all classes of mutations in the sample. We then develop in detail a model where the selection parameters across classes is treated as normally distributed with unknown mean and variance. We discuss how to find the maximum likelihood estimates of the mean and variance of the distribution of selection coefficients using an Expectation Maximization (EM) algorithm (Dempster et al., 1977) which consequently yields Empirical Bayes estimates of the selection and mutation parameters for each class of sites. We also develop a full Bayesian framework and discuss methods for sampling from the joint posterior distribution of all parameters in the model using Gibbs sampling. We then apply both the maximum likelihood and Bayesian methods to a sample data set generated under the assumptions of the model to illustrate the method.

2. Hierarchical modeling in the Poisson random field framework

Consider J alignments each consisting of S_j variable DNA sites where each alignment has a census size of n_j sequences for $1 \leqslant j \leqslant J$. The columns within each alignment are "exchangeable" in the statistical sense, so that position i in an alignment and position $i+1$ in the alignment tell us nothing about the actual position of the mutation in the genome. These columns *can* correspond to neighboring regions in the genome, but they do not necessarily need to and information on haplotype structure is not incorporated in the analysis since sites are assumed to be statistically independent (i.e., equivalent to free recombination between sites). The defining characteristic of the alignment is that all of the sites in the alignment share some biological characteristic of interest (e.g., they can be fourfold redundant sites from different genes in the genome, replacement mutations on the surface of proteins, or $G ::: C \rightarrow A :: T$ mutations in non-coding sequence) Without loss of generality, we assume an equal sample size and mutation rate within each alignment. We will refer to these alignments using the terms loci, mutational class, and alignment, interchangeably.

Table 1 shows an example of such an alignment for six haploid individuals sequenced at 10 segregating DNA sites, so that each column is an aligned DNA site and each row is an individual chromosome sampled from the population. At each site where the individual carries the derived nucleotide, the resulting entry in the matrix is 1, otherwise the entry is 0. The kth column total gives the number of individuals in the sample that carry the derived form of kth segregating site. Define the site-frequency spectrum for the jth alignment as the n_j-1-dimensional random vector, $X_j = X_{1,j}, X_{2,j}, \ldots, X_{n_j-1,j}$,

Table 1
Example of data modeled in this paper

Individual 1	0	0	1	0	0	0	0	1	1	0
Individual 2	0	0	1	0	1	0	0	0	1	0
Individual 3	0	0	0	1	0	1	0	0	0	1
Individual 4	0	1	1	0	0	0	1	1	1	0
Individual 5	1	1	1	0	0	0	0	1	1	1
Individual 6	0	0	0	0	0	0	1	0	1	0
Total	1	2	4	1	1	1	2	3	5	2

Each column is an aligned DNA site and each row is an individual sampled from the population. The (ith, jth) entry in the table is equal to 1 if the ith individual carries the derived form of the jth mutation and equal to 0, otherwise.

where the $X_{i,j}$ is the number of sites where i individuals carry derived forms of the mutation and $n_j - i$ individuals carry the ancestral form. For the sample data presented in Table 1, the site-frequency spectrum is $(4, 3, 1, 1, 1)$ because there are 4 sites where a single individual carries the derived mutation, 3 sites where two individuals carry the derived mutation, and 1 site each where 3, 4, or 5 individuals in the sample carry the derived mutations.

The key results from Hartl et al. (1994) that is relevant to this paper is the probability distribution for $X_{i,j}$. It can be shown that under the PRF model the $X_{i,j}$'s for a given alignment are jointly independent Poisson distributed random variables with mean $\theta_j F(i, \gamma_j)$:

$$p(X_{i,j} = x_{i,j} \mid \theta_j, \gamma_j) = e^{-\theta_j F(i,\gamma_j)} \frac{(\theta_j F(i,\gamma_j))^{x_{i,j}}}{x_{i,j}!}, \tag{1}$$

where γ_j is the scaled selection coefficient, $\frac{\theta_j}{2}$ is the scaled mutation rate for the region of interest, and

$$F(i,\gamma_j) = \int_0^1 \frac{1-e^{-2\gamma_j(1-q)}}{1-e^{-2\gamma_j}} \binom{n_j}{i} \times q^i (1-q)^{n_j - i} \frac{dq}{q(1-q)}. \tag{2}$$

By scaled mutation rate and scaled selection coefficient we mean

$$\gamma_j = 2N_e s_j,$$

$$\theta_j = 4N_e v_j,$$

where $2N_e$ is the effective population size, the fitness of mutants is $1 + s_j$, the fitness of non-mutants is 1, and v_j is the per-generation mutation rate for the jth locus in a standard Moran haploid model (Moran, 1962). This model is equivalent to a diploid model with additive fitnesses effects and effective population size chromosomal, N_e, since there are two chromosomes per individual. We will now turn to the issue of developing a hierarchical model for this type of data.

Hierarchical models provide a framework for modeling the dependence structure of parameters in a statistical model. In our case, we are interested in comparing the selection intensity among different loci and using this information to refine our estimates of selection for each locus. Two extreme approaches we could use are (1) to assume that the selection intensities are completely independent among genomic regions and that an estimate of selection from one region tell us nothing about selection in other regions, and (2) to assume that selection acts in the same way on all loci. As a middle ground we consider the model below in which the selection coefficients among loci are random variables drawn from some underlying distribution.

Consider a set of J loci where locus j has site frequency vector, x_j, and selection parameter γ_j, with likelihood $p(x_j \mid \gamma_j)$. We are interested in modeling variation in selection intensity using some meta-distribution that depends on a set hyperparameters ϕ so that $p(\gamma_j \mid \phi)$ is the probability distribution for selection coefficient γ_j before we have observed any data. We set up the model so that, conditional on ϕ, the γ_j's are independent and identically distributed. That is, the joint distribution of the parameters $p(\boldsymbol{\gamma} \mid \phi)$ is equal to the product of the individual marginal distributions:

$$p(\boldsymbol{\gamma} \mid \phi) = \prod_{j=1}^{J} p(\gamma_j \mid \phi), \tag{3}$$

where $\boldsymbol{\gamma}$ is the J-dimensional vector of selection coefficients, $(\gamma_1, \gamma_2, \ldots, \gamma_J)$.

In a frequentist (maximum likelihood) setting, we wish to find the vector of parameter estimates $\hat{\phi}$ that maximizes the marginal likelihood function, $L(\phi \mid x)$,

$$\begin{aligned} L(\phi \mid \mathbf{x}) &= p(\mathbf{x} \mid \phi) \\ &= \prod_{j=1}^{J} p(x_j \mid \phi) \\ &= \prod_{j=1}^{J} \int_{\gamma_i \in \Gamma} p(x_i \mid \gamma_i, \phi) p(\gamma_i \mid \phi)\, d\gamma_j \\ &= \prod_{j=1}^{J} \int_{\gamma_i \in \Gamma} p(x_i \mid \gamma_i) p(\gamma_i \mid \phi)\, d\gamma_j, \end{aligned} \tag{4}$$

where $\mathbf{x}$ is the J-dimensional vector of site-frequencies $(x_1, x_2, \ldots, x_J)$. The simplification from the second line to the third line comes from the fact that the hyperparameters affect the likelihood function only through $\boldsymbol{\gamma}$.

