Genealogies from time-stamped sequence data

Alexei Drummond, Geoff K. Nicholls, Allen G. Rodrigo, Wiremu Solomon

Auckland University, Private Bag 92019,
Auckland, New Zealand
nicholls@math.auckland.ac.nz

4/4/2002, Gregynog, UK, revised 10/1/2003, Auckland, NZ

We review and develop Bayesian statistical methods for recovering genealogical structure, population size and mutation rates from radiocarbon-dated fossil mtDNA sequence data. It is possible to obtain ages for fossil DNA sequences and their common ancestors, by fitting a population-genetic model. We describe the observation model and show how uncertainty in reconstructed parameter values may be quantified via sample-based inference. We give an example, in which errors arising from radiocarbon calibration of fossil sequences are dominated by uncertainty in the genealogy and associated population parameters. We do not discuss likely model mispecification in any detail.

1 Introduction

The remarkable thing about using genetic material to date events, is that it is possible at all. There are two clocks in the story, one ticks at the mutation rate of DNA, the other at the coalescent rate of ancestral lineages. We know the rate constant for neither clock. We seem to be pulling ourselves up by our bootstraps, obtaining rates and dates from the one data set. It turns out that in fact we cannot date without some prior knowledge of the likely value of the date we wish to estimate. However, only very limited information is needed, as we show. In this paper we review recently published statistical methodology for genetic dating. Our explicit treatment of the uncertainty arising from imprecisely dated sequences, Section 6, is new, as is our characterization of the limitations of genetic dating, presented in a couple of highlighted paragraphs in Section 4.

What are we dating here? We can estimate an age for the common ancestor of the individuals whose DNA makes up the data. The amount of divergence between the DNA sequences of the sample individuals is a measure of that age. We can also estimate the time at which an individual lived, if we have an undated DNA sequence from the individual, along with dated sequences from other individuals in the same population. In Drummond *et al.* 2002, the authors date modern-day events, using HIV sequence data, with ages known to the day. Where radiocarbon dating is used to date fossil DNA, as in Lambert *et al.* 2002, analysis to date has been conditioned on point estimates of calibrated dates. This seems to be an unimportant approximation as we show below.

What archaeological questions can we answer? We can provide estimates of dated genealogies, and in some simple cases, an estimate of the way the total population size has varied as

a function of time. It is possible that direct genetic dating of fossil DNA sequences (see Section 8 and Figure 6) may in the future be used to make very crude temporal classifications, along the lines of "ancient or modern?". However, care must be taken when gene-derived dates are used to infer cultural history. Where this link is needed, an explicit case must be made. The argument is typically based on archaeological context, but may follow from statistical considerations, as in Penny *et al.* 1993, who discuss the parallel genealogies of human languages and human genes. The genealogies of domestic animals and human and animal diseases may be of independent interest, whilst recent contact between human communities may potentially be resolved by the viruses shared by those communities, since viruses evolve at a much higher rate than their hosts. In Matisoo-Smith *et al.* 1998, the genealogy of polynesian rat mtDNA reveals prehistoric patterns of human mobility in East Polynesia. Underhill *et al.* 2001 develop a model of human contact in the wider Pacific from Y-chromosome data for Pacific peoples. Before sequence data were available, ancestral inference, in archaeology as in biology, was based on higher-level trait-based data. Ammerman and Cavalli-Sforza 1984 consider the mesolithithic-neolithic transition in Europe in the light of blood type and other trait-based data. The citations in this paragraph reflect the authors' personal interests.

We focus on recent developments in statistical methodology. Data-analytical tools, developed in the last decade, allow us to put error bars on the dates and rates reconstructed from sequence data. This is not trivial as we don't usually know how the sample specimen are related. Kuhner *et al.* 1995 used MCMC to average over genealogies and obtain an ML-estimate for one of the two rate parameters of the problem, if the other is known. Model averaging of this kind is computationally demanding, but statistically robust. Kuhner *et al.* 1995 suppose all data sequences have equal age. Importance sampling methods are used to get estimates at parameter values which were not simulated. This proved to be a weakness of the method, as the importance weights can have high variance. The Bayesian analysis given below uses MCMC to average genealogies and rate parameters simultaneously, from sequence data gathered at different times. Whilst MCMC methods have their own weaknesses, and must be used with care, they have certainly extended the range of reliable population-genetic inference. As we explain below, the output of our simulations may be used to form ML-estimates, if so desired.

The methods described below were applied in Drummond *et al.* 2002, to estimate the parameters of an HIV population, and in Lambert *et al.* 2002, to estimate the age of the common ancestor of a collection of fossil Adelie penguin bones. In earlier work, Rambaut 2000 starts with a maximum likelihood phylogeny with time-stamped sequence data at the leaves and estimates the mutation rate. This kind of analysis, which dates back to Felsenstein 1981, is not robust, but is convenient for exploratory work. Barnes *et al.* 2002 estimates the genealogy of time-stamped fossil bear sequences via parsimony.

Bayesian inference is new to population genetics, where the Kingman coalescent provides an "inevitable" prior on genealogies. Wilson and Balding 1998 and Beaumont 1999 are early examples, treating micro-satellite data. Wilson *et al.* 2003 overlaps with Drummond *et al.* 2002. Phylogenetic tree priors are more obviously subjective. Perhaps for that reason, Bayesian methods are more common in that arena. See for example Suchard *et al.* 2001 and references therein.

In this paper we lay out the methodology in the context of a model of asexual reproduction. The cells of sexually reproducing creatures contain mitochondrial organelles, which behave a bit like cells within cells. Mitochondria carry their own DNA, and that mtDNA reproduces asexually, following the maternal line. It follows that the methodology we lay out below can be used to date events in the maternal ancestral tree of creatures which reproduce sexually.

2 The Mutation-Clock

The mutation model described in this section is the (standard) independent neutral finite-sites mutation model of Felsenstein 1981.

Consider an L-site DNA sequence. For $s = 1, 2 \ldots L$, and $\mathcal{C} = \{A, C, G, T\}$, let $B_s \in \mathcal{C}$ denote the character at site s. When the organism reproduces, the DNA sequence is copied, site by site. At each copy event, there is a small chance that a copy error may occur. We will assume these errors are independent but identically distributed from one site to another. Consider the probability that an X mutates into Y at one site in the course of a single generation. A generation is a very short time compared to the timescale at which we want to work, so we will write this probability in terms a dimensionless relative-rate matrix, $Q_{X,Y}$, a mutation rate parameter μ, with units *mutations per year*, and ρ, the number of years per generation. If $X \neq Y$, the probability for the event is $\mu Q_{X,Y} \rho$. Consider what happens over many generations. Let B denote an ancestral sequence and B' a descendant sequence, and suppose the two sequences are separated by an interval of time t much larger than ρ. If we can ignore terms of order ρ/t, the probability $\Pr\{B'_s = Y | B_s = X\}$ to get a Y at site s in B', given there was a X at site s in B is

$$\Pr\{B'_s = Y | B_s = X\} = [\exp(\mu Q t)]_{X,Y}$$

The exponential function of the 4×4 matrix Q is defined by the exponential series $\exp(M) = 1 + M + MM/2! + MMM/3! \ldots$, where 1 is the 4×4 identity, and MM is just matrix multiplication. Notice that when t gets small we can approximate $\Pr\{B'_s = Y | B_s = X\}$ by the matrix $1 + \mu Q t$. The off-diagonal elements are just what we started with, $\mu Q_{X,Y} t$. The row sums of $1 + \mu Q t$ should be one (it's a transition probability), so we need $Q_{X,X} = -\sum_{Y \neq X} Q_{X,Y}$ for the diagonal elements of Q.

In the following we will suppose, without further discussion, that all the entries in Q are known. In fact we can estimate those relative rates from the data along with all the other unknowns treated here. When the absolute rate μ is high, there are plenty of mutations, so the data pins down relative rates fairly tightly. Drummond *et al.* 2002 show that this works for real data.

