BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 15 2007, pages 1962–1968
doi:10.1093/bioinformatics/btm264

Genetics and population analysis

The evolutionary forest algorithm

Scotland C. Leman[1,*], Marcy K. Uyenoyama[2], Michael Lavine[1] and Yuguo Chen[3]

[1]Institute of Statistics and Decision Sciences and [2]Department of Biology, Duke University and [3]Department of Statistics, University of Illinois at Urbana-Champaign, USA

Received on December 3, 2006; revised on April 5, 2007; accepted on May 10, 2007

Advance Access publication May 22, 2007

Associate Editor: Keith Crandall

ABSTRACT

Motivation: Gene genealogies offer a powerful context for inferences about the evolutionary process based on presently segregating DNA variation. In many cases, it is the distribution of population parameters, marginalized over the effectively infinite-dimensional tree space, that is of interest. Our evolutionary forest (EF) algorithm uses Monte Carlo methods to generate posterior distributions of population parameters. A novel feature is the updating of parameter values based on a probability measure defined on an ensemble of histories (a forest of genealogies), rather than a single tree.

Results: The EF algorithm generates samples from the correct marginal distribution of population parameters. Applied to actual data from closely related fruit fly species, it rapidly converged to posterior distributions that closely approximated the exact posteriors generated through massive computational effort. Applied to simulated data, it generated credible intervals that covered the actual parameter values in accordance with the nominal probabilities.

Availability: A C++ implementation of this method is freely accessible at http://www.isds.duke.edu/~scl13

Contact: scotland@stat.duke.edu

1 INTRODUCTION

Among the pivotal conceptual advances in evolutionary genetics is the insight that the sampling distribution of genetic variation can best be addressed from a genealogical perspective (Ewens, 1972; Kingman, 1982; Watterson, 1975). Because the ancestry of the genetic sample in turn can be inferred only from the pattern of genetic variation in the sample, analyses of DNA data have increasingly relied on sampling-based computational approaches (reviewed by Holder and Lewis, 2003). In a phylogenetic context, for example, a birth–death process describing the origin and extinction of taxa (Rannala and Yang, 1996) may determine a Bayesian prior distribution for the gene genealogy of the sample. Alternatively, relationships among genes sampled from natural populations may reflect a structured coalescence process (reviewed by Marjoram and Tavaré, 2006).

Here, we present a new Monte Carlo procedure, called the evolutionary forest (EF) algorithm, implemented in a Bayesian framework for the estimation of posterior distributions of population parameters rather than genealogies. In the application presented here, the objective is to estimate the time since divergence between closely related groups and the effective sizes of the ancestral and descendent groups (Fig. 1; compare Hey and Nielsen, 2004; Leman *et al.*, 2005; Wilson *et al.*, 2003).

*To whom correspondence should be addressed.

Most methods for the estimation of divergence times between groups treat these quantities as parameters (e.g., Hey and Nielsen, 2004; Rannala and Yang, 2003; Takahata *et al.*, 1995; Wall, 2003). As a consequence, an update of divergence time entails an update of the genealogy as well (see Nielsen and Wakeley, 2001). Incorporating time as a parameter into the Griffiths and Tavaré (1995) framework necessitates determination of the distribution of the number of ancestral lineages remaining at a specified time (Nielsen, 1998; Takahata, 1989). In contrast, we characterize divergence time as an exponentially distributed random variable [$\tau \sim \exp(\lambda)$ in Fig. 1] and estimate its rate parameter (see Leman *et al.*, 2005). This formulation preserves the character of all estimated parameters as ratios of evolutionary rates. Furthermore, it permits the separate exploration of genealogical history and evolutionary rates, a key feature of the EF algorithm.

A novel aspect of the EF algorithm is that updates of the population parameters are made on the basis of their likelihood over multiple trees (a *forest*) rather than a single tree. We show that the EF algorithm generates samples from the posterior distribution of the parameters. We assess its accuracy and efficiency through comparison to exact likelihoods computed (rather than estimated) for a small biological data set, through a study of coverage probabilities using simulated data, and in parallel with the IM package of Hey and Nielsen (2004).

2 SYSTEM AND METHODS

2.1 Pattern of genetic variation

In a set of DNA sequences sampled from two groups, we observe a number of segregating sites: homologous positions at which more than one nucleotide base occurs within the sample. We here assume the infinite sites model of mutation (see Kimura, 1969), under which all mutational events in the genealogical history of the sample since its most recent common ancestor (MRCA) have occurred at distinct nucleotide sites. We further assume the absence of recombination throughout this history. For each segregating site, we designate as the ancestral state the base in an outgroup sequence and regard any other base as a derived state. With respect to the subsample derived from each of the extant groups, we assign each derived mutation to one of three categories: present in all (fixed), present in at least one but not all (segregating), and present in none (absent) of the sequences from the group (see Leman *et al.*, 2005; Ramos-Onsins *et al.*, 2004). We denote by $n=(n_1, n_2, n_3, n_4, n_5, n_6, n_7)$ the numbers of mutations observed in the seven joint categories, at most four of which can take positive values for a single gene genealogy (Leman *et al.*, 2005).