In a Bayesian framework, we would often be interested in describing the joint and marginal posterior distributions of the parameters in the model (i.e., the selection coefficient for each locus as well as the hyperparameters). To obtain these posterior distributions, we need to specify the prior distribution on the hyperparameters, $p(\phi)$, which often takes the form of a proper distribution with "uninformative" parameters. For the model we have presented above, the joint posterior distribution is given by

$$\begin{aligned} p(\phi, \boldsymbol{\gamma} \mid \mathbf{x}) &\propto p(\mathbf{x} \mid \phi, \boldsymbol{\gamma}) p(\phi, \boldsymbol{\gamma}) \\ &= p(\mathbf{x} \mid \boldsymbol{\gamma}) p(\phi, \boldsymbol{\gamma}) \\ &= p(\phi) p(\boldsymbol{\gamma} \mid \phi) p(\mathbf{x} \mid \boldsymbol{\gamma}). \end{aligned} \tag{5}$$

Again, the step from the first line to the second line holds because $p(\mathbf{x} \mid \phi, \boldsymbol{\gamma})$, depends only $\boldsymbol{\gamma}$. The step from the second line to the third line comes directly from the definition of conditional probability.

The result in Eq. (1) provides the first level of our model. Let $\mathbf{x} = (x_1, x_2, \ldots, x_J)$ be the vector of site frequencies, $\mathbf{n} = (n_1, n_2, \ldots, n_J)$ the vector of sample sizes, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_J)$ the vector of selection coefficients, and $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_j)$ the vector of mutation rates for each of J loci $(1 \leqslant j \leqslant J)$. Note that although $\mathbf{x}$ is a "vector of vectors" it is not necessarily a matrix, since the sample sizes for the alignments may differ. The likelihood function for each locus is the product of the individual $p(x_{i,j} \mid \theta_j, \gamma_j)$ since the $x_{i,j}$'s are independent:

$$p(x_j \mid \theta_j, \gamma_j) = \frac{e^{-\theta_j \sum_{i=1}^{n_j-1} F(i,\gamma)} \theta_j^{S_j} \prod_{i=1}^{n_j-1} F(i, \gamma_j)^{x_{i,j}}}{\prod_{i=1}^{n_j-1} x_{i,j}!}, \tag{6}$$

where S_j is the number of segregating sites in the jth gene $(S_j = \sum_{i=1}^{n_j-1} x_{i,j})$.

Since we are interested in modeling variation among loci in selection, but not in mutation, it would be convenient to develop a model where the mutation rate for each locus θ_j has been integrated out of the model above. Setting a gamma prior distribution for θ_j and using Bayes rule, it is possible to find $p(x_j \mid \gamma_j)$ analytically.

Let

$$\theta_j \sim \text{Gamma}(\alpha_j, \beta_j), \tag{7}$$

where $\sim$ means "distributed as" and $\text{Gamma}(\alpha, \beta)$ represents a gamma distributed random variable with mean $\frac{\alpha_j}{\beta_j}$. By Bayes rule,

$$p(x_j \mid \gamma_j) = \frac{p(x_j \mid \gamma_j, \theta_j) p(\theta_j)}{p(\theta_j \mid x_j, \gamma_j)}. \tag{8}$$

Since the gamma distribution is the natural conjugate of the Poisson distribution, it can be shown analytically that the posterior distribution of θ_j will also be a gamma distribution with updated parameters that incorporate the data, so that

$$\theta_j \mid x_j, \gamma_j \sim \text{ Gamma } \left(\alpha_j + S_j, \beta_j + \sum_{i=1}^{n_j-1} F(i, \gamma_j) \right). \tag{9}$$

By substituting (6), (7), and (9) into (8), it follows that the marginal distribution of the data $p(x_j \mid \gamma_j)$ is a parameterized negative binomial distribution,

$$p(x_j \mid \gamma_j) = \frac{\dfrac{e^{-\theta_j \sum_{i=1}^{n_j-1} F(i,\gamma)} \theta_j^{S_j} \prod_{i=1}^{n_j-1} F(i, \gamma_j)^{x_{i,j}}}{\prod_{i=1}^{n_j-1} x_{i,j}!} \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)} \theta_j^{\alpha_j - 1} e^{-\beta_j \theta_j}}{\dfrac{(\beta_j + \sum_{i=1}^{n_j-1} F(i, \gamma_j))^{\alpha_j + S_j}}{\Gamma(\alpha_j + S_j)} \theta_j^{\alpha_j + S_j - 1} e^{-(\beta_j + \sum_{i=1}^{n_j-1} F(i,\gamma_j))\theta_j}}. \tag{10}$$

Keeping only terms that depend on γ_j, we can simplify the above equation and focus on the portion that contributes to the likelihood function for γ_j,

$$p(x_j \mid \gamma_j) \propto \frac{\prod_{i=1}^{n_j-1} F(i, \gamma_j)^{x_{i,j}}}{(\beta_j + \sum_{i=1}^{n_j-1} F(i, \gamma_j))^{(S_j + \alpha_j)}}. \tag{11}$$

It may seem that integrating out $\boldsymbol{\theta}$ by specifying independent gamma prior distributions for each θ_j would compromises maximum likelihood inference. An alternative would be to optimize the joint likelihood function $L(\boldsymbol{\theta}, \boldsymbol{\gamma} \mid \mathbf{x})$ for $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$. However, as shown below, the assumption of jointly independent gamma prior distributions for the θ_j's affects the inference on $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ minimally while greatly reducing the dimensionality of the maximization problem as long as α_j and β_j are small.

We have previously shown (Hartl et al., 1994; Bustamante et al., 2001) that the maximum likelihood estimates of γ_j and θ_j in a model where the alignments are fully independent (i.e., not the hierarchical model) are given by the point $(\widehat{\theta_j}, \widehat{\gamma_j})$ where $\widehat{\gamma_j}$ is the value of γ_j that maximizes the log-profile-likelihood function of γ_j, $l^*(\gamma_j \mid x_j)$, and $\widehat{\theta_j}$ is the conditional maximum likelihood estimate of θ_j given γ_j, $\tilde{\theta}$, evaluated at $\hat{\gamma}$. The expressions for $l^*(\gamma_j \mid x_j)$ and $\widetilde{\theta_j}$ are as follows:

$$l^*(\gamma_j \mid x_j) \propto \sum_{i=1}^{n_j-1} x_{i,j} F(i, \gamma_j) - S_j \log \sum_{i=1}^{n_j-1} F(i, \gamma_j) \tag{12}$$

and

$$\widetilde{\theta_j} = \frac{S_j}{\sum_{i=1}^{n_j-1} F(i, \gamma_j)}, \tag{13}$$

where S_j is the number of segregating sites in locus j.

