Adv. Appl. Prob. **27**, 418–442 (1995)
Printed in N. Ireland

THE GENEALOGY OF BRANCHING PROCESSES AND THE AGE OF OUR MOST RECENT COMMON ANCESTOR

NEIL O'CONNELL,* *Dublin Institute for Advanced Studies*

Abstract

We obtain a weak approximation for the reduced family tree in a near-critical Markov branching process when the time interval considered is long; we also extend Yaglom's theorem and the exponential law to this case. These results are then applied to the problem of estimating the age of our most recent common female ancestor, using mitochondrial DNA sequences taken from a sample of contemporary humans.

REDUCED BRANCHING PROCESS; NEAR-CRITICAL; YAGLOM'S THEOREM; EXPONENTIAL LIMIT LAW; MITOCHONDRIAL DNA; EVE; EVOLUTION; MUTATION

AMS 1991 SUBJECT CLASSIFICATION: PRIMARY 60J80
SECONDARY 60J85, 92A10, 92A15

1. Introduction

Let Z be a Markov branching process with mean lifetime 1 and offspring distribution ν. Let ξ be a realisation of ν and set $f(s) = \boldsymbol{E}s^{\xi}$, for $0 \leqq s \leqq 1$. Fix $t > 0$, and for each $0 \leqq s \leqq t$, define $N_t(s)$ to be the number of individuals alive at time s with descendants alive at time t. The process N_t is called the *reduced branching process,* and can be thought of as the family tree relating the individuals alive at time t. It is also referred to as the *reduced family tree.* Note that N_t is also a Markov process.

Suppose $\boldsymbol{E}\xi = 1 + \alpha/t$. In Section 2 we will show that when t is large and the time units are taken as t generations, the reduced process can be approximated by a linear pure birth process $\{N(r), 0 \leqq r < 1\}$ with jump rate $b(\alpha, r)N(r)$ at time r, where

$$b(\alpha, r) = \begin{cases} \alpha(1 - e^{-\alpha})^{-1}(1 - r)^{-1}, & \alpha \neq 0, \\ (1 - r)^{-1} & \alpha = 0. \end{cases} \tag{1}$$

This generalises a result due to Fleischman and Siegmund-Schultze [11] and Durrett [5], which dealt specifically with the critical case, $\alpha = 0$.

In the supercritical case, when the process is not 'close' to critical, individuals are typically distantly related. For example, it follows from results of Bühler [2], Zubkov [44] and Durrett [5], that if $0 \leqq D_t \leqq t$ denotes the time of death of the most recent

Received 7 June 1993; revision received 17 January 1994.

* Postal address: DIAS, 10 Burlington Road, Dublin 4, Ireland.

This work constitutes part of the author's Ph.D. thesis, completed at the University of California at Berkeley, and was supported in part by NSF grants MCS90-01710 and DMS91-58583.

common ancestor of two randomly chosen individuals alive at time t, and $E\xi > 1$, then $(D_t/t \mid Z(t) > 0) \to 0$ in probability as $t \to \infty$. We will see later that this fact can be extrapolated (in some sense) from our result by letting $\alpha \to \infty$.

In the subcritical case ($E\xi < 1$) individuals typically have very recent ancestors: in this case, $(D_t/t \mid Z(t) > 0) \to 1$ in probability (see, for example, [5], [44]). Again this can be seen from our result by letting $\alpha \to -\infty$.

Essentially what we are doing here is describing the continuum of possibilities in between, which arise when the process is close to critical.

We will also introduce analogues of Yaglom's theorem and the exponential limit law for near-critical branching processes. These are well-known facts about critical branching processes and can be summarised as follows. If Z is a critical Markov branching process with mean lifetime 1 and offspring variance $\sigma^2 > 0$, then as $t \to \infty$, writing P^1 for the law of the process started with one individual,

$$tP^1(Z(t) > 0) \to 2/\sigma^2, \tag{2}$$

and

$$P^1\left\{\exp\left(-\lambda \frac{2Z(t)}{\sigma^2 t}\right) \,\middle|\, Z(t) > 0\right\} \to \frac{1}{1+\lambda}, \tag{3}$$

for $\lambda > 0$.