Now, if we knew the ancestral sequence B, the final sequence B' and the mutation rate μ we could estimate the time t between the initial sequence and the final sequence. The likelihood for t would be

$$\Pr\{B' | B, \mu, t\} = \prod_{s=1}^{L} [\exp(\mu Q t)]_{B_s, B'_s} . \tag{1}$$

We do not have the ancestral sequence B, but we do know something about B, even before we see B'. The sequence B has itself evolved from a sequence of great antiquity. The proportions of A's, C's, G's and T's in B are determined by the relative rates, $Q_{A,C}$ *etc*, at which these bases mutate into one another. It follows that each character in B is a draw from the equilibrium of the mutation process. Let $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ denote this equilibrium distribution (so, $\pi Q = 0$). Since the mutation process acts independently at each site, the probability distribution for B itself is $\Pr\{B\} = \prod_s \pi_{B_s}$.

Of course, knowing B is a draw from the equilibrium of the mutation process doesn't help us pin down t. The likelihood obtained by summing out the unknown B gives us back the equilibrium distribution for B', independent of t. We need to know more about B and we can learn that by looking at its other descendants.

3 The Coalescent-Clock

The continuous time model described in this section, called the Kingman coalescent, is described in two classic papers, Kingman 1982a and Kingman 1982b. The process is extended to serial times in Rodrigo and Felsenstein 1999.

Consider a population of N_e individuals reproducing asexually. We will assume that the population size is constant in time (this assumption may be replaced by any other assumption which reasonably restricts the set of allowed population size histories, for example, to exponential growth at an unknown rate). Consider a pair of generations, and suppose the i'th individual in generation one produces n_i offspring in generation two. We model the evolution of a genealogy in the following way. We suppose each individual in generation two chooses its parent uniformly at random from the individuals in generation one, and independent of the choices made by its peers in generation two. This model, which is equivalent to imposing a multinomial distribution for the vector $(n_1, n_2, \ldots n_{N_e})$ of family sizes, is called the Wright-Fisher population model.

In order to simulate $K-1$ generations of an ideal Wright-Fisher population, take a piece of paper and mark a square lattice of dots, N_e dots across by K dots up. See Figure 1. Connect each dot by a directed edge pointing to a randomly chosen dot in the row above.

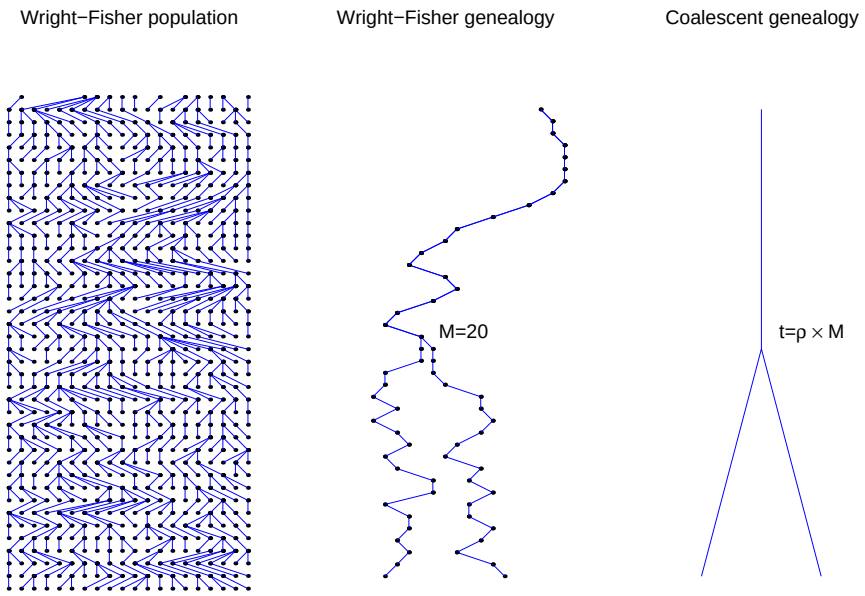

Figure 1: The Wright-Fisher and Kingman coalescent processes, with generation time ρ. (left) $K = 40$ generations of a Wright-Fisher population of $N_e = 20$ individuals, (center) the ancestral lineages of contemporary individuals 5 and 15 in the population at left coalesce $M = 20$ generations back, (right) the continuous time coalescent tree summarizing the discrete time ancestry in the centre.

The "present" generation is the row at the bottom of the page, and the earliest is the row at the top. A directed edge leads away from each dot in the array. A lineage is a directed sequence of edges.

Now pick two individuals from the population in the present without reference to their ancestry (*ie* pick two dots from the bottom row without looking at the edges connected to them). Consider the number, M say, of generations back to their first common ancestor. The Wright-Fisher model determines a geometric distribution for M, so that $\Pr\{M = m|N_e\} = p \times (1-p)^{m-1}$ with $p = 1/N_e$ (p is the probability two individuals have a common parent in the previous generation, so $p(1-p)^{m-1}$ is the probability the lineages don't coalesce for $m-1$ generations, and then do coalesce in the m'th). Notice that the mean number of generations back to the common ancestor of two individuals in the present is just N_e, the population size.

When N_e is large, the distribution of the time t back to coalescence is well approximated by an exponential. Recall ρ is the number of years per generation. Let λ be the rate, with units *coalescent events per year*, defined by $p = \lambda\rho$. Now m generations is $t = m\rho$ years, so when $0 < \rho \ll t$ and $N_e \gg 1$, we have $(1 - 1/N_e)^{t/\rho} \simeq \exp(-t/N_e\rho)$ and $t \sim \text{Exp}(\lambda)$. Readers familiar with population genetics should note that our λ (which equals $1/(N_e\rho)$) is equivalent to the expression "$1/\Theta$" found elsewhere in the literature.

Now consider what happens if we select not two, but n individuals. Also, we may select individuals from different generations. The ancestral lineages form a tree, with n leaf nodes and $n-1$ ancestral nodes, corresponding to $n-1$ coalescent events. At the root of the tree is the node corresponding to the most recent common ancestor of all n individuals at the leaf nodes. As we trace back from the present to the root, the lineages arising from the leaves coalesce one by one until there are just two lineages, which coalesce at the root. The common ancestral tree of n individuals is called their genealogy, and denoted g. It is a tree graph with distinguishable leaf vertices. The tree is drawn upside down, so that the altitude of a vertex is proportional to its age.

The coalescent process we have described determines a probability distribution for g, that is, a probability density for the joint distribution of the tree coalescent times and topology. If we follow any particular pair of lineages back in time, and ignore the rest, the pair we are following behave according to the rule we worked out for $n = 2$: they coalesce at instantaneous rate λ. With this one rule (and a bit of notation) we can write down the probability density for any particular tree g. As we trace back from the most recent leaf to the root, the number of lineages decreases by one at each coalescent event, and increases by one at each leaf. The number of lineages is a constant in each interval of time between consecutive vertices of the tree.

Imagine simulating the tree from its leaves up to the root. Number the vertices of the tree in order by age from $i = 1$ up to $i = 2n-1$ at the root and assign the vertices ages t_i. Suppose k_i lineages are present in interval $[t_i, t_{i+1})$. The rate R_i for coalescent events in interval i is a constant, $R_i = k_i(k_i-1)\lambda/2$ (the number of distinct pairs multiplied by the rate for each pair). If interval i ends with a coalescence event then its length, $\tau_i = t_{i+1} - t_i$ say, is an exponential variate with mean $1/R_i$. The pair of lineages which coalesce at the top of the interval is chosen from $k_i(k_i-1)/2$ pairs, so the probability density for that coalescent event was $\lambda \exp(-R_i\tau_i)$. If interval i ends with a leaf then there was no coalescence event in that interval, and that happens with probability $\exp(-R_i\tau_i)$. These events are independent so the probability density $f_G(g|\lambda)$ for the whole tree is a product. Each interval contributes a factor like $\exp(-R_i\tau_i)$ and the $n-1$ intervals terminated by coalescent events each contribute an extra λ factor. The probability density for the joint distribution of tree coalescent times and topology is then

$$f_G(g|\lambda) = \lambda^{n-1} \prod_{i=1}^{2n-2} e^{\lambda \frac{k_i(k_i-1)}{2}(t_{i+1}-t_i)}. \tag{2}$$

Return to the comments at the end of Section 2. The mutation model gives us a likelihood for the age of an organism if we knew something about the DNA sequences of that organism and its descendants, and the mutation rate. Evidence for the DNA sequence of the ancestral organism is obtained by putting the DNA sequences of its descendants together with prior information about the likely tree structure. The Wright-Fisher population model determines a prior for trees. So, we can expect to be able to date coalescent events, if we have DNA sequences for individuals at the leaves of the genealogy, and know the two parameters μ and λ. Can we reconstruct and date a genealogy from the DNA sequences of its leaf individuals if we have the prior and observation models, but the two rate parameters and the genealogy itself are unknown?