Taking the logarithm of (11), we see that the integrated log-likelihood function $l(\gamma_j \mid x_j)$ is proportional to

$$l(\gamma_j \mid x_j) \propto \sum_{i=1}^{n_j-1} x_{i,j} \log F(i,\gamma_j) - (S_j + \alpha_j) \times \log\left(\sum_{i=1}^{n_j-1} F(i,\gamma_j) + \beta_j\right). \quad (14)$$

Clearly if α_j and β_j are equal to zero, Eqs. (12) and (14) are identical. This shows that the maximum likelihood of γ_j for the integrated likelihood and the joint-likelihood is the same if α_j and β_j are equal to zero. This means that for of modeling variation in γ_j's, no information is lost by integrating out the θ_j's if α_j and β_j are small.

Likewise, in (9) it is shown that the distribution of θ_j given γ_j and x_j is a gamma distribution with parameters $\alpha_j + S_j$ and $\beta_j + \sum_{i=1}^{n_j-1} F(i,\gamma_j)$. The maximum likelihood estimate of θ_j in the integrated model will equal the mode of this distribution, which one can show to be $\frac{\alpha_j+S_j-1}{\beta_j+\sum_{i=1}^{n_j-1} F(i,\gamma_j)}$. If the number of segregating sites is large and α_j and β_j are small, this will be very close to the MLE of θ_j for the joint likelihood function given in Eq. (13).

Completing the first level of the hierarchical model, we note that the conditional posterior distribution for γ_j is given by

$$p(\gamma_j \mid \phi, x_j) \propto p(x_j \mid \gamma_j) p(\gamma_j \mid \phi), \quad (15)$$

where $p(\gamma_j \mid \phi)$ is the probability density function for the meta-distribution of selection coefficients.

Since selection coefficients are defined on the real line (i.e., in theory they can range from $-\infty$ to ∞) one model that is appropriate mathematically for the distribution of selection coefficients is that they are independent and identically normally distributed with mean μ and variance σ^2:

$$p(\gamma_j \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\gamma_j-\mu)^2}{2\sigma^2}}. \quad (16)$$

The independence assumption yields J independent conditional posterior distributions

$$p(\gamma_j \mid \mu, \sigma^2, x_j) \propto p(x_j \mid \gamma_j) p(\gamma_j \mid \mu, \sigma^2) \quad (17)$$

The next two sections of the paper will focus on how to proceed with statistical inference for the normal model above from Bayesian and frequentist perspectives.

3. Frequentist inference

In a frequentist framework, we wish to find the maximum likelihood estimates of μ and σ^2, that is, values of μ and σ^2, denoted $\hat{\mu}$ and $\widehat{\sigma^2}$, that maximize

$$p(\mathbf{x} \mid \mu, \sigma^2) = \prod_{j=1}^{J} \int_{-\infty}^{\infty} p(x_j \mid \gamma_j) \frac{1}{\sqrt{\pi 2\sigma^2}} e^{-\frac{(\gamma_j-\mu)^2}{2\sigma^2}} \, d\gamma_j. \quad (18)$$

To perform this optimization, we use the Expectation Maximization algorithm (Dempster et al., 1977).

The Expectation Maximization (EM) algorithm is an iterative method for maximum-likelihood fitting of parametric statistical models (Dempster et al., 1977). In EM the maximization problem at hand is framed in terms of some data augmentation problem. That is the problem is posed in such a way that the addition of some key missing data makes maximization of the likelihood function "easy". The EM algorithm consists of two iterated steps. In the expectation step (E-) one calculates the expected value of the log-likelihood function of the *augmented* data (i.e., the log-likelihood function of the observed and missing data) with respect to the distribution of the missing data, given the observed data and the previous M- step estimates of the parameters to be estimated. This (E-) function is denoted $Q(\phi, \phi^{(i-1)})$ where $\phi^{(i)}$ is the parameter estimates at step i. In the maximization step (M-) one maximizes $Q(\phi, \phi^{(i-1)})$ with respect to ϕ.

Let x represent the observed data in some parametric statistical model and ϕ the parameters of the model, where x and ϕ can be either scalars or vectors. Our objective is to maximize the likelihood function of the observed data, $p(x \mid \phi)$. Define y as the "missing" or unobserved data. In each iteration i of the EM algorithm, the goal is to find $\phi^{(i)}$ such that

$$\phi^{(i)} = \underset{\phi}{\operatorname{argmax}}\ Q(\phi, \phi^{(i-1)}), \quad (19)$$

where

$$Q(\phi, \phi^{(i-1)}) = \mathbb{E}[\log p(x, y \mid \phi) \mid x, \phi^{(i-1)}] \quad (20)$$

$$= \int_{y \in Y} \log p(x, y \mid \phi) p(y \mid x, \phi^{(i-1)}) \, dy. \quad (21)$$

For the hierarchical PRF model outlined in the previous section, the function to maximize is the marginal likelihood function of the hyperparameters of the meta-distribution of selection coefficients, $L(\phi \mid \mathbf{x}) = p(\mathbf{x} \mid \phi)$. This means we want to find the vector ϕ that maximizes

$$p(\mathbf{x} \mid \phi) = \int_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}} p(\mathbf{x} \mid \boldsymbol{\gamma}, \phi) p(\boldsymbol{\gamma} \mid \phi) \, d\boldsymbol{\gamma}. \quad (22)$$

where ϕ is an h-dimensional vector of hyperparameters of the meta-distribution ($h<J$), $\mathbf{x}$ is the vector of site-frequency spectra for the J loci in the study and $\boldsymbol{\gamma}$ is the vector of selection coefficients. We use the ϕ notation for generality, since we can specify a series of meta distributions for $\boldsymbol{\gamma}$; in the case of the normal distribution model specified above, $h = 2$ and $\phi = (\mu, \sigma^2)$. As will be

shown below, the missing data in this problem is the vector of selection coefficients.

The $Q(\phi, \phi^{(i-1)})$ function for the hierarchical PRF model is

$$Q(\phi, \phi^{(i-1)}) = \int_{\gamma \in \boldsymbol{\Gamma}} \log(p(\mathbf{x}, \gamma \mid \phi)) p(\gamma \mid \phi^{(i-1)}, \mathbf{x})\, d\gamma \quad (23)$$

$$= \int_{\gamma \in \boldsymbol{\Gamma}} \log(p(\mathbf{x} \mid \gamma) p(\gamma \mid \phi)) \times p(\gamma \mid \phi^{(i-1)}, \mathbf{x})\, d\gamma \quad (24)$$

where the step from (23) to (24) comes from the definition of conditional probability [i.e., $p(\mathbf{x}, \gamma \mid \phi) = p(\mathbf{x} \mid \gamma, \phi) p(\gamma \mid , \phi)$] and the fact that ϕ affects $\mathbf{x}$ only through γ so that $p(\mathbf{x} \mid \gamma, \phi) = p(\mathbf{x} \mid \gamma)$. Note that the expectation taken in (23) and (24) is over the joint conditional posterior distribution of γ_j's. Given the ϕ, the conditional posterior distribution of the γ_j's are independent, so that $p(\gamma \mid \phi^{(i-1)}, \mathbf{x})$ factors into the product of the individual conditional posterior distributions. We will return to this point shortly.