Table 1 shows the counts observed for a non-coding region (*DPS2002*) in samples obtained by Machado *et al.* (2002) from natural

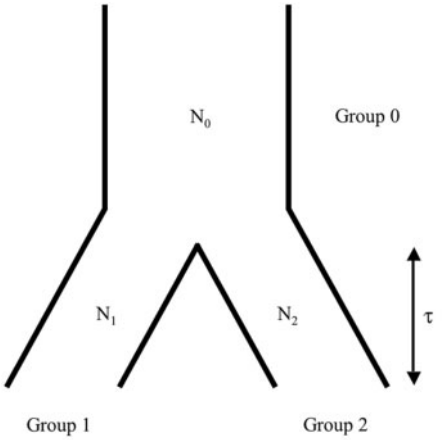

Fig. 1. Divergence of extant groups 1 and 2 from ancestral group 0. N_i denotes the effective number of genes in group i and τ the time since the division of the gene pool.

Table 1. Classification of segregating mutations and counts at the DPS2002 locus for sequences generated by Machado *et al.* (2002)

Count	Group 1	Group 2	*Dpe*[1]/*Dpp*[2]	*Dpe*/*Dpb*[3]
n_1	Segregating	Absent	16	16
n_2	Fixed	Absent	5	6
n_3	Absent	Segregating	65	18
n_4	Absent	Fixed	0	2
n_5	Segregating	Segregating	0	0
n_6	Fixed	Segregating	1	0
n_7	Segregating	Fixed	0	0

[1]*D.persimilis*, 13 sequences.
[2]*D.p.pseudoobscura*, 19 sequences.
[3]*D.p.bogotana*, 13 sequences.

populations of the fruit fly *Drosophila persimilis* and two subspecies of *D.pseudoobscura*, *D.p.pseudoobscura* and *D.p.bogotana* (see Leman *et al.*, 2005, for a description of the counting procedure). This vector of summary statistics constitutes the data (D) on which we base the estimation of population parameters of the evolutionary process.

2.2 Evolutionary model

We present a brief summary of the evolutionary model, described in full in Leman *et al.* (2005). Figure 1 depicts the divergence of extant groups 1 and 2 from ancestral group 0 τ time units in the past. An effective number of N_i genes constitutes the gene pool of group i ($i = 0, 1, 2$).

A genealogical history of the sample constitutes an ordered list of speciation, mutation and coalescence events (compare Griffiths and Tavaré, 1995). In the post-speciation phase, a lineage ancestral to the sampled genes exists in either group 1 or group 2, and in the post-speciation phase, it exists in group 0. We further classify a lineage according to whether its descendants in the sample occur exclusively in group 1 (type 1), exclusively in group 2 (type 2) or in both groups (type 3). To account for the seven kinds of mutations (Table 1), we need only specify the numbers of the three lineage types at any point in the genealogy:

$$\boldsymbol{S} = (l_{11}, l_{22}, l_{01}, l_{02}, l_{03}), \tag{1}$$

for l_{ij} the number of type j lineages that presently reside in group i. For example, a mutation that occurs only in genes sampled from group 1 must have arisen on a type 1 ancestral lineage. While the definition of the state space of the Griffiths and Tavaré (1995) approach under the infinite-sites model includes the mutational state of the lineages, our description (1) specifies only lineage type and number, with the mutations placed on the genealogical history in a separate step.

Transitions in state reflect speciation and coalescence between lineages residing in the same deme. Initiated in state (L_1, L_2, 0, 0, 0), for L_i the number of genes sampled from group i ($i = 1, 2$), the system evolves backward in time within this subspace under a pure death (coalescence) process until the speciation event. Speciation induces a jump to the subspace (0, 0, l_{01}, l_{02}, l_{03}), for $l_{0i} = l'_{ii}$, in which the prime denotes the most ancient post-speciation state. At the moment of speciation, $l_{03} \equiv 0$, reflecting that in the absence of between-group migration, only type 1 or type 2 lineages can exist. Generation of an ancestral lineage of type 3, by coalescence between lineages of different types (1/2, 1/3 or 2/3), can occur only in the pre-speciation phase. The process terminates in the MRCA (type 3) of the sampled genes.

We model the time since the speciation event as an exponentially distributed random variable, $\tau \sim \exp(\lambda)$, for λ the per-generation rate of speciation, and the time to a mutation in any lineage as an exponentially distributed random variable with parameter u, the per-gene, per-generation rate of mutation (see Leman *et al.*, 2005). We assume complete selective neutrality, which implies equal rates of coalescence among lineages residing in the same group, irrespective of mutational state. Proceeding backward in time, the waiting time to coalescence between any pair of the $l_i = \sum_j l_{ij}$ genetic lineages present in group i ($i = 0$, 1, 2) has an exponential distribution with parameter $\binom{l_i \cdot}{2}/N_i$ (Kingman, 1982). Independence among the processes of speciation, mutation and coalescence implies that the waiting time to the most recent event also has an exponential distribution, with parameter equal to the sum of the rates of the processes. In the post-speciation phase, for example, the most recent event corresponds to speciation with probability

$$\frac{\lambda}{\binom{l_1 \cdot}{2}/N_1 + \binom{l_2 \cdot}{2}/N_2 + l_{1\cdot}u + l_{2\cdot}u + \lambda} = \frac{\Lambda}{l_{1\cdot}(l_{1\cdot} - 1)/\theta_1 + l_{2\cdot}(l_{2\cdot} - 1)/\theta_2 + l_{1\cdot} + l_{2\cdot} + \Lambda} \tag{2}$$

for

$$\Lambda = \lim \lambda/u$$
$$\theta_i = \lim 2N_i u,$$

in which the limits denote infinitesimal rates of change (λ, u, $1/N_i \to 0$). These considerations define a Markov chain, initiating with the observed sample and terminating in the MRCA, with transition rates corresponding to ratios of the evolutionary rate parameters.