4 Inference for rate parameters

In this section we write down the posterior probability distribution for those unknown parameters of interest which may be estimated from the sequence data of n individual organisms. Exact sequence ages are assumed. We defer treatment of radiocarbon calibration to Section 6.

It is convenient at this point to drop the time ordering of vertex labels in g. We will want to make small independent changes to the ages of vertices of g, and we would like them to keep their names as we vary their ages. Let I $[Y]$ denote the set of leaf [ancestral] node labels. Let $t(g) = (t_1, t_2 \dots t_{2n-1})$ where t_i is the age of vertex i in genealogy g. Split the vector $t(g)$ into two vectors, $t_I = (t_{I_1}, t_{I_2} \dots t_{I_n})$ and $t_Y = (t_{Y_1}, t_{I_2} \dots t_{Y_{n-1}})$. Let $R \in Y$ denote the label of the root node. Let $E(g)$ denote the edge set of g, with the convention $\langle i, j\rangle \in E(g) \;\Rightarrow\; t_i \geq t_j$. A genealogy is determined by its edge set and vertex times, $g = (E, t)$. Let B be a $(2n-1) \times L$ array of DNA characters. A row of B corresponds to a DNA sequence. Let $B_{i,:}$ denote the ith row of B and $B_{i,s} \in \mathcal{C}$, $(\mathcal{C} = \{A, C, G, T\})$ denote the character at site s in the DNA sequence for vertex i. Let B_I and B_Y denote the sub-arrays of leaf and ancestral node sequences respectively.

We can think of the coalescent process, which determines the tree-genealogy g, as laying down the railway tracks, along which the mutation process runs. The root sequence is drawn from the equilibrium of the mutation process. The transition probability $\Pr\{B_{j,s} = b | B_{i,s} = a, \mu, t_i - t_j\}$ of Equation (1) carries the sequence down from one node to the next down the tree. The probability $\Pr\{B_I, B_Y | g, \mu\}$ for the mutation process acting over tree g to generate ancestral sequences B_Y and leaf data B_I is then

$$\Pr\{B_I, B_Y | g, \mu\} = \prod_{s=1}^{L} \pi_{B_{R,s}} \prod_{\langle i,j\rangle \in E(g)} \left[e^{Q\mu(t_i - t_j)} \right]_{B_{i,s}, B_{j,s}} . \tag{3}$$

Our data D are the n dated sequences $D = \{t_i, B_{i,:}\}, i \in I$. The tree topology, E, the $n-1$ undated sequences $\{t_j, B_{j,:}\}, j \in Y$ and the mutation and coalescent rate parameters μ and λ are unknown. Let $x = (\mu, \lambda, E, t_Y, B_Y)$ denote the set of unknowns in this problem. Our inference is based on the posterior probability density $f_{X|D}(x|B_I, t_I)$. Suppose a prior density $p(\mu, \lambda)$ is given. We write the posterior as a product of the conditional probabilities determined by the mutation and coalescent processes,

$$f_{X|D}(x|B_I, t_I) \propto \Pr\{B_I, B_Y | g, \mu\}\ f_G(g|\lambda)\ p(\mu, \lambda). \tag{4}$$

Note that we write g where E, t_I and t_Y appear together.

We now discuss the problem of deciding a prior for μ and λ. This is a density on just two variables, both of which are scale parameters. There should be little mystery in the business. The difficulties treated below arise because we choose to illustrate our methods for a diffuse prior, so we take a paragraph to justify this choice. First, we wish to establish sampling methodology, and the MCMC sampling problems we consider become more difficult as we use a more diffuse prior. Secondly, any careful Bayesian inference must make some model sensitivity analysis, and a straightforward way to do this is to probe the data with more and less informative priors. For that reason the diffuse priors we consider here may be of use in a more informed analysis. Thirdly and finally, naive application of a diffuse prior leads to an improper posterior for the parameters of interest. It is useful to warn against this error, and to identify readily available scientific knowledge which is sufficient to fill the gap.

In Drummond *et al.* 2002 we show that a fairly straightforward MCMC scheme is adequate for inferential problems of real interest. Earlier work (Kuhner *et al.* 1995) treated the estimation of λ and g if μ was known, and of μ and g if λ was known. Drummond *et al.* 2002 estimate μ, λ and g jointly, a substantially harder problem.

Joint estimation is not possible if the data are exactly contemporaneous, that is, if $t_i = t_j$ for all $i, j \in I$. Shift the zero of time so that $t_i = 0, i \in I$ and fix a real constant $c > 0$. Consider Equation (4) and the transformation $x \to cx$ defined by $cx = (\mu/c, \lambda/c, E, ct_Y, B_Y)$. The rates go down as the tree depth goes up. The factors $\Pr\{B_I, B_Y|g, \mu\}$ and $f_G(g|\lambda)d^{n-1}t_Y$ are invariant under this transformation. The only c-dependence left in the posterior is in the prior distribution for μ and λ. In other words, the data tells us nothing about c that we didn't already know. Here is another way to think about the problem. The transformation $x \to cx$ does not scale the leaf node times, just the ancestral node times, so it is not in general simply a change in the units of time. However, for equal-time leaves $x \to cx$ is indistinguishable from a change of units for time since the leaves are at time zero, and c times zero is zero.

This argument does not go through when leaf vertices are not all contemporaneous. In Equation (4), c does not cancel in factors involving edges connecting leaf and ancestral nodes. The time offset between leaves makes one time scale special, and time-scale invariance is lost in Equation (4). However, although the qualitative property of identifiability is present whenever leaf times are not exactly contemporaneous, we can expect a great deal of uncertainty in the scale factor c when the leaf spacing is slight. In particular, if we take c very large, the spacing between leaves is small compared to the time scale for events in the tree, and the identifiability problem present for equal time-leaves reappears.

> **Warning** The posterior density $f_{X|D}$ of Equation (4) is improper for $p(\mu, \lambda)$ proportional to $1/(\mu\lambda)$.

The warning tells us that "naive" non-informative inference is not possible for the joint estimation of μ, λ and g. A proof of this result is given in Appendix A. The basic problem is that the factors $\Pr\{B_I, B_Y|g, \mu\}$ and $f_G(g|\lambda)$ do not go to zero sufficiently fast as $c \to \infty$ to yield a finite integral over g, μ and λ.

What state of knowledge does determine a proper posterior? Of course this is in a certain sense straightforward. Any proper $p(\mu, \lambda)$ will do the job and in most applications a very little elicitation will determine such a prior. However, as we mentioned above, it is useful, for the purpose of sensitivity analysis, and for challenging MCMC algorithms, to consider very diffuse priors. It is appealing to biologists to allow states at $\mu \to 0$ and $\lambda \to 0$, at least in exploratory analysis. However, a conservative bound t_R^* on the maximum age of the root was readily approved.

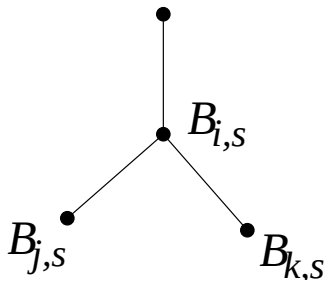

Figure 2: A subtree with characters $B_{i,s}$, $B_{j,s}$, $B_{k,s}$ at site s of the sequences at vertices i, j, and k respectively.