Given the fact that $\log(p(\mathbf{x} \mid \gamma) p(\gamma \mid \phi)) = \log p(\mathbf{x} \mid \gamma) + \log p(\gamma \mid \phi)$ and that $p(\mathbf{x} \mid \gamma)$ does not depend on ϕ (the vector of parameters we want to maximize), we can rewrite $Q(\phi, \phi^{(i-1)})$ as

$$Q(\phi, \phi^{(i-1)}) = C_i(\gamma, \mathbf{x}, \phi^{(i-1)}) + \int_{\gamma \in \boldsymbol{\Gamma}} \log(p(\gamma \mid \phi)) \times p(\gamma \mid \phi^{(i-1)}, \mathbf{x})\, d\gamma, \quad (25)$$

where

$$C_i(\gamma, \mathbf{x}, \phi^{(i-1)}) = \int_{\gamma \in \boldsymbol{\Gamma}} \log(p(\mathbf{x} \mid \gamma)) p(\gamma \mid \phi^{(i-1)}, \mathbf{x})\, d\gamma \quad (26)$$

is a constant that does not depend on ϕ.

To maximize (25) we need to find the vector of parameter estimates $\phi^{(i)}$ such that all components of the vector of partial derivatives of $Q(\phi, \phi^{(i-1)})$ evaluated at $\phi^{(i)}$ are 0 (i.e., $\frac{\partial Q(\phi, \phi^{(i-1)})}{\partial \phi_i} \big|_{\phi = \phi^{(i)}} = 0$ for $i = 1, \ldots, h$). How to perform this maximization depends on the model specified for the hyper-distribution. For the case of the normal model specified in the previous section, $\phi = (\mu, \sigma^2)$, we can find $\phi^{(i)}$ analytically in terms of the posterior means and variances of $p(\gamma_j \mid \mu^{(i-1)}, \sigma^{2\,(i-1)}, x_j)$. To do so, we need to find

$$\frac{\partial Q(\mu, \sigma^2, \mu^{(i-1)}, \sigma^{2\,(i-1)})}{\partial \mu} = \mathbb{E}\left[\frac{\partial \log p(\gamma \mid \mu, \sigma^2)}{\partial \mu} \bigg| \gamma, \mu^{(i-1)}, \sigma^{2\,(i-1)} \right] \quad (27)$$

and

$$\frac{\partial Q(\mu, \sigma^2, \mu^{(i-1)}, \sigma^{2\,(i-1)})}{\partial \sigma^2} = \mathbb{E}\left[\frac{\partial \log p(\gamma \mid \mu, \sigma^2)}{\partial \sigma^2} \bigg| \gamma, \mu^{(i-1)}, \sigma^{2\,(i-1)} \right]. \quad (28)$$

It is a matter of differentiation and algebra to show that

$$\frac{\partial \log p(\gamma \mid \mu, \sigma^2)}{\partial \mu} = -\frac{1}{\sigma^2} \sum_{j=1}^{J} (\gamma_j - \mu) \quad (29)$$

$$\frac{\partial \log p(\gamma \mid \mu, \sigma^2)}{\partial \sigma^2} = -\frac{J}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{j=1}^{J} (\gamma_i - \mu)^2 \quad (30)$$

$$= -\frac{J}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{j=1}^{J} (\gamma_i^2 - 2\,\gamma_i \mu + \mu^2). \quad (31)$$

As noted above, the expectations in (27) and (28) are taken over the individual conditional posterior distributions of the γ_j's. This amounts to replacing γ_j and γ_j^2 in (30) and (31) with $\mathbb{E}[\gamma_j]$ and $\mathrm{VAR}[\gamma_j] + \mathbb{E}^2[\gamma_j]$, respectively, where $\mathbb{E}[\gamma_j]$ is the expectation and $\mathrm{VAR}[\gamma_j]$ the variance of the conditional posterior distribution, $p(\gamma_j \mid x_j, \mu^{(i-1)}, \sigma^{2\,(i-1)})$. Hence,

$$\mu^{(i)} = \sum_{j=1}^{J} \frac{\mathbb{E}[\gamma_j]}{J} \quad (32)$$

$$\sigma^{2\,(i)} = \sum_{j=1}^{J} \frac{\mathrm{VAR}[\gamma_j]}{J} + \sum_{j=1}^{J} \frac{(\mathbb{E}[\gamma_j] - \mu^{(i)})^2}{J}. \quad (33)$$

In words, Eqs. (32) and (33) state that at the end of each iteration, the new maximum likelihood estimate of the mean of the distribution of selection coefficients is the mean of the expected values of the individual conditional posterior distributions, and the MLE of the variance is the sum of the average of the conditional variances plus the variance of the expected values of the conditional distributions.

Once the EM algorithm has converged (i.e., once there is no change in the log-likelihood from iteration to iteration) the MLEs of μ and σ^2, $\hat{\mu}$ and $\widehat{\sigma^2}$, will be equal to $\mu^{(i)}$ and $\sigma^{2\,(i)}$. One added benefit of using this EM algorithm is that at convergence we have also found $\gamma_j^{\mathrm{EB}} = \mathbb{E}[\gamma_j \mid x_j, \hat{\mu}, \widehat{\sigma^2}]$ the expected value of the γ_j's given the maximum likelihood estimates of μ and σ^2. The γ_j^{EB} values are known as the Empirical Bayes estimates of the γ_j. Likewise, we can also compute the Maximum a posteriori estimate of γ_j, γ_j^{MAP}, which is defined as the mode of the distribution $p(\gamma_j \mid \mathbf{x}, \hat{\mu}, \widehat{\sigma^2})$.

4. Bayesian inference

To define a Bayesian analog of the hierarchical model outlined above, we need to specify prior distributions for μ and σ^2. A logical framework to employ is the conjugate model:

$$\mu \mid \sigma^2 \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0} \right), \quad (34)$$

$$\sigma^2 \sim \text{Inv-}\chi^2(v_0, \sigma_0^2), \quad (35)$$

where $\mu_0, \kappa_0, v_0, \sigma_0^2$ are the parameters of the prior distributions for μ and σ^2, and Inv-χ^2 refers to an inverse χ^2 distribution.

The conditional posterior distribution of μ for the conjugate normal model is

$$\mu \mid \sigma^2, \gamma \sim N\left(\frac{\frac{\kappa_0}{\sigma^2}\mu_0 + \frac{J}{\sigma^2}\bar{\gamma}}{\frac{\kappa_0}{\sigma^2} + \frac{J}{\sigma^2}}, \frac{1}{\frac{\kappa_0}{\sigma^2} + \frac{J}{\sigma^2}}\right), \tag{36}$$

where $\bar{\gamma}$ is the arithmetic average of the γ's.