Adopting a Bayesian approach, we seek to characterize the posterior distribution of the population parameters $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \Lambda)$:

$$p(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}),$$

for D the observed mutational counts (Table 1). The likelihood corresponds to a marginal over genealogical histories (t):

$$p(D|\boldsymbol{\theta}) = \sum_{t \in \Omega_T} p(D, t|\boldsymbol{\theta}), \tag{3}$$

for Ω_T the space of T, the random variable representing the genealogical history of the sample. For a specified history, $p(D, t|\boldsymbol{\theta})$ is determined from the Markovian transition probabilities (2).

The number of possible histories ($|\Omega_T|$) is finite in the absence of migration (and countable in the presence of migration). Uyenoyama and Takebayashi (2004) derived a recursion in probability generating functions from which the likelihood (3) can be determined by explicit enumeration of all possible histories. For all but very small data sets, the vastness of Ω_T makes this approach infeasible.

2.3 The EF approach

Here we show that the EF algorithm directly addresses the marginal in (3) by sampling from the joint posterior distribution $p(t, \boldsymbol{\theta}|D)$.

Like a number of other genealogy-based methods, the EF algorithm incorporates a Gibbs sampler (Gelfand and Smith, 1990; Geman and Geman, 1984) with Metropolis–Hastings (Hastings, 1970; Metropolis *et al*., 1953) steps. In particular, sufficient alternate sampling from $p(\boldsymbol{\theta}|t, D)$ and $p(t|\boldsymbol{\theta}, D)$ yields samples from the joint distribution $p(t, \boldsymbol{\theta}|D)$. A feature that distinguishes the EF approach from comparable genealogy-based methods is the assessment of parameter values with respect to a collection (*forest*) of trees rather than a single tree. Consideration of multiple trees, even though the actual sample had a single genealogical history, represents a form of data augmentation (Tanner and Wong, 1987), a standard approach to missing data problems.

We define the space of forests by

$$\Omega_F = \left\{ \biguplus_{i=1}^{K} t_i : \ t_i \in \Omega_T \right\},$$

for K a fixed number of trees. Operator $\biguplus$ denotes union to form a multiset, an unordered collection which preserves multiplicity: for example, $t_1 \biguplus t_2 \biguplus t_1 \biguplus t_3 \biguplus t_4 = \{t_1, t_1, t_2, t_3, t_4\}$. For each forest $f \in \Omega_F$ and each tree $t \in f$, $M(t, f)$ denotes the multiplicity of t in f. The joint posterior distribution of the forest (and parameter set) is defined to be proportional to the sum of the posterior probabilities on the trees (and parameters) in the forest. We explicitly assign the probability of the joint posterior distribution of f and $\boldsymbol{\theta}$ as

$$q(f, \boldsymbol{\theta}|D) = C \sum_{t \in f} p(t, \boldsymbol{\theta}|D), \tag{4}$$

for C a normalizing constant:

$$1/C = \binom{n+K-1}{K} K/n, \tag{5}$$

for n the total number of genealogical histories ($|\Omega_T|$). This expression reflects that each tree appears in a proportion K/n of the $\binom{n+K-1}{K}$ multisets that exist (the number of ways of placing K indistinguishable balls into n distinguishable urns; Feller, 1950). We note for a tree having multiplicity $M(t,f)$, its contribution to the sum in Equation (4) is $M(t, f)p(t, \boldsymbol{\theta}|D)$, so $q(f, \boldsymbol{\theta}|D)$ is the sum over *every* tree in the forest.

A corollary of (4) gives the essence of the EF approach: the marginal posterior distribution of $\boldsymbol{\theta}$ over forests corresponds to that over samples of single trees. Explicitly,

$$\begin{aligned} \sum_{f \in \Omega_F} q(f, \boldsymbol{\theta}|D) &= \sum_{f \in \Omega_F} \sum_{t \in f} Cp(t, \boldsymbol{\theta}|D) \\ &= p(\boldsymbol{\theta}|D) \sum_{f \in \Omega_F} \sum_{t \in f} Cp(t|\boldsymbol{\theta}, D) \\ &= p(\boldsymbol{\theta}|D), \end{aligned} \tag{6}$$

reflecting that for $R = 1/C$ denoting the multiplicative size increase of the forest space over the tree space, $\sum_{f \in \Omega_F} \sum_{t \in f} Cp(t|\boldsymbol{\theta}, D) = R \sum_{t \in \Omega_T} Cp(t|\boldsymbol{\theta}, D) = RC = 1$. Consequently, updating parameter values with respect to forests sampled from $q(f, \boldsymbol{\theta}|D)$ generates draws of θ from the correct distribution.

3 ALGORITHM

Our algorithm initializes the sampling with a set of parameter values,

$$\boldsymbol{\theta}^{(0)} = (\theta_0^{(0)}, \theta_1^{(0)}, \theta_2^{(0)}, \Lambda^{(0)}),$$

and a forest of trees,

$$f^{(0)} = \{t_1^{(0)} \ldots t_K^{(0)}\},$$

generating each tree independently under $\boldsymbol{\theta}^{(0)}$, and samples alternately between $q(\boldsymbol{\theta}|f, D)$ and $q(f|\boldsymbol{\theta}, D)$. We use a random Gibbs sampler (Robert and Casella, 1999, Chapter 7), proposing a move in parameter space according to $\nu \sim \text{Bernoulli}(p)$ ($p = 1/2$ in our present implementation) and in forest space otherwise. Acceptance of parameter moves, proposed as a block under a Gaussian random walk, depends on the likelihood over the current forest. Proposing a move in forest space entails replacement of a single tree in the current forest with a tree independently generated under the current set of parameters (see Leman *et al*., 2005). We summarize the ith ($i = 1, 2, \ldots$) cycle of the algorithm:

Sample $\nu \sim \text{Bernoulli}(p)$.