> **Encouragement** Suppose the data B_I contains at least two non-identical sequences. Let positive constants μ^*, λ^* and t_R^* be given. The posterior density $f_{X|D}$ of Equation (4) is a proper probability density if $p(\mu, \lambda)$ is proportional to $1/(\mu\lambda)$ and the conditions $\mu < \mu^*$, $\lambda < \lambda^*$ and $t_R < t_R^*$ apply.

This result is established in Appendix B. Why is it worth stating? We saw that integration of $f_{X|D}$ along any ray $\{(\mu/c, \lambda/c, ct_Y); c > 0\}$ is undefined. Divergences arise in two additional limits, $\mu \to \infty$ and $\lambda \to \infty$. These wildly unphysical states cause problems. Firstly

$$\lim_{\mu\to\infty} \Pr\{B_I, B_Y|g, \mu\} = \prod_{s=1}^{L} \prod_{i=1}^{2n-1} \pi_{B_{i,s}}.$$

At high mutation rates there is essentially an instantaneous mutational equilibrium, and the likelihood goes to a non-zero constant corresponding to the equilibrium base frequencies. Second, the coalescent density $f_G(g|\lambda)$ concentrates on very short trees as $\lambda \to \infty$, giving rise to a non-integrable divergence. In terms of the original Wright-Fisher model, $\lambda = 1/(N_e\rho)$, so this divergence arises in the zero population limit. Before we do anything else, we must eliminate these states, and we do that with $\mu < \mu^*$, $\lambda < \lambda^*$. We are now ready to deal with the divergence along rays $\{(\mu/c, \lambda/c, ct_Y); c > 0\}$. These rays are truncated by the root age bound $t_R < t_R^*$. Since the space of states is now closed and bounded, any bounded prior gives a proper posterior, so our encouragement refers only to the scale invariant prior, $1/(\mu\lambda)$.

Notice that a maximum likelihood estimate of λ and μ, from the integrated likelihood surface $\Pr\{B_I|\mu, \lambda\}$, can be made using draws from $f_{X|D}$. An estimate of the marginal posterior surface for μ and λ becomes an estimate of the likelihood surface $\Pr\{B_I|\mu, \lambda\}$ (up to an overall normalization) by simply dividing out by $p(\mu, \lambda)$.

5 Pruning

Before we continue, we note that we have the option, in this inference, to sum out the (typically) uninteresting unknown ancestral sequences B_Y and work with state $x = (\mu, \lambda, E, t_Y)$ and likelihood $\Pr\{B_I|g, \mu\}$. We may compute the sum $\Pr\{B_I|g, \mu\} = \sum_{B_Y} \Pr\{B_I, B_Y|g, \mu\}$ over all $B_Y \in \{A, C, G, T\}^{(n-1)\times L}$ numerically, without recourse to Monte Carlo, using the pruning algorithm of Felsenstein 1981. Vertex j is a child vertex of vertex i in genealogy g if there exists an edge $\langle i, j\rangle \in E(g)$. Suppose vertex i has child vertices j and k (see Figure 2).

For $a \in \mathcal{C}$, the likelihood $\Pr\{B_I|g,\mu,B_{i,s}=a\}$ can be written in terms of the corresponding likelihoods evaluated at j and k,

$$\begin{aligned}\Pr\{B_I|g,\mu,B_{i,s}=a\} &= \sum_{b\in\mathcal{C}}\left[e^{\mu Q(t_i-t_j)}\right]_{a,b}\Pr\{B_I|g,\mu,B_{j,s}=b\} \quad \times \\ &\quad \sum_{b'\in\mathcal{C}}\left[e^{\mu Q(t_i-t_k)}\right]_{a,b'}\Pr\{B_I|g,\mu,B_{k,s}=b'\}\end{aligned}$$

The likelihood at the root $\Pr\{B_I|g,\mu,B_{R,s}=a\}$ is defined by a recursion of the above expression down the tree to the leaves. Let $\mathbb{I}_{\mathcal{E}}$ denote the indicator function for the event $\mathcal{E}$. If j is a leaf then $j \in I$ and $\Pr\{B_I|g,\mu,B_{j,s}=b\} = \mathbb{I}_{B_{j,s}=b}$. There are two sums over four elements at each level of the recursion. The integrated likelihood is given in terms of the equilibrium frequencies, π, defined in Section 2, by

$$\Pr\{B_I|g,\mu\} = \prod_{s=1}^{L}\sum_{b\in\mathcal{C}}\pi_b \Pr\{B_I|g,\mu,B_{R,s}=b\}.$$

6 Uncertainty in fossil sequence dates

There is no discussion, in the literature to date, of the likely impact of radiocarbon calibration errors on genetic inference. However, in paired studies which we omit from the present work, in which we alternately treat and ignore the uncertainty due to radiocarbon calibration, we find that the effect is not important (so, for example, Lambert *et al.* 2002 are correct to ignore the issue). By far the greatest part of the uncertainty in reported ages comes from the uncertainty in the rate parameters and genealogy (and, no doubt, model mispecification error). There are around $(\mu L)^{-1}$ years (*ie*, about 2000 years in Section 8) between mutations on a single lineage. This sets a lower bound on the order of magnitude of the size of the error bars for all age estimates (both leaves and ancestral nodes) at a value far above typical radiocarbon uncertainty. However, since mutation rates vary from species to species, researchers should keep an eye on this source of uncertainty.

In the following we explain how to treat radiocarbon calibration as an explicit part of the population-genetic inference. For each leaf $i \in I$, let T_i, y_i and σ_i denote respectively the unknown true age, the conventional radiocarbon age, and measurement error associated with DNA sequence $B_{i,:}$. Let $d(\tau)$ denote the radiocarbon calibration curve with age dependent error $\sigma(\tau)$ as published in Stuiver *et al.* 1998. We fit the standard radiocarbon observation model (described, for example, in Buck *et al.* 1991) for the data y_I, that is

$$y_i \sim d(T_i) + \epsilon(T_i) + \epsilon_i,$$

where $\epsilon(T_i) \sim N(0,\sigma(T_i)^2)$ and $\epsilon_i \sim N(0,\sigma_i)$ are unknown additive Gaussian noise variates. In the absence of sequence data, the posterior density $f_{T_I}(t_I|y_I) \propto f_{Y_I|T_I}(y_I|t_I)f_{T_I}(t_I)$, for the calibrated ages t_I is given in terms of a radiocarbon likelihood $f_{Y_I|T_I}$ and a prior f_{T_I}.

We elicit a prior density f_{T_I} as follows. Suppose the effective population $N_e(\tau)$ is a function of age τ. Suppose that age termini A and P are available, so that for $i \in I$, $A \le t_i \le P$ is prior knowledge, but that otherwise each data sequence $B_{i,:}$ might equally well belong to any individual in the population history from P to A. The state of knowledge described above is, therefore, represented by a prior density

$$f_{T_I}(t_I) = \prod_{i\in I}\frac{N_e(t_i)}{\int_A^P N_e(\tau)d\tau}$$

defined for $t_I \in [A, P]^n$ (recall, n leaf labels in I). Where λ is estimated as a function of time, it will be necessary to model the action of taphonomy and specimen selection on recovered fossil DNA. In what follows we assume N_e is a constant and ignore selection due to taphonomy. In this setting the above considerations lead to $f_{T_I}(t_I) \propto 1$, the constant prior. This form is used throughout the radiocarbon literature, from Buck *et al.* 1991 onwards. We choose it, in the example which follows, not because we are reaching for some default, non-informative prior, but because it is computed from a simple explicit model (of the kind described in Nicholls and Jones 2001) of the processes which realize the parameters in question. See Drummond *et al.* 2002 for a discussion of the age-dependent case.