In Section 3 we will apply our results to the problem of estimating the age of our most recent common ancestor, more affectionately known these days as 'Eve'. (Assuming of course such an ancestor exists! Note the common ancestor of all humans need not be a human: it could be more like a chimpanzee, or even a fish. To question the existence of such an ancestor is simply beyond the scope of this paper. We remark however that, although nothing is certain, it would be remarkable and peculiar to the accepted theory of evolution if we did not have a common ancestor. A simple alternative would lead one logically to one of two possibilities: either humans and chimpanzees are not related at all; or there exist two humans and a chimpanzee such that the first human shares a common ancestor with the chimpanzee, but not with the other human. The latter possibility is considered implausible, and so repeating the argument we are reduced to the conclusion that no two species are related, contrary to the theory of evolution.) If T denotes the age of Eve and we model the female population as a Markov branching process with offspring mean $1 + \alpha/T$ ($\alpha \in \mathbb{R}$), then we can simultaneously estimate α and T using mitochondrial DNA data from a sample of contemporary human beings; assuming that the evolution of mitochondrial DNA is neutral, its inheritance is maternal, and that the mutation rate is constant and known. Most of the molecular studies so far (see, for example, [13], [35], [40], [41], [43]) have argued for a recent African genesis (as recent as 200 000 years ago), a theory that some paleontologists feel to be inconsistent with the fossil evidence (see, for example, [39]). The molecular evidence is generally presented in the form of a reconstructed tree relating a collection of

aligned mitochondrial DNA sequences, taken from a sample of contemporary humans, from which conclusions are drawn. We will argue that it is not necessary to construct a tree in order to make inferential statements about the *age* of Eve, and in Section 3.2 present an alternative approach, where the genealogy is modelled via a branching process. Similar arguments about Eve, also avoiding tree constructions, can be found in [19] and [17]. In Section 3.3 we evaluate the performance of our estimators, and finally, in Section 3.4, apply our methods to some data.

We conclude the paper with suggestions for further research in this area, in particular those motivated by the Eve problem. For example, it would be desirable to have a branching process analogue of Ewen's sampling formula. This is a formula giving the joint distribution of the numbers of distinct genetic types in a finite sample taken from a population that has been evolving according to the infinite-alleles Wright–Fisher model with neutral mutations. Such a formula would perhaps allow one to simultaneously estimate α, T *and* the mutation rate from the data; although it is open to question whether there is sufficient information contained in the data to do this effectively. Some progress in this direction has been made by Taib [36].

2. Main results

For each $t \geqq 0$, let $Z_t(s)$ be a Markov branching process with mean lifetime 1 and offspring distribution ξ_t, with $\boldsymbol{E}\xi_t = 1 + \alpha/t + o(1/t)$ and $\text{var}\,(\xi_t) = \sigma^2 + o(1/t) < \infty$, where $\alpha \in \mathbb{R}\backslash\{0\}$. (We assume $\alpha \neq 0$ for notational convenience only—the corresponding results for the critical case can be extrapolated by letting $\alpha \to 0$.) We will be considering the genealogy of this process for fixed α and large t: for this reason we refer to it as the *near-critical* case. A good general reference on near-critical branching processes is the book of Jagers [18], pp. 63–70, 199–206. We denote by $\boldsymbol{P}^x \equiv \boldsymbol{P}_t^x$ the law of the process Z_t started at x, suppressing the subscript for notational convenience, and write $\boldsymbol{E}^x$ for the corresponding expectations. Set $f_t(s) = \boldsymbol{E}^1 s^{Z_t(1)}$. It is important to note (see, for example, [12]) that the embedded (discrete-time) process $\{Z_t(n), n \in \mathbb{Z}_+\}$ is a Galton–Watson process with offspring mean $1 + \alpha/t + o(1/t)$, variance $\sigma^2 + o(1/t)$, and generating function $f_t(s)$. For $r > 0$, set $p_{x,r,t} = \boldsymbol{P}^x(Z_t(rt) > 0)$. We will assume throughout this section that $(Z_t^2(1) \mid Z_t(0) = 1)$ is uniformly integrable in t.