- $\nu = 1$ (parameter space move) Propose parameter updates $\boldsymbol{\theta}^{(c)} \sim N(\boldsymbol{\theta}^{(i-1)}, \sigma^2 I)$. Set $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(c)}$ with probability
$$\min\left(1, \frac{q(\boldsymbol{\theta}^{(c)}|D, f^{(i-1)})}{q(\boldsymbol{\theta}^{(i-1)}|D, f^{(i-1)})}\right); \tag{7}$$
otherwise, $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^{(i-1)}$.
- $\nu = 0$ (forest space move)
Generate a new tree $t^{(c)}$ from $g(t|\boldsymbol{\theta}^{(i)}, D)$; sample t_j uniformly from the trees in $f^{(i-1)}$ and assign $f^{(c)} = t^{(c)} \biguplus f^{(i-1)} \setminus t_j$. Accept the proposed forest ($f^{(i)} = f^{(c)}$) with probability
$$\min\left(1, \frac{q(f^{(c)}|D, \boldsymbol{\theta}^{(i)}) M(t^{(c)}, f^{(c)}) g(t_j|\boldsymbol{\theta}^{(i)}, D)}{q(f^{(i-1)}|D, \boldsymbol{\theta}^{(i)}) M(t_j, f^{(i-1)}) g(t^{(c)}|\boldsymbol{\theta}^{(i)}, D)}\right); \tag{8}$$
otherwise, $f^{(i)} = f^{(i-1)}$.

Iteration of this algorithm asymptotically results in draws from $q(f, \boldsymbol{\theta}|D)$.

In the parameter updates, the acceptance ratio (7) corresponds to

$$\frac{q(\boldsymbol{\theta}^{(c)}|D, f^{(i-1)})}{q(\boldsymbol{\theta}^{(i-1)}|D, f^{(i-1)})} = \frac{q(D, f^{(i-1)}|\boldsymbol{\theta}^{(c)}) p(\boldsymbol{\theta}^{(c)})}{q(D, f^{(i-1)}|\boldsymbol{\theta}^{(i-1)}) p(\boldsymbol{\theta}^{(i-1)})},$$

in which all distributions on the right are known. In particular, the definition of the forest distribution (4) implies

$$q(D, f|\boldsymbol{\theta}) = C \sum_{t \in f} p(D, t|\boldsymbol{\theta}).$$

The acceptance probability for updates to the forest (8) reflects that multiple trees in a forest can specify identical genealogical histories. For example, the probability of proposing $f^{(c)}$ from $f^{(i-1)}$ is

$$\frac{M(t_j, f^{(i-1)})}{K} g(t^{(c)}|\boldsymbol{\theta}^{(i)}, D),$$

for $g(t|\boldsymbol{\theta}, D)$ the proposal distribution for trees (see Leman *et al*., 2005), which guarantees the consistency of the tree. That is, we will have $p(D, t|\boldsymbol{\theta}) > 0$ for all t generated under $g(\cdot)$. The conditional forest distribution,

$$q(f|\boldsymbol{\theta}, D) = \frac{q(D|f, \boldsymbol{\theta}) q(f|\boldsymbol{\theta})}{q(D|\boldsymbol{\theta})},$$

is known up to $q(D|\boldsymbol{\theta})$, which cancels out from (8) because it depends on neither the current nor the proposed forest.

In the current implementation of EF, the user decides on the basis of trace plots whether the process has converged. Dependence among the parameters ($\boldsymbol{\theta}$) necessitates convergence to the joint stationary distribution of all parameters.

After that point, every Δth draw from the Markov chain is stored, for Δ determined by the user. In the application described here, we assigned $\Delta = 1/p$, for $p = 1/2$ the probability of proposing updates of the population parameters, so that the interval between draws corresponds to the expected number of iterations before an attempted parameter update.

4 IMPLEMENTATION

4.1 Application to simulated data

To assess convergence to the target distribution, we generated simulated data under specified parameters and determined the coverage probabilities of credible intervals estimated using the EF algorithm. For each data set, we generated population parameters uniformly for $\Lambda \in (0, 5]$ and $\boldsymbol{\theta}_i \in (0, 30]$ ($i = 0, 1, 2$). We determined counts for the seven joint summary statistics (Table 1) by simulating the coalescence process described in Section 2.2 under these parameters, starting with five samples from each of the extant groups. This process was repeated until we obtained 100 independent data sets, each generated under an independent set of population parameters $\boldsymbol{\theta}$ and each containing at least 15 segregating mutations.