We must modify Equation (4) to take into account the uncertainty arising from the calibration. Our data D are the n dated sequences $D = \{y_i, B_{i,:}\}, i \in I$. Let $x = (\mu, \lambda, E, t_I, t_Y, B_Y)$ denote the set of unknowns in this problem. The leaf times t_I have joined the set of unknowns. The revised posterior is

$$f_{X|D}(x|B_I, y_I) \propto \Pr\{B_I, B_Y|g, t_I, \mu\}\ f_G(g|t_I, \lambda)\ f_{Y_I|T_I}(y_I|t_I)\ f_{T_I}(t_I)\ p(\mu, \lambda) \qquad (5)$$

Notice that undated leaves introduce the possibility of an improper posterior, since the likelihood for an undated leaf, attached at an age, τ say, greater than the root age t_R, does not go to zero as $\tau \to \infty$. Some upper bound on the leaf (or root) age must be provided as prior knowledge. If the data is sufficiently informative it is possible to set this upper bound to an extremely conservative value. The mass of probability in the upper tail of the age distribution is then negligible. This is the case in the example of Section 8.

7 MCMC

We have implemented an MCMC algorithm generating $X \sim f_{X|D}$. In fact we made three more or less independent implementations of the entire MCMC scheme. One, in MatLab, does not represent sequences B_Y on ancestral vertices, using the above pruning scheme to eliminate those variables. This first implementation samples the marginal posterior distribution for μ, λ and g obtained by summing B_Y out of $f_{X|D}$. A second implementation, also in MatLab, does represent the ancestral sequences in the state, placing them on an equal footing with μ, λ and g in the Monte Carlo. A third Java implementation uses pruning. The multiple implementations were used for checking and debugging, and to investigate the relative efficiency of pruning. Pruning proved to be particularly helpful at high mutation rates, where there is real uncertainty in the ancestral sequences.

We describe in Appendix C a collection of MCMC updates for the case where ancestral sequences are an explicit part of the MCMC simulation. We give details for those updates which are in our opinion difficult to compute, or interesting in other respects. Updates for the two implemented MCMC samplers which use pruning are described in Drummond *et al.* 2002. Updates of the kind discussed below, in particular, updates which treat sequences at ancestral nodes explicitly, may be found elsewhere, for example, Wilson and Balding 1998 and Wilson *et al.* 2003. An even more "explicit" treatment may be found in Beaumont 1999, where individual mutation events are represented. Wilson and Balding 1998 and Beaumont 1999 treat microsatellite data, as opposed to DNA sequence data.

For readers who wish to apply the methods described in this paper to data sets of their own, the MEPI software

`http://www.cebl.auckland.ac.nz/mepi/index.html`

makes the business as straightforward as one could reasonably hope.

8 Example

By way of example, we consider a synthetic problem set up to resemble the problem treated in Lambert *et al.* 2002. We allow for the uncertainty arising from the simultaneous estimation of mutation and coalescent rates and genealogy (μ, λ and g), and from the calibration of radiocarbon-dated mtDNA sequences. Apart from Drummond *et al.* 2002, Lambert *et al.* 2002 is the only published analysis to take into account uncertainty arising from the simultaneous estimation of μ, λ and g. We illustrate "genetic dating" of leaf sequences and common ancestors.

Lambert *et al.* 2002 treat fossil sequences from penguins. They sequenced the mitochondrial HVRI region using material from 96 ancient bone samples, up to around 6500 years in age, and 380 blood samples from modern birds at 13 Antarctic locations. Their analysis is based on 352 aligned sites in the 96 fossil sequences and an unpublished subset of the modern sequences. We simulate $n = 22$ sequences of length $L = 400$ with no gaps. We leave two sequences (from the bottom and middle of the genealogy) completely undated, in order to illustrate genetic dating. The dated sequences in the data allow us to say something about the unknown ages of the undated sequences. Lambert *et al.* 2002 do not publish an estimate of the effective population size. We suppose $N_e = 1000$, an order of magnitude for populations of this sort. They assume a generation time of $\rho = 5.5$ years and estimate a mutation rate of around 10^{-6} mutations per site per year.

Let Λ, M, G and T_I denote the synthetic true coalescent and mutation rates, synthetic true genealogy and leaf times. In line with Lambert *et al.* 2002, we choose $\Lambda = 1/5500$ and $M = 10^{-6}$. We distribute the 22 true leaf times T_I uniformly between the present and 11000BP, and, for $i \in I$, simulate synthetic radiocarbon data $y_i \sim f_{Y|T}(\cdot|T_i)$. Synthetic sequence data is drawn by simulating a true tree $G \sim f_G(\cdot|\Lambda, T_I)$ (see Figure 3 (left side)), drawing synthetic root characters $B_{R,s} \sim \pi$ for each $s = 1, 2 \ldots L$, and then simulating leaf sequences $B_I \sim \Pr\{\cdot|G, M, B_{R,\cdot}\}$ by simulating the mutation process $\Pr\{B_{j,s} = b|B_{i,s} = a, M, t_i - t_j\}$ down each edge $\langle i, j\rangle \in E(G)$ from the root to the leaves. In order to make the inference proper, we impose upper limits, $\mu^* = 1$ mutation per site per year, $\theta^* = 1/5.5$ (so $N_e \geq 1$) and $t_R^* = 40000$. Lower limits are all zero. The first two bounds are almost completely uninformative. In fact the Monte Carlo did not visit any of the bounds. The posterior probability of states in the vicinity of the bounds is negligible.

We carry out sample-based Bayesian inference, simulating $X_k \sim f_{X|D}, k = 0, 1 \ldots K$ with $K = 5 \times 10^6$ using the MCMC scheme of Appendix C. The MCMC is started with a tree drawn from f_G, the coalescent prior. A tree g, sampled from the posterior (simply the last tree in the run) is shown in Figure 3 (right side). The MCMC output for the slowest mixing statistic (that is, the state function $h(x)$ with the greatest integrated autocorrelation time) can be seen in Figure 4 along with its autocorrelation function, and its large-lag asymptotic variance ($\pm 2\sigma$). The run contains about 400 effective independent samples. For details of these convergence diagnostics, see Geyer 1992.

In Figure 5 we present scatter plots of posterior samples (μ, t_R) and (λ, t_R) with the "true" values indicated by cross-hairs. The points lie on hyperbola, reflecting the fact that points in parameter space on the ray $(\mu/c, \lambda/c, ct_Y), c > 0$ are not well distinguished by the data.

No radiocarbon dates were provided for leaves one and fifteen of the synthetic true tree in Figure 3. Marginal posterior distributions for the ages of the two undated sequences are given in Figure 6. The reconstructed age distributions of Figure 6 have a width which goes down as μ and the aligned sequence length L increase. In the present setting, the low accuracy is driven by the relatively low mega-faunal mutation rate μ, and by uncertainty in

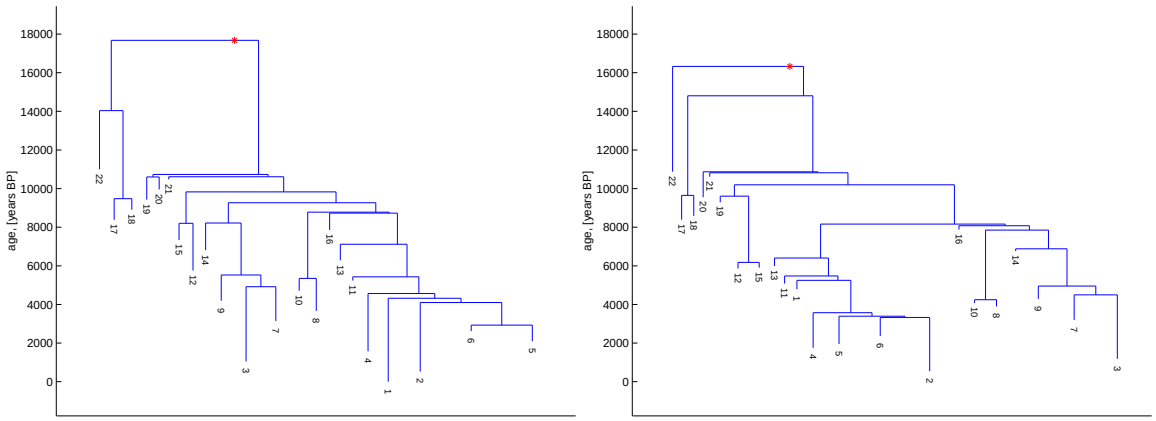

Figure 3: (left) Synthetic data, the true genealogy. Leaf labels correspond to distinct fossil sequences. (right) A genealogy sampled from the posterior distribution.