Our first result describes the rate at which $p_{x,r,t} \to 0$ when $t \to \infty$, and our second is an 'exponential limit law' for near-critical Markov branching processes. We combine them in one theorem. The corresponding well-known results for critical processes (see, for example, [1], pp. 19–20) can be extrapolated by letting $\alpha \to 0$.

Theorem 2.1

(i) *As* $t \to \infty$, $p_{x,r,t} \sim a_r x/t$, *where*

$$a_r = \frac{2\alpha}{\sigma^2}(1 - e^{-\alpha r})^{-1}.$$

(ii) *If* $Z_t(0)/t \Rightarrow 0$ *as* $t \to \infty$, *then for* $\lambda > 0$, $x \in \mathbb{Z}_+ \setminus \{0\}$,

$$\boldsymbol{E}^x(\exp\{-\lambda Z_t(rt)/t\} \mid Z_t(rt) > 0) \to \frac{b_r}{b_r + \lambda},$$

where $b_r = e^{-\alpha r} a_r$. *The limit law is exponential with parameter* b_r.

Proof. By a well-known theorem of Feller [9], Jiřina [20], and Lindvall [23], the process $\{t^{-1} Z_t([rt]), r \geqq 0\}$ converges in law, as $t \to \infty$, to a diffusion W with generator

$$\tfrac{1}{2}\sigma^2 x \frac{d^2}{dx^2} + \alpha x \frac{d}{dx}, \tag{4}$$

provided $Z_t(0)/t \Rightarrow W(0)$. Denote by $\mathbb{R}^w$ the law of the process W started at w. Thus, since Z_t is branching,

$$\begin{aligned} p_{x,r,t} &= 1 - \boldsymbol{P}^x(Z_t(rt) = 0) \\ &= 1 - \boldsymbol{P}^1(Z_t(rt) = 0)^{x/t} \\ &\sim 1 - \mathbb{R}^1(W(r) = 0)^{x/t}. \end{aligned}$$

Now suppose $(X, \mathbb{Q}^x)$ is a BESQ$^0(x)$ process; that is, a diffusion on $\mathbb{R}_+$ with generator $2x d^2/dx^2$, started at x. Using the space-time transformation of Pitman and Yor [30], Example 6.1, the $\mathbb{R}^w$-law of the process W is the same as the $\mathbb{Q}^{4w/\sigma^2}$-law of the process

$$\left\{\frac{\sigma^2 e^{\alpha r}}{4} X\left(\frac{1 - e^{-\alpha r}}{\alpha}\right), r \geqq 0\right\}.$$

Thus (see, for example, [21], p. 100),

$$\begin{aligned} \mathbb{R}^1(W(r) = 0) &= \mathbb{Q}^{4/\sigma^2}\left\{X\left(\frac{1 - e^{-\alpha r}}{\alpha}\right) = 0\right\} \\ &= \exp\{-a_r\}, \end{aligned}$$

and (i) follows.

To prove (ii), we apply the diffusion approximation and space-time transformation once more:

$$\begin{aligned} \boldsymbol{E}^x \exp\{-\lambda Z_t(rt)/t\} &= [\boldsymbol{E}^1 \exp\{-\lambda Z_t(rt)/t\}]^{x/t} \\ &\simeq [\mathbb{R}^1 \exp\{-\lambda W(r)\}]^{x/t} \\ &= \left\{\mathbb{Q}^{4/\sigma^2} \exp\left[-\lambda \frac{\sigma^2 e^{r\alpha}}{4} X\left(\frac{1 - e^{-r\alpha}}{\alpha}\right)\right]\right\}^{x/t} \\ &= \exp\left[-\frac{x}{t} \frac{\lambda a_r}{b_r + \lambda}\right]. \end{aligned}$$

Combining this with (i) we see that as $t \to \infty$,

$$\boldsymbol{E}^x(\exp\{-\lambda Z_t(rt)/t\} \mid Z_t(rt) > 0) \to \frac{b_r}{b_r + \lambda},$$

which we recognise as the Laplace transform of an exponential distribution with parameter b_r. This completes the proof of the theorem.