For each data set, we generated draws from the joint posterior distribution (4) using the EF algorithm with $K = 50$ trees per forest and prior distributions $\Lambda \sim \text{Unif}(0,5]$ and $\boldsymbol{\theta}_i \sim \text{Unif}(0,30]$ ($i = 0, 1, 2$). Convergence of the Markov chain (burn-in) was ascertained by visual inspection of the trace plots. In all cases, burn-in required fewer than 5000 MCMC iterations. After burn-in, 100 000 samples were collected for posterior inference. Table 2 shows that the proportion of data sets for which the estimated 95% credible intervals contained the actual parameter values accorded well with the nominal probability. The absence of significant departures from expectation under a binomial distribution suggests convergence of the EF algorithm to the correct distribution.

4.2 Comparison to exact likelihoods based on actual data

For counts listed in the *Dpe*|*Dpb* column of Table 1, we used the method of Uyenoyama and Takebayashi (2004) to compute the exact likelihood curve for the speciation rate parameter Λ under assignment of the remaining parameters to the maximum-likelihood estimates (MLEs) reported by Leman *et al*. (2005). In Figure 2,the solid curve corresponds to this true curve, scaled by the diffuse prior distribution [$\Lambda \sim N(0.1, 50^2)$, truncated at zero] used in the EF analysis. The histograms summarize 120,000 draws, after a burn-in period of 100 000 iterations, from the posterior distribution estimated by the EF algorithm with forests of different sizes ($K = 1$, 10, 25) and using a random walk proposal for updating Λ with SD set to 1.

Similarly, for counts listed in the *Dpe*|*Dpp* column of Table 1, Figure 3 compares the true posterior distribution for θ_2, under prior distribution $N(10, 50^2)$ which has been truncated at zero, to 120 000 draws from the posterior distribution estimated by the EF algorithm using the same prior distribution with forest sizes of 1, 5, 50 and 100 trees and a random walk proposal for updating θ_2 with SD set to 5.

Both Figures 2 and 3 indicate increasing agreement between the exact and estimated posterior distribution under larger forests. Smaller forest sizes appear to require longer burn-in periods for both the *Dpe*|*Dpb* and *Dpe*|*Dpp* data sets. For example, convergence of parameter θ_2 for the *Dpe*|*Dpp* data set required about 50 million iterations for forests comprising a single tree ($K = 1$) and only 120 000 iterations for $K = 100$.

For θ_2, the parameter with the broadest posterior distribution, computation of the exact posterior (solid curve in Fig. n2post) took over four months using the recursive method of Uyenoyama and Takebayashi (2004). Using forests of 50 trees, the EF results required ~3 h (all computations were performed on a Macintosh Dual 2.5 GHz PowerPC G5). Computing exact posterior distributions for all four parameters would require on the order of $4^4/12 = 21.33$ years. In contrast, computation time for the full joint distribution of multiple parameters under the EF algorithm shows only trivial increases with the number of parameters. Rather, computation time increases linearly with the number of trees in the forest, reflecting both increased dimension of forest space and smaller perturbations of the current forest induced by the substitution of single trees.

Table 2. Coverage probabilities of 95% credible intervals

Population parameter	Coverage probability
Λ	0.93
θ_0	0.96
θ_1	0.96
θ_2	0.96

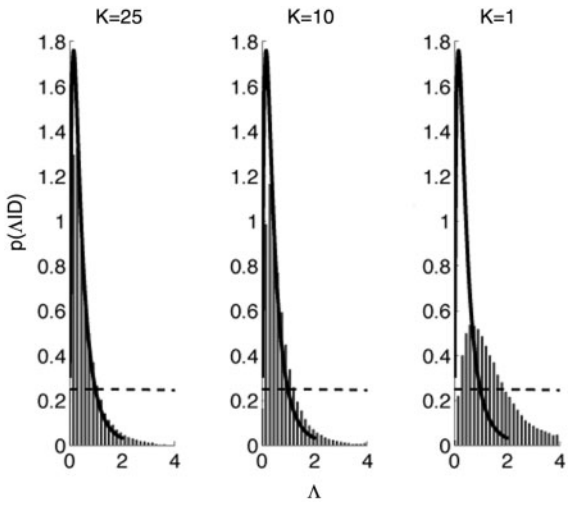

Fig. 2. Comparison between exact posterior distributions (curves) and estimated posterior distributions for Λ (histograms) using the EF algorithm with forest sizes of 1, 10 and 25 trees, with the remaining parameters assigned to their estimated maximum-likelihood values ($\theta_0 = 2.31$, $\theta_1 = 2.91$, $\theta_2 = 3.51$; Leman *et al*., 2005). Dashed curves depict the prior distributions.

4.3 Acceptance probability

A key innovation of the EF algorithm is the updating of population parameters in accordance with their effects on the average likelihood over multiple trees rather than a single tree. Because a single tree independently sampled from the vast tree space is unlikely to improve the likelihood, all MCMC-based