The marginal posterior distribution of σ^2 is

$$\sigma^2 \mid \gamma \sim \text{Inv-}\chi^2(v_J, \sigma_J^2), \tag{37}$$

where

$$v_J = v_0 + J, \tag{38}$$

$$v_J\sigma_J^2 = v_0\sigma_0^2 + (J-1)s^2 + \frac{\kappa_0 J}{\kappa_0 + J}(\bar{\gamma} - \mu_0)^2 \tag{39}$$

and s^2 is the sample variance of the γ's. For this model, if κ_0 and v_0 are chosen to be small and σ^2 to be large, the prior distribution will be uninformative. Our next task is to develop a method for sampling from the joint posterior distribution $p(\gamma_1, \gamma_2, \ldots, \gamma_J, \mu, \sigma \mid \mathbf{x})$. A natural method to implement for the normal conjugate model is Gibbs sampling.

4.1. Gibbs sampling

The Metropolis–Hastings algorithm Metropolis et al., 1953; Hastings, 1970) provides a general method for generating a Markov chain with some stationary distribution of interest by sampling from some proposal distribution and accepting new proposed values with particular probabilities. In our case, we are interested in sampling from the joint posterior distribution of the parameters and hyperparameters, $p(\boldsymbol{\gamma}, \mu, \sigma^2 \mid \mathbf{x})$. An efficient way to sample from this distribution is to use a form of the Metropolis–Hastings algorithm known as Gibbs sampling or alternating conditional sampling. In this case, each iteration of the algorithm will consist of $J+2$ steps in which we sample from each component of $p(\boldsymbol{\gamma}, \mu, \sigma^2 \mid \mathbf{x})$ conditional on the previous values of the other components and we update the parameter with probability 1. We will denote the samples at each iteration by adding a subscript t to $\boldsymbol{\gamma}, \mu$, and σ^2 so that $\gamma_{t.j}$, μ_t, and σ_t^2 are the values of γ_j, μ, and σ^2, respectively, at iteration t and $\boldsymbol{\gamma}_t = (\gamma_{t.1}, \gamma_{t.2}, \ldots, \gamma_{t.J})$. In particular, to implement Gibbs sampling we used the following algorithm.

1. Start at iteration $t = 1$ with $\mu = \mu_1$ and $\sigma^2 = \sigma_1^2$ as given starting values.
2. Given μ_t and σ_t^2, calculate the conditional posterior distribution $p(\gamma \mid \mu_t, \sigma_t^2, x_j)$ numerically on a grid of points as given by (17), normalize to unity, and use the inverse-cdf method to draw a sample, $\gamma_{t,j}$, from $p(\gamma_j \mid \mu_t, \sigma_t^2, x_j)$ for each locus. Since the γ_j's are conditionally independent given μ and σ^2, it does not matter in which order the $\gamma_{t,j}$'s are sampled.
3. *Optional*: For $j = 1, \ldots, J$, sample θ_j from $p(\theta_j \mid \gamma_j, x_j)$ using standard techniques for sampling from a Gamma distribution and the result from (9).
4. Increment t by 1.
5. Given the $\boldsymbol{\gamma}_{t-1}$, sample σ_t^2 from $p(\sigma^2 \mid \boldsymbol{\gamma}_{t-1})$ (37) using standard techniques for sampling from an inverse Gamma distribution.
6. Given σ_t^2, sample μ^t from $p(\mu \mid \sigma_t^2, \boldsymbol{\gamma}_{t-1})$ (36) using standard techniques for sampling from a normal distribution.
7. Repeat steps 2–6 until some desired convergence criteria is met.

The Gibbs sampler for the model outlined above was implemented in an ANSI C program available from *www.bscb.cornell.edu/Homepages/Carlos_Bustamante/*.

4.1.1. Convergence criteria

To monitor convergence of the MCMC chains we chose to use the $\sqrt{\hat{R}}$ statistic discussed by Gelman et al. (1997). This measure compares the between- and within-sequence variances for multiple Markov chains started at different starting points. The general "rule of thumb" is to run the chains until $\sqrt{\hat{R}}$ is below 1.2 (or some other value "close" to 1) and then discard the first half of the observations up to that point from each chain. Combining the data from all the chains, one has mostly independent samples from the joint-posterior distribution $p(\boldsymbol{\gamma}, \mu, \sigma \mid \mathbf{x})$.

5. Example

To illustrate how the model can be used to analyze variation in selection among genomic regions or mutational classes, we will focus the remainder of the paper on a particular example. The site-frequency data in Fig. 1 were simulated by first drawing 25 selection coefficients ($\gamma_j, 1 \leqslant j \leqslant 25$) from a normal distribution with $\mu = 2$ and $\sigma = 2.22$ and then for each site-frequency class drawing a Poisson random variable with probabilities given by Eq. (1). For simplicity we used the same sample size ($n = 15$) for each mutational class as well as the same mutation rate ($\theta = 25$).

For maximum likelihood estimation of μ and σ^2 we used the EM algorithm with nine different starting points. On a desktop computer with a Pentium III 900 MHz processor, it took less than 10 min to run all nine analyses using a grid of 200 points for the posterior distribution of each gamma. The algorithm in all cases converged to the fixed point $\hat{\mu} = 2.63$ and $\hat{\sigma} = 2.12$. We increased the grid size to 500 points and the results remained unchanged. Note that confidence intervals for the MLEs based on the joint likelihood function do