For each t and $0 \leqq s < t$, define (the reduced process) $N_t(s)$ to be the number of individuals alive at time s (in the process Z_t) having descendants alive at time t. Note that for each t, N_t is a time-inhomogeneous Markov branching process. Our main result is the following. In the statement of the theorem, $D_{\mathbb{Z}_+}[0, 1)$ denotes the space of càdlàg paths in $\mathbb{Z}_+$, parametrised by the unit interval; the reader should note that weak convergence in this case requires only convergence of finite-dimensional distributions.

Theorem 2.2. *As* $t \to \infty$, *the sequence of processes* $\{N_t(rt), 0 \leqq r < 1\}$ *converges in distribution in* $D_{\mathbb{Z}_+}[0, 1)$ *to a linear pure birth process* $\{N(r), 0 \leqq r < 1\}$ *with jump rate* $b(\alpha, r)N(r)$ *at time* r, *where*

$$b(\alpha, r) = \alpha(1 - e^{-\alpha})^{-1}(1 - r)^{-1},$$

provided $N_t(0) \Rightarrow N(0)$.

Proof. By Theorem 2.1, as $t \to \infty$,

$$\begin{aligned} \boldsymbol{P}^x(N_t(0) = 1 \mid Z_t(t) > 0) &= \boldsymbol{P}^x(N_t(0) = 1)p_{x,1,t}^{-1} \\ &= xp_{1,1,t}(1 - p_{1,1,t})^{-1}p_{x,1,t}^{x-1} \\ &\to 1. \end{aligned}$$

Applying Theorem 2.1 again we see that as $t \to \infty$,

$$\begin{aligned} \boldsymbol{P}^x(N_t(rt) = k \mid N_t(0) = 1) &= \boldsymbol{P}^x(N_t(rt) = k \mid Z_t(t) > 0) \\ &= p_{x,1,t}^{-1} \sum_{j=k}^{\infty} \boldsymbol{P}^x(Z_t(rt) = j)\binom{j}{k}p_{1-r,t}^k(1 - p_{1-r,t})^{j-k} \\ &\simeq p_{x,1,t}^{-1}p_{x,r,t}\int_0^{\infty} b_r e^{-b_r u}(k!)^{-1}u^k a_{1-r}^k e^{-a_{1-r}u}\, du \\ &\to \frac{a_r(-a_{1-r})^k}{k!\, a_1} G_r^k(a_{1-r}) \\ &=: p_r(k), \end{aligned}$$

for any $x \in \mathbb{Z}_+ \setminus \{0\}$, where G_r is the Laplace transform of the exponential distribution with rate b_r, and G_r^k denotes the kth derivative of G_r. It is easy to check that for $k \geqq 1$,

$$p_r(k) = q_r(1 - q_r)^{k-1},$$

where

$$q_r = \frac{e^{-\alpha r} - e^{-\alpha}}{1 - e^{-\alpha}}.$$

This shows that for each $0 \leqq r < 1$, $(N_t(rt) \mid N_t(0) = 1)$ converges in distribution to a random variable $N^1(r)$ with $\boldsymbol{P}(N^1(r) = k) = p_r(k)$. (The random variable $N^1(r) - 1$ has a geometric distribution with success probability q_r.) It follows from scaling that for $0 \leqq r_1 < r_2 < 1$,

$$(N_{(1-r_1)t}((r_2 - r_1)t) \mid N_{(1-r_1)t}(0) = 1) \Rightarrow N^1\left(\frac{r_2 - r_1}{1 - r_1}\right).$$

Therefore, since $(N_t(r_2,t) \mid N_t(r_1 t) = k)$ is just the sum of k independent copies of

$$(N_{(1-r_1)t}((r_2 - r_1)t) \mid N_{(1-r_1)t}(0) = 1),$$

it must converge in distribution to the sum of k independent copies of $N^1[(r_2 - r_1)/(1 - r_1)]$. This shows that the finite-dimensional distributions of the process $\{N_t(rt), 0 \leqq r < 1\}$ converge. Since each sample path is monotone increasing and for each r, the collection $\{N_t(rt), t \geqq 0\}$ is tight, it follows that $\{N_t(rt), 0 \leqq r < 1\}$ converges weakly in $D_{\mathbb{Z}_+}[0, 1)$ to a Markov process $\{N(r), 0 \leqq r < 1\}$, whose transition probabilities are determined by