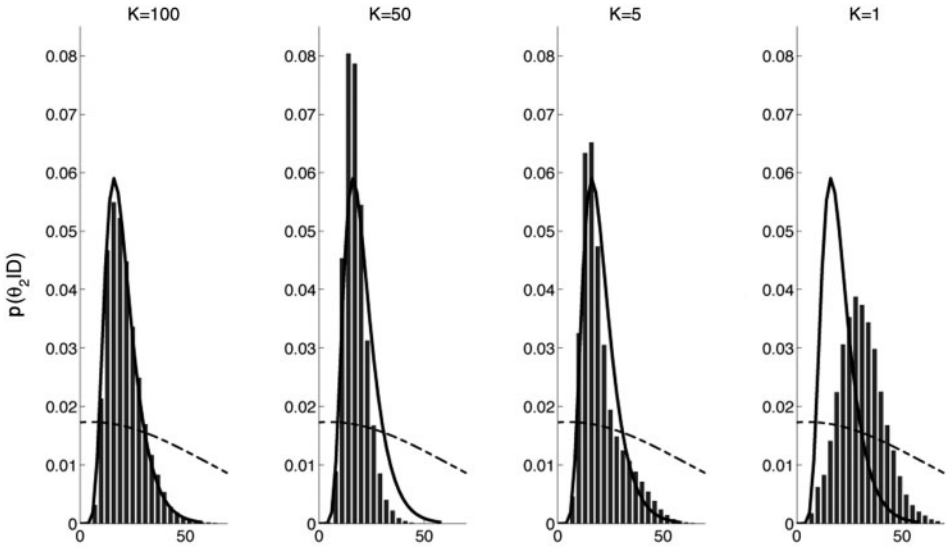

Fig. 3. Comparison between exact posterior distributions (solid curves) and estimated posterior distributions for θ_2 (histograms) with forest sizes of 1, 5, 50 and 100 trees, with the remaining parameters assigned to their maximum-likelihood estimates ($\Lambda = 0.12$, $\theta_0 = 0.81$, $\theta_1 = 2.71$; Leman *et al.*, 2005). Dashed curves depict the prior distributions.

methods currently in wide use restrict proposals in tree space to local moves: by breaking and repositioning a single branch, for example. Increasing forest size can accelerate the rate of exploration of genealogies under the EF algorithm.

Proposed updates to the forest entail substitution of a single tree by an independently generated tree. Because increasing forest size reduces the effect of this substitution, the rate of acceptance increases as well. For arbitrarily large forests,

$$\lim_{K\to\infty} \frac{q(D, f^{(c)}|\boldsymbol{\theta}^{(i)})}{q(D, f^{(i-1)}|\boldsymbol{\theta}^{(i)})} = 1. \quad (9)$$

In the limit, the ratio of the proposal distributions $M(t, f)g(t|\theta, D)$ comes to dominate the Metropolis–Hastings ratio (8).

To illustrate this effect of forest size, we monitored the acceptance probability of moves in forest space using the *D. persimilis* and *D.p.pseudoobscura* samples (*Dpe*/*Dpp* column in Table 1). We assigned the population parameters to their maximum-likelihood values (see caption to Fig. 4) and limited MCMC moves to forest space ($p = 0$ in the algorithm described in Section 3). Figure 4 shows an increase in the acceptance rate with forest size, suggesting that larger forests promote more rapid exploration of forest space. Because it is often this step that limits the rate of convergence of MCMC analyses, larger forests may promote more rapid convergence of the full joint posterior distribution, over both forests and parameters.

4.4 Comparisons to IM

EF resembles IM (Hey and Nielsen, 2004) in its Bayesian approach and model of speciation (Fig. 1). IM differs from EF in its use of entire nucleotide sequences rather than summary statistics; estimation of divergence time T (τ in Fig. 1) as a parameter rather than as the expectation of a random variable with parameter λ; and incorporation of parallel tempering (Geyer, 1991) of multiple chains, each specifying a single tree, rather than a single chain specifying a forest of trees. Applied to the two *Drosophila* data sets (Table 1), these methods showed marked differences in convergence properties, but once converged, gave similar parameter estimates.

In our application of IM, we specified three parallel tempering chains and storage of every 100 th draw (10-fold higher than the default). For the uniform prior distributions assumed by IM, we specified $uT \in (0, 100]$ and $\theta_i \in (0, 300]$ ($i = 0, 1, 2$). We accepted the default values for the remaining options. Our EF results derived from forests comprising $K = 50$ trees and truncated diffuse Gaussian prior distributions for $\Lambda \sim N(.1, 30^2)$ and $\theta_i \sim N(10, 100^2)$ ($i = 0, 1, 2$). We compared the Λ estimate from EF to the inverse of the uT estimate from IM, reflecting that $E[\tau] = 1/\lambda$ under the model incorporated into EF (Fig. 1).

Applied to the larger data set (*Dpe*/*Dpp* in Table 1), IM failed to converge within three weeks, apparently reflecting low acceptance rates of proposed updates to the gene tree and divergence time (τ in EF or T in IM). For example, updating of the gene tree ranged from less than 1 in 10 000 to 1 in 9, suggesting slow exploration of tree space. Because the various elements of the model are highly interdependent, non-convergence of some parameters reduces confidence in apparently stable parameter estimates. A marginal assessment of estimates previously reported by Hey and Nielsen (2004) and rescaled to our units ($\Lambda = 0.15$, $\theta_0 = 1.1$, $\theta_1 = 2.4$, $\theta_2 = 21.6$) indicated non-significant departures from our estimates (Table 3), generated by EF in ~5h.

Applied to the smaller data set (*Dpe*/*Dpb* in Table 1), IM generated trace plots that suggested convergence within four days, but we continued the run because subsets of samples from the posterior distributions appeared to differ markedly. Differences among subsets persisted after seven additional days, but we nevertheless accepted the estimates generated after that point. EF required ~3h for convergence, under random walk proposal SDs of (1, 3, 3, 3) for updating ($\Lambda, \theta_0, \theta_1, \theta_2$), respectively. Table 4 indicates close correspondence between both the posterior modes and credible intervals generated by the two methods.