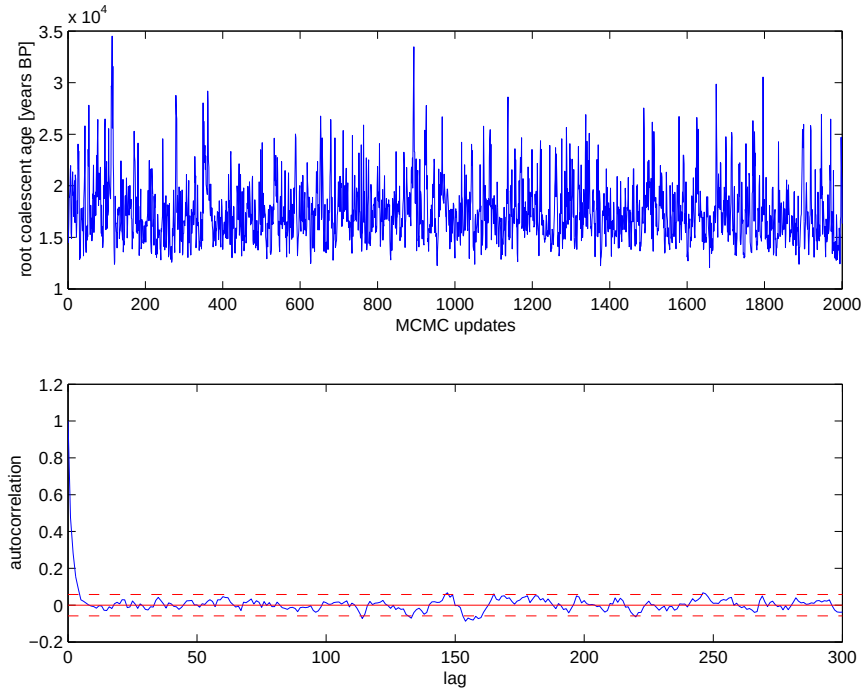

Figure 4: (top) MCMC output for the slowest mixing statistic, the root coalescent time t_R. The x-axis is MCMC updates ($\times 2500$). The autocorrelation time of this statistic was about 13000 updates. The total time for the run shown was around 18 hours, in a MatLab implementation, on a laptop purchased in 2001. (bottom) The serial autocorrelation function of the trace above. Horizontal lines show asymptotic variance at $\pm 2\sigma$ (Geyer 1992).

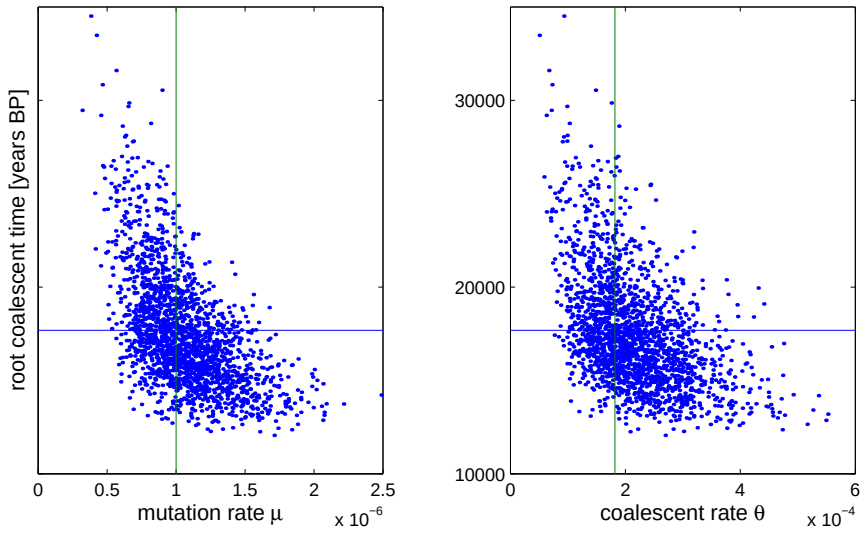

Figure 5: Scatter plots of posterior samples of (μ, t_R) (left) and (λ, t_R) (right). Cross-hairs indicate true parameter values. Note: $\lambda = 1/(N_e\rho)$, see Section 3 for details.

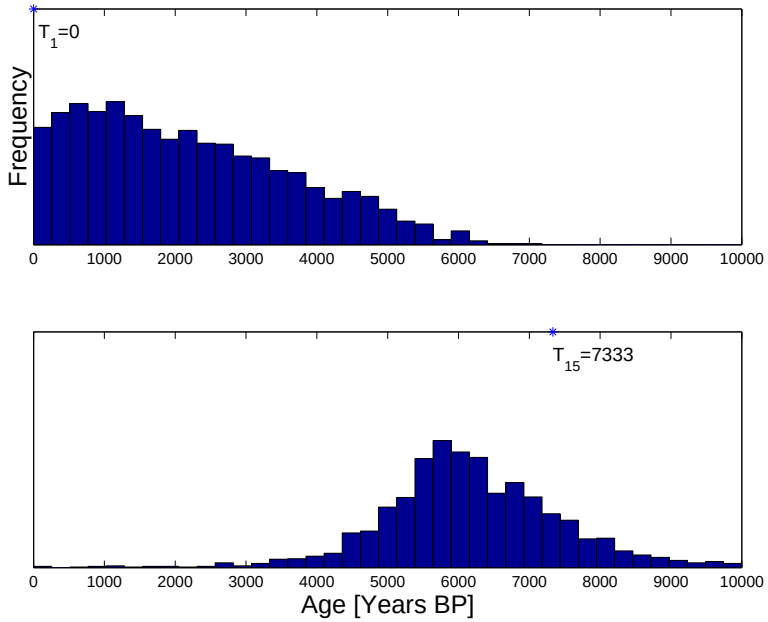

Figure 6: Ancient or modern? Marginal posterior distributions for the ages of undated mtDNA sequences. (top) leaf 1 of Figure 3, (bottom) leaf 15. Synthetic true ages indicated are indicated with an asterix.

the value of that rate. For species of particular archaeological interest (we have in mind the polynesian rat) some of this uncertainty could be removed using independent measurements of mutation rates.

The model we are fitting makes a number of assumptions which are unlikely to hold for this population. The real population size is not constant. The animals are distributed in breeding colonies which intermingle, but are not panmictic. The mutation process is correlated along the sequence, and is subject to selection. Because we ignore these properties, immediate chronometric conclusions cannot be drawn from an analysis of the kind presented above. The aim here is illustrate sample-based genealogical inference in a simple setting. Nevertheless, the model we are fitting is the natural null model for this kind of problem. Departures from this model may be expected, and in future work evidence for such features will no doubt be sought. However those future model comparison studies will be made relative to this model, or something very similar, and will need to make a fit of the kind made in this section.

References

Ammerman, A. J., and Cavalli-Sforza, L. L. 1984. *The Neolithic transition and the genetics of population in Europe*. Princeton: Princeton University Press.

Barnes, I., Matheus, P., Shapiro, B., Jensen, D., and Cooper, A. 2002. Dynamics of pleistocene population extinctions in Beringian brown bears. *Science*, **195**, 2267–2270.

Beaumont, M.A. 1999. Detecting population expansion and decline using microsatellites. *Genetics*, **153**, 2013–2029.

Buck, C.E., Kenworthy, J.B., C.D.Litton, and Smith, A.F.M. 1991. Combining archaeological and radiocarbon information: a Bayesian approach to calibration. *Antiquity*, **65**, 808–821.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., and Solomon, W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, **161**, 1307–1320.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.

Geyer, C. J. 1992. Practical Markov chain Monte Carlo (with discussion). *Statist. Sci.*, **7**, 473–511.

Green, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

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

Kingman, J. F. C. 1982a. The coalescent. *Stoch. Proc. Appl.*, **13**, 235–248.

Kingman, J. F. C. 1982b. On the genealogy of large populations. *J. Appl. Probab.*, **19A**, 27–43.

Kuhner, M. K., J., Yamato, and Felsenstein, J. 1995. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, **140**, 1421–1430.