$$(N(r_2) \mid N(r_1) = k) \stackrel{\mathrm{d}}{=} \sum_{i=1}^{k} N^i\left(\frac{r_2 - r_1}{1 - r_1}\right),$$

where for each r, $N^i(r)$ are independent and identically distributed with

$$\boldsymbol{P}(N^1(r) = k) = p_r(k).$$

For $k \geqq 3$, it can be easily checked that $r^{-1} p_r(k) \to 0$ as $r \to 0$, so the limit process N almost surely has no jumps of size $\geqq 2$. Since N must inherit the branching property, it follows that N is a linear pure birth process with jump rate $b(\alpha, r)N(r)$ at time r, where

$$\begin{aligned} b(\alpha, r) &= \lim_{h \to 0} h^{-1} P(N(r + h) = 2 \mid N(r) = 1) \\ &= \lim_{h \to 0} h^{-1} p_{h/(1-r)}(2) \\ &= \alpha(1 - e^{-\alpha})^{-1}(1 - r)^{-1}, \end{aligned}$$

as required.

Remark. An alternative description of the limiting process in Theorem 2.2 is via the following construction. Let η be a random variable taking values in $[0, 1)$ with

$$\boldsymbol{P}(\eta > r) = \frac{e^{-r\alpha} - e^{-\alpha}}{1 - e^{-\alpha}}. \tag{5}$$

The process evolves as follows. Start with $N(0)$ individuals. Each individual alive at time $0 \leqq r < 1$ lives for a time equal in distribution to $(1-r)\eta$ and is then replaced by two individuals. (Individuals live and die independently of one another.) Let $N(r)$ be the number of individuals alive at time r.

This reveals a kind of *stickbreaking* structure. In particular, consider the sequences of branch times along a single line of descent $(T_1, T_2, T_3, \cdots)$. If $\{\eta_i\}$ is a collection of independent copies of η, then the sequence

$$\left(\eta_1, \frac{\eta_2}{1-\eta_1}, \frac{\eta_3}{1-\eta_2-\eta_1}, \cdots\right) \tag{6}$$

has the same law as $(T_1, T_2, T_3, \cdots)$. The sequence (6) is an example of a *stickbreaking scheme.* Stickbreaking schemes arise naturally in a variety of settings, and have in recent times become a topic of much interest. For example, Pitman and Yor [31] and Perman et al. [27] have classified schemes that arise when the inter-jump times of a stable subordinator on the unit interval are sampled without replacement, with a bias according to their size (this is called *size-biased sampling*). Similar schemes can be used to approximate the distribution of the respective proportions of distinct alleles in a large population, ordered according to the ages of the alleles, when the population evolves according to the infinite-alleles Wright–Fisher model for neutral evolution (see, for example, [4], [15] and references therein).

Theorem 2.2 can be applied to help answer various genealogical questions. For example, the next result describes the degree of relationship of two randomly chosen individuals at time t. Let D_t be the time of death of the most recent common ancestor of the two individuals (we adopt the convention $D_t = 0$ if they have no common ancestor).

Theorem 2.3. *For* $0 \leqq r < 1$, $x \in \mathbb{Z}_+ \setminus \{0\}$,

$$\lim_{t\to\infty} \boldsymbol{P}(D_t > rt \mid N_t(0) = x) = \frac{2q_r^x}{(x-1)!}\{-(1-q_r)^{-x} - F(x-1, 1-q_r)\},$$

where

$$q_r = \frac{e^{-r\alpha} - e^{-\alpha}}{1-e^{-\alpha}},$$

and $F: \mathbb{Z}_+ \times (0,1) \to \mathbb{R}$ *is defined by*

$$F(n, y) = \frac{\partial^n}{\partial y^n}\left\{\frac{\log(1-y)}{y^2}\right\}.$$