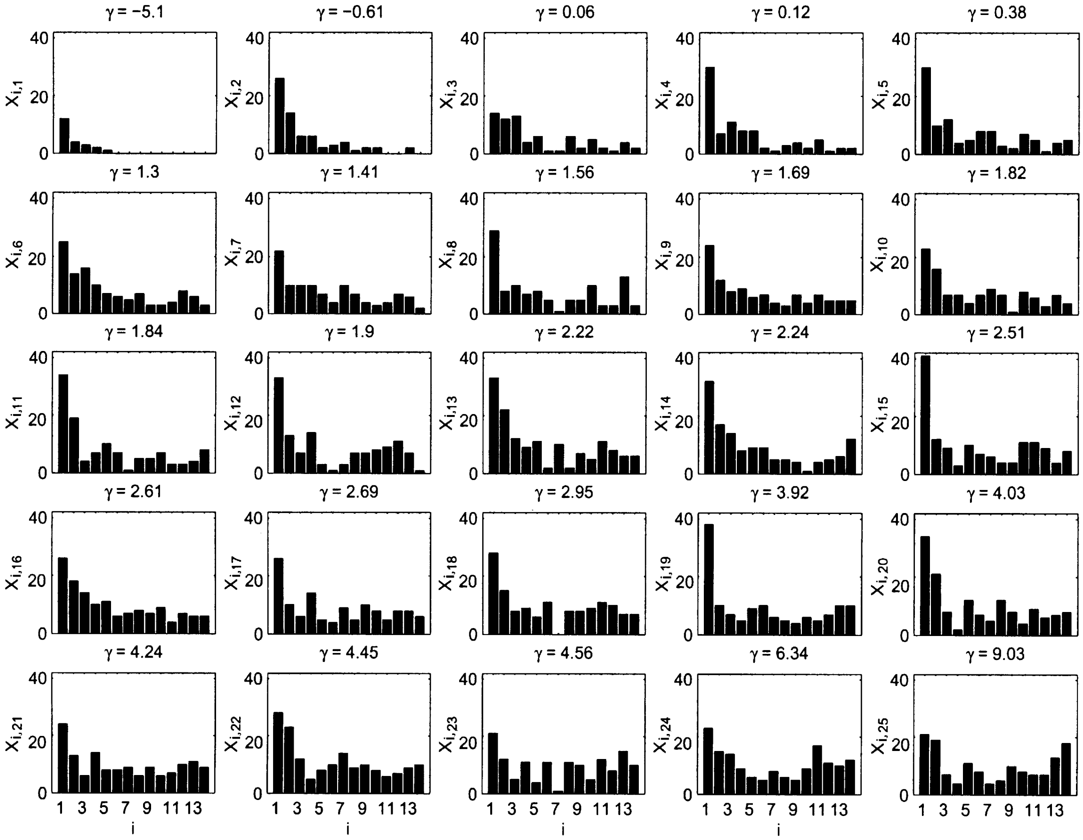

Fig. 1. Simulated site-frequency data for 25 genes.

contain the true values of the parameters, $\mu = 2$ and $\sigma = 2.22$.

For fitting the Bayesian model we used Gibbs sampling as described above for nine different starting points with $\alpha_j = \beta_j = 0$ for all chains. We ran each of the nine chains for 10,500 steps using a grid of 200 points on $[-10, 30]$ for the posterior distribution of each γ. This analysis took 4.5 h to run on the same computer described above. The $\sqrt{\hat{R}}$ was well below 1.2 after the first 100 iterations. To be conservative, we discarded the first 500 draws of each chain. Fig. 2 shows the first 500 iterations for μ and σ to show that the chains had mixed well before the point at which samples were retained. Fig. 3 shows histograms of 500 bins each to summarize the marginal posterior distributions of μ and σ as estimated from the 90,000 total draws in the nine MCMC chains. The quantiles of these distribution as well as the posterior mean and variance are summarized in Table 2. We see that the true mean and true variance are well within the region encompassed by the posterior distribution.

Figs. 4 and 5 illustrate the performance of the frequentist and Bayesian point estimators of γ_j and θ_j introduced in this paper. In general, the Maximum a posteriori estimate (MAP) and the Emperical Bayes estimate (EB) are quite similar to each other and to the mean and median of the posterior distribution. In Fig. 5, we see that the maximum likelihood estimate of γ_j for some genes is closer to the true value of the parameter than either the MAP or the EB estimate, but in certain cases the MLE yields a value much greater than the true value (e.g., $j = 25$). Likewise, we note that for 24 out of 25 of the mutation classes, both the 95% credibility intervals and the 95% frequentist confidence intervals for γ_j based on the log-likelihood function contain their true value. One clear advantage of Bayesian approach in this regard is that the lengths of the credibility intervals are, on average, noticeably shorter than the corresponding confidence intervals based on the profile-log-likelihood functions.

One method that could be used to estimate the mean and variance of the distribution of selection coefficients is to calculate the average of the individual maximum likelihood estimates and use the sample variance to construct a confidence interval for the true mean. The average of the MLEs is 2.98 and the sample variance is 16.425, resulting in a confidence interval for the true mean of $(1.31, 4.65)$ which does contain the true mean of the distribution. Note, however, that both the maximum likelihood estimate of μ as well as the posterior mean and median are closer to the true value than this ad hoc estimate. Note, also, that since

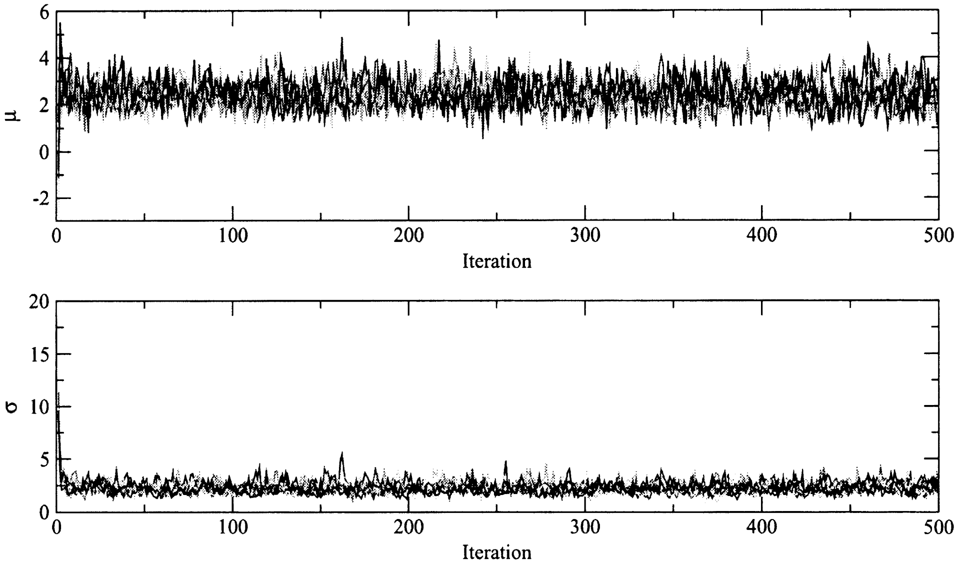

Fig. 2. Time-series plots of the first 500 draws of nine Markov Chains used to estimate the joint posterior distribution of the parameters in the Hierarchical PRF model.

Table 2
Summary of maximum likelihood estimates of μ and σ as well as posterior mean (PM), variance (PV), and quantiles of the posterior distributions of μ and σ in the example dataset.

Parameter	MLE	PM	PV	Quantiles of posterior distribution				
				2.5%	25.0%	50.0%	75.0%	97.5%
μ	2.63	2.42	0.30	1.42	2.05	2.40	2.77	3.58
σ	2.12	2.28	0.25	1.48	1.93	2.22	2.57	3.43

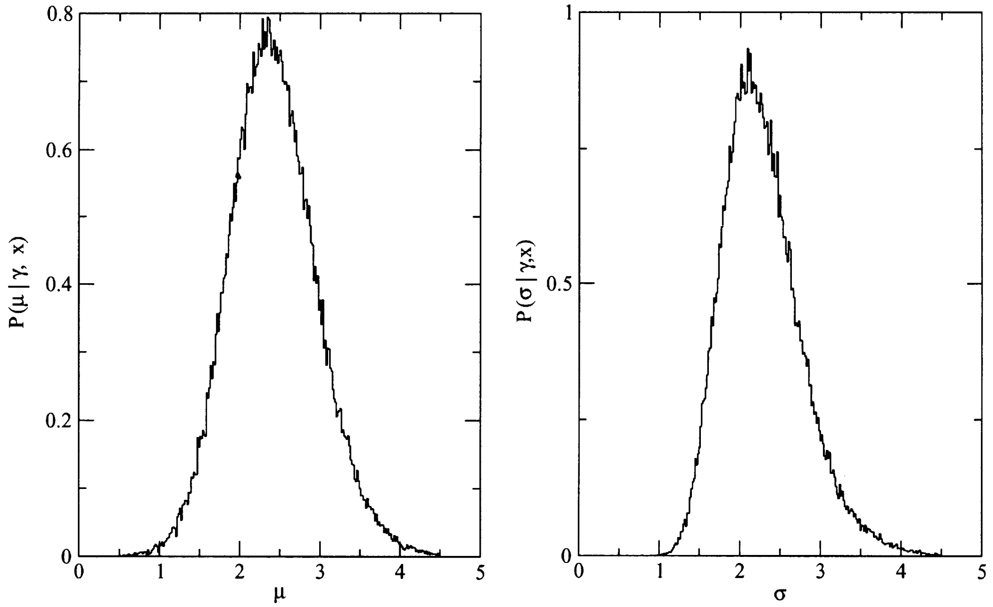

Fig. 3. Marginal posterior distributions of μ and σ as estimated from 9 chains of 10,000 draws each (90,000 draws total).