Lambert, D. M., Ritchie, P. A., Millar, C. D., Holland, B., Drummond, A. J., and Baroni, C. 2002. Rates of evolution in ancient DNA from Adelie penguins. *Science*, **195**, 2270–2273.

Matisoo-Smith, E., Roberts, R.M., Irwin, G.J., Allen, J.S., Penny, D., and Lambert, D.M. 1998. Patterns of prehistoric human mobility in Polynesia revealed by mitochondrial DNA from the Pacific rat. *Proceedings of the National Academy of Sciences*, **95(25)**, 15145–15150.

Mau, B., Newton, M. A., and Larget, B. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, **55**, 1–12.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. 1953. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Nicholls, G. K., and Jones, M. 2001. Radiocarbon dating with temporal order constraints. *J. R. Statist. Soc. C*, **50**, 503–521.

Penny, D., Watson, E., and Steel, M. 1993. Trees from languages and genes are very similar. *Systematic biology*, **42**, 382–384.

Rambaut, A. 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics*, **16**, 395–399.

Rodrigo, A. G., and Felsenstein, J. 1999. Coalescent approaches to HIV population genetics. *Page 233 of:* Crandall, K. (ed), *Molecular Evolution of HIV*. Baltimore: Johns Hopkins University Press.

Stuiver, M., Reimer, P.J., E. Bard, J.W. Beck, Burr, G.S., Hughen, K.A., Kromer, B., McCormac, F.G., Plicht, J.v.d., and Spurk, M. 1998. Intcal98 radiocarbon age calibration, 24,000-0 cal BP. *Radiocarbon*, **40**, 1041–1083.

Suchard, M. A., Weiss, R. E., and Sinsheimer, J. S. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. *Molecular Biology and Evolution*, **18**, 1001–1013.

Suomela, P. 1976. *Construction of nearest neighbour systems.* Ph.D. thesis, Department of Mathematics, University of Helsinki.

Tierney, L. 1994. Markov chains for exploring posterior distributions. *Ann. Statist.*, **22**, 1701–1728.

Underhill, P. A., Passarino, G., Lin, A. A., Marzuki, S., Cavalli-Sforza, L. L., and Chambers, G. 2001. Maori origins, Y-chromosome haplotypes and implications for human history in the Pacific. *Human mutation*, **17**, 271–280.

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

Wilson, I. J., Weale, M. E., and Balding, D. J. 2003. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *J. R. Statist. Soc. A*, **166**, 1–33.

Appendix A: Warning

In this appendix we show that the posterior density $f_{X|D}$ of Equation (4) is improper when $p(\mu, \lambda) \propto 1/(\mu\lambda)$. Let

$$P(A|B_I) = \int_{A\cap\Omega} f_{X|D}(x|B_I)dx$$

Let $\epsilon > 0$ and a compact subset A of Ω be given, satisfying $P(A|B_I) > 0$ and, for each $(\mu, \lambda, E, t_Y, B_Y) \in A$, $\mu > 0$, and $\min(t_Y) > \max(t_I) + \epsilon$. For each $c > 1$ let $cA = \{cx : \ x \in A\}$, and, for $g = (E, t_Y, t_I)$, $cg = (E, ct_Y, t_I)$. We now show that $P(A|B_I) = P(cA|B_I)$. First, $\Pr\{B_Y, B_I|g, \mu\} = \Pr\{B_Y, B_I|cg, \mu/c\}$ from Equation (3). Next, consider $f_G(cg|\lambda/c)$ in Equation (2). The number of lineages present at the time t_i of coalescent node i depends on the times of all other nodes in the tree, that is, $k_i = k_i(t_Y, t_I)$ in the rate $k_i(k_i - 1)/2$ for coalescence at time t_i. However, A is defined so that, for $x \in A$, $k_i(t_Y, t_I) = k_i(ct_Y, t_I)$ for all $i = 1, 2 \ldots 2n - 1$ and all $c > 1$, and consequently $f_G(cg|\lambda/c) = f_G(g|\lambda)/c^{n-1}$. The priors $d\mu/\mu$ and $d\lambda/\lambda$ are scale invariant, and volume element dt_Y contributes c^{n-1} so the change of variables $x' = cx$ in $P(cA|B_I)$ gives us back $P(A|B_I)$. Now, there exists a sequence, $1 < c_1 < c_2 < c_3 \ldots$ with the property that $A, c_1A, c_2A, c_3A \ldots$ are mutually disjoint. Since $P(A|B_I) > 0$, and $P(c_nA|B_I) = P(A|B_I)$ for each $n = 1, 2, 3 \ldots$, it follows that P is improper.

Appendix B: Encouragement

In this appendix we show that the conditions given in our encouragement in Section 4 determine a proper posterior. First we bound $\Pr\{B_Y, B_I|g, \mu\}$ away from one, for $t_R \le t_R^*$. Since the B_I are not identical, there is at least one mutation over g. The probability to get the sequences B_I and B_Y is less than the probability that there is at least one mutation (since B_I implies a mutation), which is one minus the probability for no mutations on g,

$$\begin{aligned}\Pr\{B_Y, B_I|g, \mu\} &\le 1 - \sum_{b\in\mathcal{C}} \pi_b e^{\mu Q_{b,b}|g|} \\ &\le 1 - \sum_{b\in\mathcal{C}} \pi_b \min_{b\in\mathcal{C}} e^{\mu Q_{b,b}(2n-2)t_R^*}\end{aligned}$$

where $|g| = \sum_{\langle i,j\rangle\in E(g)} |t_i - t_j|$ is the total edge length. Note that $Q_{b,b}$ is negative and $|g| \le (2n-2)t_R^*$, as the greatest tree length is less than the number of edges times the maximum edge length. Let Γ denote the set of all genealogies g allowed for leaf times t_I and given $t_R \le t_R^*$. Integration dg involves integration $d^{n-1}t_Y$ and summation over all distinct tree topologies E. The normalizing constant $Z = \int f_{X|D}dx$ is

$$\begin{aligned}Z &= \int_0^{\mu^*}\int_\Gamma\int_0^{\lambda^*} \left[\sum_{B_Y\in\mathbb{B}} \Pr\{B_Y, B_I|g, \mu\}\right] f_G(g|\lambda)p(\lambda, \mu)d\lambda dg d\mu \\ &\le 4^{(n-1)L}\int_0^{\mu^*} \frac{1}{\mu}\left(1 - \min_{b\in\mathcal{C}} e^{\mu Q_{b,b}(2n-2)t_R^*}\right)d\mu \\ &\quad\times \int_\Gamma\int_0^{\lambda^*} \lambda^{n-2}\exp\left(-\frac{\lambda}{2}\sum_{i=1}^{2n-2} k_i(k_i-1)(t_{v(i)+1} - t_{v(i)})\right)d\lambda dg\end{aligned}$$

The integral over μ is finite. The integration $d\lambda dg$ is a sum, over tree topologies, which are finite in number. Each term of that sum is given by the integral $d\lambda dt_Y$ of a bounded function over a bounded domain. It follows that the posterior is proper.

Appendix C: Markov chain Monte Carlo

In this Appendix we give an MCMC scheme for states with ancestral sequences. For our example, we suppose $p(\lambda, \mu) = (\lambda\mu)^{-1}$. Let Ω denote the space of states $x = (\mu, \lambda, E, t_Y, B_Y)$. We specify a Markov chain X_n, $n = 0, 1, 2 \ldots$, with states, $X_n \in \Omega$, and equilibrium $f_{X|D}$. Metropolis *et al.* 1953 Hastings 1970 and Green 1995 define a class of Monte Carlo update algorithms which determine a transition matrix stationary with respect to a given target distribution.