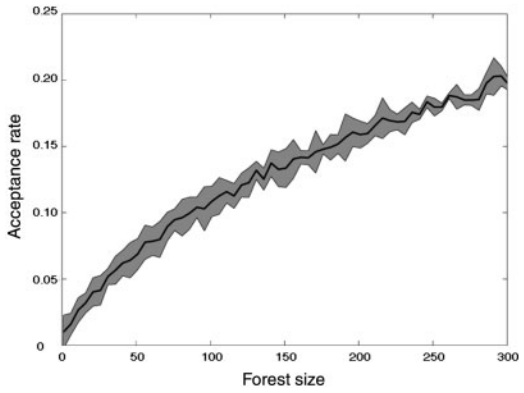

Fig. 4. Acceptance rate increases with forest size. Black line: mean acceptance rate from six MCMC runs, each comprising 10 000 samples for each point. Gray region: pointwise 95% confidence interval around the mean for the six runs. Population parameters were fixed at the maximum-likelihood values indicated in Figure 3, with the addition of $\theta_2 = 18.21$.

IM appeared to require substantial thinning, with parameter updates appearing to show substantial autocorrelations in spite of retention of draws at intervals 10-fold longer than the default. Because parameter likelihoods are conditional on the current tree, infrequent updating of the tree exacerbates autocorrelations.

4.5 Comparison to a previously studied importance sampling procedure

Leman *et al*. (2005) analyzed the data set in Table 1, using importance sampling (IS) approximations to obtain MLEs of the population parameters (θ). While the EF and IS algorithms generated comparable estimates, the EF algorithm requires substantially less manual tuning. In particular, the IS algorithm estimates the likelihood function under a set of driving parameter values (see Felsenstein *et al*., 1999), the optimal choice of which corresponds to the modal values. To address this problem, the IS algorithm implements a two-phase procedure to search for the mode: the first phase uses random searching to generate initial values for the gradient-descent search of the second phase. After this tuning step, which may require several days, the IS method estimates likelihood values for θ on a grid and uses interpolating splines to construct a smooth likelihood surface.

In contrast, the EF method requires little manipulation by the user. Because the EF algorithm constantly updates the parameter values θ, it shows little dependence on the initial values. Sampling from the posterior distribution $p(\boldsymbol{\theta}|D)$ obviates the need for interpolation. Further, the EF algorithm makes much less demands on memory (RAM) because it stores only the relatively small number of trees which constitute the forest, rather than the great many trees required for fast and reliable estimation of the likelihood function.

5 DISCUSSION

5.1 Relationship to parallel tempering

Running multiple chains simultaneously can increase performance of MCMC algorithms (Geyer, 1991). Parallel tempering methods use an energy function transformation of the probability distribution to flatten the posterior distribution. This device improves mixing by increasing movement between regions of high posterior density (or mass). While hot chains promote sampling efficiency, sampling from the target distribution must be done on cool chains. In particular, Altekar *et al*. (2004) exploit this technique in their Metropolis coupled Markov chain Monte Carlo (MC^3) algorithm for the estimation of gene trees, and Hey and Nielsen's (2004) IM program for the estimation of population parameters incorporates it as well. Ascertainment of convergence can be difficult in this setting, often requiring considerable *a priori* knowledge of the target distribution.

The EF algorithm shares with parallel tempering the strategy of flattening tree space: the effect of the replacement of any single tree diminishes as forest size increases (9). A key advantage of the EF algorithm over tempering is the preservation of the marginal distribution of the population parameters (6). This property obviates the additional complexity of running multiple chains, each examining a single tree at a time. In the EF algorithm, computational complexity increases only linearly with forest size (K).

Table 3. Parameter estimates and credible intervals for the *D. persimilis* and *D.p.pseudoobscura* data set (*Dpe*|*Dpp*) under the EF method

Parameter	Posterior mode	95% credible interval
Λ	0.21	(0.05, 1.4)
θ_0	5.4	(0.9, 27.9)
θ_1	4.2	(2.3, 11.1)
θ_2	15.4	(9.4, 54.4)

Posterior inferences based on 1 000 000 samples.

Table 4. Parameter estimates and credible intervals for the *D.persimilis* and *D.p.bogotana* data set (*Dpe*|*Dpb*) under the EF and IM methods

	Posterior mode		95% credible interval	
Parameter	EF	IM	EF	IM
Λ	0.19	0.14	(0.05, 1.2)	
θ_0	2.5	2.4	(0.5, 51.3)	(0.39, 44.7)
θ_1	3.4	2.8	(1.7, 10.0)	(2.0, 11.0)
θ_2	4.0	5.6	(2.0, 11.0)	(3.0, 21.5)

Posterior inferences based on 1 000 000 samples.

5.2 Marginalization and efficiency

In genealogy-based approaches for the estimation of population parameters, the ancestry of the sample provides the context for the characterization of properties of the sampling distribution. Some methods (e.g. Kuhner *et al*., 1995; Wilson and Balding, 1998) entail full reconstruction of gene trees, including topology and all node ages, prior to their marginalization. Alternatively, a genealogical history under the Griffiths and Tavaré (1995) approach specifies only the relative order of evolutionary events, a definition that in itself marginalizes

entire equivalence classes of fully resolved trees. Our EF algorithm replaces marginalization over tree space with marginalization over the augmented forest space. In spite of the higher dimension of forest space, our experience with the application of the EF algorithm to the actual and simulated data sets described here suggests more rapid exploration of forest space ($K > 1$) than tree space ($K = 1$), perhaps reflecting the greater connectedness of forests that differ by few trees.