Proof. We imitate the proof of the corresponding result in [5] for the critical case. Let $\boldsymbol{P}_{t,s,k}$ denote the probability that two individuals chosen randomly at time t have

the same ancestor at time s, given $N_t(s) = k$. Let $X_1(s, t), \cdots, X_k(s, t)$ be independent and identically distributed random variables with the same distribution as $(Z_t(s) \mid Z_t(s) > 0)$. If we let

$$S_k(s, t) = X_1(s, t) + \cdots + X_k(s, t),$$

then

$$\boldsymbol{P}_{t,s,k} = k\boldsymbol{E}[X_1(s, t)/S_k(s, t)]^2.$$

By Theorem 2.1(ii), for each i and $0 \leqq r < 1$, $X_i(rt, t)$ converges in distribution, as $t \to \infty$, to an exponentially distributed random variable with mean b_r^{-1}, which we denote by $X_i(r)$. So by bounded convergence we have

$$\boldsymbol{P}_{t,rt,k} \to k\boldsymbol{E}[X_1(r)/S_k(r)]^2,$$

as $t \to \infty$, where

$$S_k(r) = X_1(r) + \cdots + X_k(r).$$

But the random variable $kX_1(r)/S_k(r)$ has (by definition) an F-distribution with 1 and k degrees of freedom (this does not depend on the value of $\boldsymbol{E}X_1(r)$). In particular,

$$\boldsymbol{E}[kX_1(r)/S_k(r)]^2 = \frac{2k}{k+1}.$$

Combining this with Theorem 2.2 we have as $t \to \infty$,

$$\begin{aligned}\boldsymbol{P}(D_t > rt \mid N_t(0) = x) &= \sum_{k=1}^{\infty} \boldsymbol{P}_{t,rt,k}\boldsymbol{P}(N_t(rt) = k \mid N_t(0) = x) \\ &\to \sum_{k=1}^{\infty} \frac{2}{k+1}\boldsymbol{P}(N(r) = k \mid N(0) = x) \\ &\to \sum_{k=1}^{\infty} \frac{2}{k+1}\binom{k-1}{x-1}q_r^x(1-q_r)^{k-x},\end{aligned}$$

and the result follows.

Remarks. 1. The limiting process in Theorem 2.2 can be represented as a deterministic time change of a (time-homogeneous) Yule process. If $\{Y(t), t \geqq 0\}$ is a Yule process with branching rate 1, then the process

$$\left\{Y\left(\log\left[\frac{1-e^{-\alpha}}{e^{-r\alpha}-e^{-\alpha}}\right]\right), 0 \leqq r < 1\right\}$$

has the same law as N.

2. The corresponding results for the critical case [5], [11] can be extrapolated from Theorems 2.2 and 2.3 by letting $\alpha \to 0$. In this case the jump rate of the limiting pure linear birth process is $N(r)/(1-r)$ at time $0 \leqq r < 1$, and the stickbreaking distribution η is uniform on $[0, 1)$.

3. For $\theta > 0$, the process $\{N_{\theta t}(rt), 0 \leqq r < \theta\}$ converges weakly in $D_{\mathbb{Z}_+}[0, \theta)$ to an inhomogeneous Yule process $\{N^\theta(r), 0 \leqq r < \theta\}$ with branching rate $b(\alpha\theta, r\theta^{-1})$ $N^\theta(r)$ at time r. This follows from Theorem 2.2 by scaling. If $\alpha > 0$, then as $\theta \to \infty$, this limit becomes a Yule process, consistent with our earlier result [25] for the skeleton of a branching process (where essentially we let $\theta \to \infty$ before taking the diffusion limit as $t \to \infty$; here the limits are reversed).

4. In the 'strictly' supercritical case, where $\boldsymbol{E}\xi_t = m > 1$ for all t, individuals are distantly related in the following sense: if D_t is defined as in Theorem 2.3, then $(D_t/t \mid Z_t(t) > 0) \to 0$ in probability, as $t \to \infty$ (see, for example, [2], [5]). This fact can be extrapolated from Theorem 2.3, by letting $\alpha \to +\infty$. Intuitively, we can also see this from Theorem 2.2, noting that the stickbreaking distribution η converges to a point mass at zero as $\alpha \to +\infty$.

In the strictly subcritical case, where $\boldsymbol{E}\xi_t = m < 1$ for all t, individuals typically have very recent common ancestors: in this case $(D_t/t \mid Z_t(t) > 0) \to 1$ in probability, as $t \to \infty$ (see, for example, [44]). This can also be extrapolated from Theorem 2.3, by letting $\alpha \to -\infty$.