Suppose $X_n = x$. A value for X_{n+1} is computed using the Metropolis-Hastings algorithm. First, a candidate state x' is generated by randomly perturbing x in some way. An operation of type m is chosen at random from a list $m = 1, 2 \ldots \mathcal{M}$ of operation types. The state x' is generated. This is implemented by drawing uniform random variates $u = (u^{(1)}, u^{(2)} \ldots)$ according to according to a density $q_m(u)$, and computing some function $x' = x_m(x, u)$. For example, to do a random walk update to μ with constant window size $z > 0$, draw $u \sim U(0, 1)$ and set $\mu' = \mu + z(2u - 1)$. Consider now the reverse operation. Suppose the draw $u' \sim q_m$ maps x' back to x, so that $x = x_m(x', u')$. In the random walk example, $u' = 1 - u$. Secondly, we accept the candidate, and set $X_{n+1} = x'$ with probability

$$\alpha_m(x'|x) = 1 \ \wedge \ \frac{f_{X|D}(x'|B_I, t_I) q_m(u')}{f_{X|D}(x|B_I, t_I) q_m(u)} \left| \frac{\partial(x', u')}{\partial(x, u)} \right|,$$

(where $a \wedge b$ equals a if $a < b$ and otherwise b) for update type m. If the candidate is not accepted, we set $X_{n+1} = x$, so the state of the chain is unchanged.

We may choose the proposal scheme $q_m, x_m, m = 1, 2 \ldots M$ as we please, subject to conditions outlined, for example, in Tierney 1994. The role of the Jacobean factor is clarified in Green 1995 in a general setting. As an example, in the random walk update above, the relevant block of the Jacobian matrix is $\begin{bmatrix} 1 & 2 \\ 0 & -1 \end{bmatrix}$ so the absolute value of the determinant is equal one and the acceptance probability for the random walk update to μ is $1 \wedge f_{X|D}(x'|B_I, t_I)/f_{X|D}(x|B_I, t_I)$. In our MCMC we need to use an update in which this Jacobian factor is not equal one. We saw, in Section 4, that the posterior density can be expected to posses a ridge, which we may move along using the operator $x' = cx$. The update then is as follows. Suppose $X_n = x$. Choose $c \sim U(1/2, 2)$. Set $x' = (\mu/c, \lambda/c, E, ct_Y, B_Y)$. This may result in $x' \notin \Omega$ (for example scaled ancestral node ages ct_Y may violate of the parent-child age order relation for edges in E). If this is the case x' will be rejected at the next step. If $x' \in \Omega$, the candidate is admissible, and the acceptance probability is $1 \wedge c^{n-5} f_{X|D}(x'|B_I, t_I)/f_{X|D}(x|B_I, t_I)$. Let us see how the factor c^{n-5} arises. Since $t'_Y = ct_Y$, $\mu' = \mu/c$, $\lambda' = \lambda/c$ and $c' = 1/c$ (so that $x = c'x'$), $\partial(x', c')/\partial(x, c)$ has diagonal $(c, \ldots [n-1 \text{ repeats}] \ldots c, c^{-1}, c^{-1}, -c^{-2})$. The off-diagonal elements are zero, except the last column which contains non-zero elements. The determinant of this matrix is $-c^{n-5}$.

Some of the parameters of the problem may feasibly be Gibbs-sampled (Suomela 1976). For parameter $p \in x$, let $x_{-p} = x \setminus \{p\}$ denote x with p omitted. The conditional density of $\lambda|x_{-\lambda}, D$ in $f_{X|D}$ is

$$\lambda|x_{-\lambda}, D \ \sim \ \lambda^{n-2} e^{-\beta\lambda} \mathbb{I}_{\lambda \leq \lambda^*}$$

where

$$\beta = \frac{1}{2} \sum_{i=1}^{2n-2} k_i(k_i - 1)(t_{v(i)+1} - t_{v(i)}).$$

Here $v(i)$ is a mapping from the unordered node labels of Section 4 back to the age-ordered node labels of Section 3. This is just a Gamma$(n-1, \beta)$ density on $\lambda \leq \lambda^*$.

As discussed above, we have the option to sum out the ancestral sequences B_Y from the posterior distribution, using the pruning algorithm. If that is done, B_Y does not arise in x, the Monte-Carlo state. If we choose not to prune, so B_Y is part of x, then we need some MCMC update for the conditional distribution of $B_Y|B_I, g, \mu$ determined by $f_{X|D}$. This conditional distribution is in fact a Markov Random Field (MRF), in which each of the $n-1$, L-component variables $B_i \in \mathcal{C}^L, i \in Y$ is conditionally independent of the rest, given the sequences at its neighbors i_1, i_2 and i_3 on the tree. The neighbors of vertex i in tree g are those vertices to which it is connected by an edge in $E(g)$. In the update below, i_2 and i_3 are i's child vertices. The root vertex of g is the child of a vertex of infinite age. The MRF may be simulated by the following Gibbs update. Select $i \in Y$ uniformly at random. For each $s = 1, 2 \ldots L$ and $b \in \mathcal{C}$, calculate the 4-components

$$\mathcal{B}_b = \left[e^{Q\mu(t_{i_1} - t_i)}\right]_{B_{i_1,s},b} \left[e^{Q\mu(t_i - t_{i_2})}\right]_{b,B_{i_2,s}} \left[e^{Q\mu(t_i - t_{i_3})}\right]_{b,B_{i_3,s}}$$

of the vector $\mathcal{B}(B_{-i,s}, g, \mu) = (\mathcal{B}_A, \mathcal{B}_C, \mathcal{B}_G, \mathcal{B}_T)$. Draw $B_{i,s} = b$ with probability $\mathcal{B}_b / Z_{\mathcal{B}}$ where $Z_{\mathcal{B}}(B_{-i,s}, g, \mu) = \sum_{a \in \mathcal{C}} \mathcal{B}_a$. The acceptance probability is equal one.

It is necessary to have some topology changing update, so that tree-space is explored. We make some small random modification of the tree topology, so that the new state x' is equal x up to $E \to E'$. For example, we can choose two edges $\langle i, i' \rangle$ and $\langle j, j' \rangle$ in E and reconnect them as $\langle i, j' \rangle$ and $\langle j, i' \rangle$. If the resulting tree is not admissible the candidate state will be rejected. The probability to generate E' from E in this way is just the probability to chose the two edges by which they differ, so the acceptance probability is $1 \wedge f_{X|D}(x'|B_I, t_I) / f_{X|D}(x|B_I, t_I)$. Where simulation is made with ancestral sequences an explicit part of the Monte Carlo state, we improve the candidate's chances if we draw new sequences at vertices i and j, using the above Gibbs proposal in the new tree. This is not exactly a Gibbs update, since the conditional distributions $\Pr\{B_i|B_{-i}, g', \mu\}$ and $\Pr\{B_i|B_{-i}, g, \mu\}$ involved in the forward and reverse proposals are normalized on different trees. The acceptance probability $1 \wedge \prod_{k=i,j} \prod_{s=1}^{L} Z_{\mathcal{B}}(B_{-k,s}, g', \mu) / Z_{\mathcal{B}}(B_{-k,s}, g, \mu)$ does not quite collapse down to one.

When we account for the uncertainty in the ages of the sequence data, as in Sections 6 and 8, we need an update varying leaf times t_I, which are otherwise fixed. We use a suite of updates, suggested by our experience in Nicholls and Jones 2001 with MCMC for the posterior distribution of radiocarbon calibration. We omit the Hastings ratios from this paper as they are simply the constant prior density Hastings ratios of Nicholls and Jones 2001 weighted by the likelihood ratio

$$\frac{\Pr\{B_I, B_Y|g, t'_I, \mu\}\ f_G(g|t'_I, \lambda)}{\Pr\{B_I, B_Y|g, t_I, \mu\}\ f_G(g|t_I, \lambda)}.$$

We include a number of other move types, including other topology-changing tree operations. We have a random walk move for node age, acting on a single randomly chosen ancestral node. This move generates its candidate by selecting a new time for the node at random between the time of its parent, and oldest child. We have experimented with a wide range of other moves. However, whilst it is easy to think up computationally demanding updates which improve the convergence and mixing rates per Markov-chain update, it is harder to find updates that improve the convergence and mixing rates per CPU second. Certain move types which may be of value have not been considered. In particular, updates of the kind described in Mau *et al.* 1999 which are natural in the cophentic matrix tree representation, were not considered, though they seem promising.