Reduction of the data to summary statistics represents another form of marginalization, now over data sets that exhibit the values of the statistics observed in the data set at hand. This reduction may entail little loss of information about the parameters of interest for summary statistics that are in some sense 'close' to sufficient for those parameters. Here, we have based the estimation on counts of mutations under the infinite-sites model in the absence of recombination (compare Ramos-Onsins *et al*., 2004; Wakeley and Hey, 1997). Unlike most methods of this kind, ours explicitly accounts for the dependence among the summary statistics that is induced by the occurrence of the mutations on the same genealogical tree (Leman *et al*., 2005). While we have not attempted to address the issue of sufficiency analytically, our empirical experience suggests that EF exhibits better convergence properties and generates comparable estimates for divergence time and effective sizes from summary statistics in much less time than the full-sequence IM method (Hey and Nielsen, 2004).

Beyond the features shared with our present implementation, IM can accommodate multiple loci and other models of mutation. Extending our method to multiple independent loci by assigning the joint likelihood to the product of likelihoods across loci, as in IM, would appear relatively straightforward. Accommodating more general models of mutation (a general time-reversible model allowing multiple hits at a site, for example) might entail specification of branch lengths, as in IM (which would complicate the marginalization of genealogy), or incorporation of mutational state into the genealogical history (1), as in the Griffiths and Tavaré (1995) approach.

ACKNOWLEDGEMENTS

We thank the reviewers for their constructive comments. Funding from the National Institutes of Health (GM 37841, M.K.U.) and the National Science Foundation (DMS-0503981, Y.C.) provided partial support for this study.

Conflict of Interest: none declared.

REFERENCES

Altekar,G. *et al*. (2004) Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, **20**, 407–415.

Ewens,W.J. (1972) The sampling theory of selectively neutral alleles. *Theor. Pop. Biol*., **3**, 87–112.

Feller,W. (1950) *An Introduction To Probability Theory And Its Applications*. Vol. I, John Wiley & Sons, New York.

Felsenstein,J. *et al*. (1999) Likelihoods on coalescents: A Monte Carlo sampling approach to inferring parameters from population samples of molecular data. In Seillier-Moiseiwitsch,F. (ed.) *Statistics in Molecular Biology and Genetics*. Institute of Mathematical Statistics and American Mathematics Society Haywood, CA, pp. 163–185.

Gelfand,A.E. and Smith,A.F.M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc*., **85**, 398–409.

Geman,S. and Geman,D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell*., **6**, 721–741.

Geyer,C.J. (1991) Markov chain Monte Carlo maximum likelihood. In Keramidas,EM. (ed.) *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*. Interface Foundation of North America Fairfax Station, VA, pp. 156–163.

Griffiths,R.C. and Tavaré,S. (1995) Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Math. Biosci*., **127**, 77–98.

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

Hey,J. and Nielsen,R. (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis. *Genetics*, **167**, 747–760.

Holder,M. and Lewis,P.O. (2003) Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev. Genet*., **4**, 275–284.

Kimura,M. (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, **61**, 893–903.

Kingman,J.F.C. (1982) On the genealogy of large populations. *J.Appl. Prob*., **19**, 27–43.

Kuhner,M.K. *et al*. (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, **140**, 1421–1430.

Leman,S.C. *et al*. (2005) Likelihoods from summary statistics: recent divergence between species. *Genetics*, **171**, 1419–1436.

Machado,C.A. *et al*. (2002) Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and close relatives. *Mol. Biol. Evol*., **19**, 472–488.

Marjoram,P. and Tavaré,S. (2006) Modern computational approaches for analysing molecular genetic variation data. *Nat. Rev. Genet*., **7**, 759–770.

Metropolis,N. *et al*. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys*., **21**, 1087–1092.

Nielsen,R. (1998) Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. *Theor. Pop. Biol*., **53**, 143–151.

Nielsen,R. and Wakeley,J. (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.

Ramos-Onsins,S.E. *et al*. (2004) Multilocus analysis of variation and speciation in the closely related species *Arabidopsis halleri* and *A. lyrata*. *Genetics*, **166**, 373–388.

Rannala,B. and Yang,Z. (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol*., **43**, 304–311.

Rannala,B. and Yang,Z. (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, **164**, 1645–1656.

Robert,C.P. and Casella,G. (1999) *Monte Carlo Statistical Methods*. Springer-Verlag, New York.

Takahata,N. (1989) Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics*, **122**, 957–966.

Takahata,N. *et al*. (1995) Divergence time and population size in the lineage leading to modern humans. *Theor. Pop. Biol*., **48**, 198–221.

Tanner,M.A. and Wong,W.H. (1987) Calculation of posterior distributions by data augmentation (with discussion). *J. Am. Stat. Assoc*., **82**, 528–550.

Uyenoyama,M.K. and Takebayashi,N. (2004) A simple method for computing exact probabilities of mutation numbers. *Theor. Pop. Biol*., **65**, 271–284.

Wakeley,J. and Hey,J. (1997) Estimating ancestral population parameters. *Genetics*, **145**, 847–855.

Wall,J.D. (2003) Estimating ancestral population sizes and divergence times. *Genetics*, **163**, 395–404.

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

Wilson,I.J. and Balding,D.J. (1998) Genealogical inference from microsatellite data. *Genetics*, **150**, 499–510.

Wilson,I.J. *et al*. (2003) Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *J. R. Stat. Soc. A*, **166**, 155–201.