the variance of the MLEs is a composite of the variance between the γ_j's as well as the estimation variance of each γ_j, it cannot be used to estimate the variance between γ_j without including some term to account for the variance of each estimate. We have previously shown (Bustamante et al., 2001) that the asymptotic

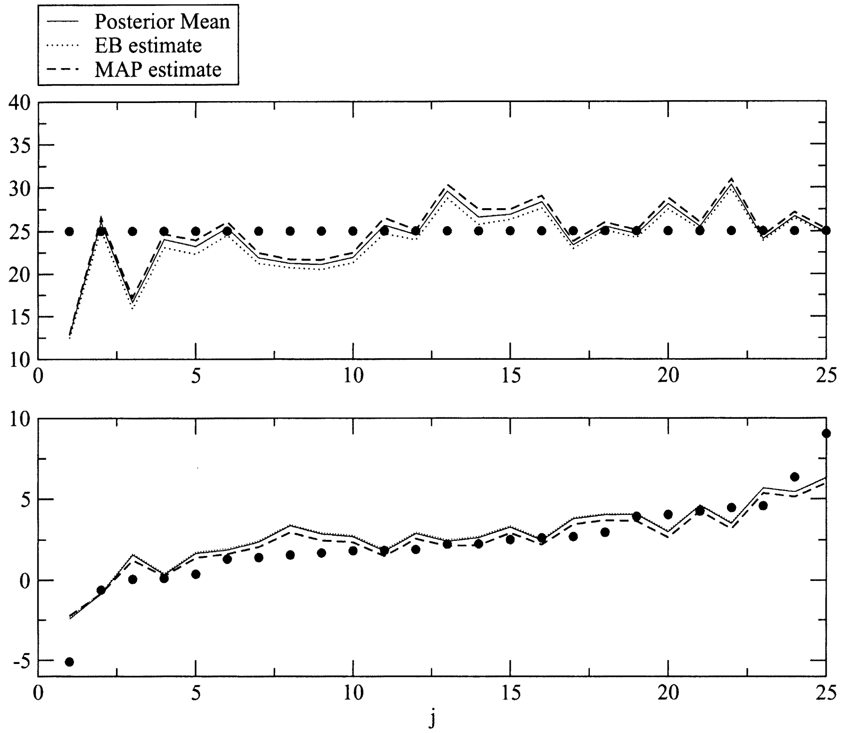

Fig. 4. Posterior mean, emperical Bayes estimate, and maximum a posteriori estimates (MAP) of γ_j and θ_j for the simulated data as compared to their true value.

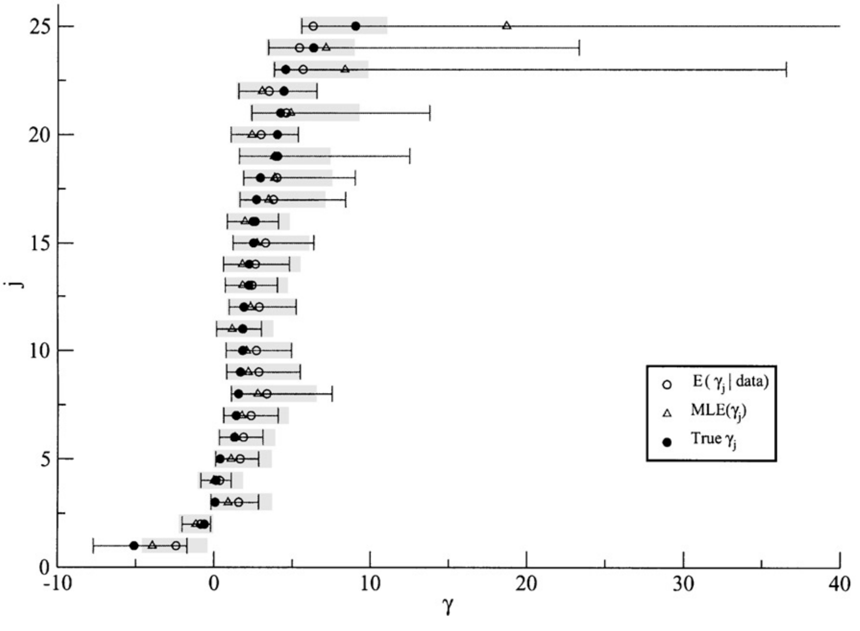

Fig. 5. Coverage of confidence intervals and credibility intervals for γ parameters using the simulated data. The black lines detail 95% confidence intervals based on the set of values of γ_j that are within 1.92 profile-log-likelihood units from the MLE, and the grey lines detail 95% highest posterior credibility intervals.

variance of $\hat{\gamma}$ (which would be the natural estimate of the estimation variance to use) is a poor estimate of the true variance if the number of segregating sites is not very large ($\approx$2500 for each locus). If we simply treat the MLEs as data, we would therefore tend to over-estimate the variation in selection among classes of mutations. Therefore, both the maximum likelihood and the Bayesian methods are superior for estimating the

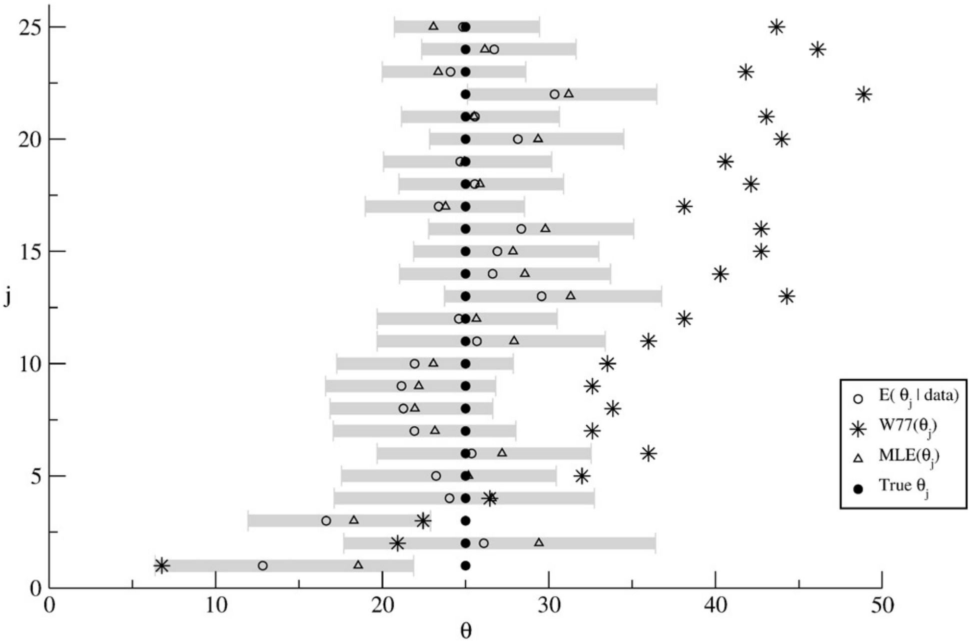

Fig. 6. Coverage of 95% Credibility Intervals for θ parameters using the simulated data and performance of point estimators.

distribution of selection coefficients than simply using the MLEs as data.

Lastly, we turn to the estimation of θ_j's. We see in Figs. 4 and 6 that the maximum likelihood estimates, the MAP estimates, and the estimates from the posterior mean are all quite close to each other and, in general, closer to the true value of $\theta = 25$ for each gene than θ_W. The reason we included Watterson's (1975) estimate of θ is that it is the most commonly used estimate of θ and because we have previously shown (Bustamante et al., 2001) that under neutrality for the PRF model, θ_W is the maximum likelihood estimate of θ. The reason our approach can improve upon the estimate so dramatically is that the estimate of θ_j is completely dependent on the estimate of γ_j. As long as one has some reasonable estimate of γ_j, one will always improve upon an estimate of θ_j that assumes neutrality.

For this example, it is clear that maximum likelihood and Bayesian analyses give quite similar results. The advantage of using the maximum likelihood method is that it is much faster to run than the Bayesian method and therefore makes the study of power and other statistical properties of the system computationally tractable. The advantage of the Bayesian method is that it is more flexible than the maximum likelihood method and can easily be modified to accommodate more complicated models. Both methods are superior to simply using the MLEs as data and estimating the distribution of γ_j's from $\widehat{\gamma}_j$'s. We have also implemented the same model above in a program using a Metropolis–Hastings algorithm to sample from the γ_j's which does not require calculating the posterior distribution of the γ_j's on a grid. This method gave quite similar results to Gibbs sampler but took much longer to run. We hope to focus a subsequent publication on the power and robustness of the model described here.

References

Akashi, H., 1999. Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics 151, 221–238.

Braverman, J.M., Hudson, R.R., Kaplan, N.L., Langley, C.H., Stephan, W., 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140, 783–796.

Bustamante, C.D., Wakeley, J., Sawyer, S., Hartl, D.L., 2001. Directional selection and the site-frequency spectrum. Genetics 159, 1779–1788.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. Roy. Statist. Soc. Ser. B, Methodological 39, 1–37.

Donnelly, P., Nordborg, M., Joyce, P., 2001. Likelihoods and simulation methods for a class of nonneutral population genetics models. Genetics 159, 853–867.

Ewens, W.J., 1972. The sampling theory of selectively neutral alleles. Theor. Popul. Biol. 3, 87–112.

Fay, J.C., Wu, C.I., 2000. Hitchhiking under positive Darwinian selection. Genetics 155, 1405–1413.

Fu, Y.X., 1994. Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. Genetics 138, 1375–1386.

Fu, Y.X., 1996. New statistical tests of neutrality for DNA samples from a population. Genetics 143, 557–570.

Fu, Y.X., Li, W.H., 1993. Statistical tests of neutrality of mutations. Genetics 133, 693–709.

Gelman, A., Carlin, J.S., Stern, H.S., Rubin, D.B., 1997. Bayesian Data Analysis. Chapman & Hall, Boca Raton, FL.

Hartl, D.L., Moriyama, E.N., Sawyer, S.A., 1994. Selection intensity for codon bias. Genetics 138, 227–234.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–109.

Hudson, R.R., Kaplan, N.L., 1988. The coalescent process in models with selection and recombination. Genetics 120, 831–840.

Kaplan, N.L., Darden, T., Hudson, R.R, 1988. The coalescent process in models with selection. Genetics 120, 819–829.

Kaplan, N.L., Hudson, R.R., Langley, C.H., 1989. The "hitchhiking effect" revisited. Genetics 123, 887–899.

Kaplan, N., Hudson, R.R., Iizuka, M, 1991. The coalescent process in models with selection, recombination and geographic subdivision. Genet. Res. 57, 83–91.

Krone, S.M., Neuhauser, C., 1997. Ancestral processes with selection. Theor. Popul. Biol. 51, 210–237.

Lewontin, R.C., 1974. The Genetic Basis of Evolutionary Change. Columbia University Press, New York, NY.

Lewontin, R.C., Krakauer, J., 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74, 175–195.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations for fast computing machines. J. Chem. Phys. 21, 1087–1092.

Moran, P.A.P., 1962. The statistical processes of evolutionary theory. Clarendon Press, Oxford.

Neuhauser, C., Krone, S.M., 1997. The genealogy of samples in models with selection. Genetics 145, 519–534.

Nielsen, R., 2001. Statistical tests of selective neutrality in the age of genomics. Heredity 86, 641–647.

Sawyer, S.A., Hartl, D.L., 1992. Population genetics of polymorphism and divergence. Genetics 132, 1161–1176.

Sawyer, S.A., Dykhuizen, D.E., Hartl, D.L., 1987. Confidence interval for the number of selectively neutral amino acid polymorphisms. Proc. Natl. Acad. Sci. USA 84, 6225–6228.

Simonsen, K.L., Churchill, G.A., Aquadro, C.F., 1995. Properties of statistical tests of neutrality for DNA polymorphism data Genetics 141, 413–429.

Slade, P.F., 2000a. Most recent common ancestor probability distributions in gene genealogies under selection. Theor. Popul. Biol. 58, 291–305.

Slade, P.F., 2000b. Simulation of selected genealogies. Theor. Popul. Biol. 57, 35–49.

Slade, P.F., 2001. Simulation of "hitch-hiking" genealogies. J. Math. Biol. 41, 41–70.

Slatkin, M., 1994. An exact test for neutrality based on the Ewens sampling distribution. Genet. Res. 64, 71–74.

Slatkin, M., 1996. A correction to the exact test based on the Ewens sampling distribution. Genet. Res. 68, 259–260.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585–595.

Watterson, G.A., 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7, 256–276.

Wright, S., 1949. Adaptation and selection. In: Jepson, G.G.S.G.L., Mayr, E. (Eds.), Genetics, Paleontology and Evolution, Princeton, Princeton, NJ, 1949 pp.

Wright, S., 1969. Evolution and the Genetics of Populations, Vol. 2: The Theory of Gene Frequencies. University of Chicago Press, Chicago.