5. Similar and related results for general branching processes can be found in [32], [36], [44], for branching diffusion processes in [5], [33], and for superprocesses in [3], [6]. For an excellent review of the vast literature on genealogical processes in population genetics models, see [38].

6. Theorem 2.1 actually holds more generally for Bellman–Harris (age-dependent) branching processes (with mean lifetime 1 and bounded third offspring moment). This can be seen in the proof: it relies on the Feller branching diffusion approximation, and this was extended to the age-dependent case by Jagers [16]. Presumably, Theorems 2.2 and 2.3 can also be generalised to include the age-dependent case.

3. Application to mtEve

In recent times there has been much interest and controversy regarding the age and whereabouts of our most recent common female ancestor, Eve. Wilson and Cann [43] and Vigilant et al. [40], [41] claim that she probably lived in Africa about 200 000 years ago; Hasegawa and Horai [13] estimate the age to be 280 000 years. More recently, Stoneking et al. [35] published an estimate of 135 000 years. These estimates are based on mtDNA (mitochondrial DNA) data collected from contemporary humans. On the other hand, a number of paleontologists (see, for example, [39]) consider this theory to be inconsistent with the fossil evidence; they claim that if indeed we do have a common ancestor, then she must have walked the Earth at

least a million years ago. Popular accounts of the Eve controversy can be found in [39], [43] and [22].

We will only concern ourselves here with the age, not the whereabouts, of Eve.

3.1. *The molecular evidence for a recent African genesis.* We will explain here briefly how the argument for a recent African genesis is presented in [40]. The methods used in [13], [35], [41], [43] are similar.

The data used is a collection of aligned nucleotide sequences, each approximately 600 base pairs (sites) in length, sampled from the hypervariable segment in the control region of the human mitochondrial genome, of 189 individuals from around the world. There are four nucleotides: adenine, thymine, guanine and cytosine. A typical sequence might be coded as follows:

TTCTTTCCATG GGGAAGCAGA · · · CCTAACCAGA.

It is assumed that these sequences are maternally inherited. This is more or less the case in reality, since the mitochondria are located in the cytoplasm of the cell (as opposed to the nucleus) and the sperm contributes very little cytoplasm to the fertilised egg. In fact, it has been shown experimentally that the number of mitochondria in the egg is of the order of 10^5, whereas the sperm contains only about 50 mitochondria. It is also assumed that mtDNA is neutral from the standpoint of natural selection. In other words, the specific makeup of your mtDNA does not influence your reproductive tendencies. The general feeling is that there are certain sequence patterns that are fatal, and these can be ignored since they disappear at once from the gene pool; but apart from that, neutrality is a reasonable assumption.

A *substitution* is said to occur if one of the nucleotides in the sequence is replaced by another, and the new sequence is inherited. According to the *molecular clock hypothesis* substitutions occur randomly along lineages at a constant rate, and rates along different lineages are the same. The *genetic distance,* or *divergence,* between two such sequences is defined to be the proportion of sites at which the sequences differ. Among humans this is typically less than 5% in the control region of mtDNA. Vigilant et al. [40] found the average divergence between the humans in their sample, and a sample chimpanzee, to be about 15%.

Given the data, a phylogenetic tree relating the individuals in the sample is constructed, either by maximum likelihood [13] or by *parsimony* [35], [40], [41], [43]; and rooted using the chimpanzee as an *outgroup.* Maximum likelihood is not possible with samples as large as the one of Vigilant et al. [40].

The parsimony method is essentially an algorithm designed to minimise the total number of mutations on the constructed tree. Once the tree relating the sampled individuals is constructed, it is rooted at the point that is closest in genetic distance to the outgroup, which in this case is a common chimpanzee. We would like to

point out that there are several arguments for, and against, the use of parsimony; the same is true for the outgroup method. It seems that the former is best suited to the case where a small, genetically homogeneous sample is considered, in which case the resulting tree is often quite similar to that obtained by maximum likelihood. The advantage of parsimony over maximum likelihood is its relative computability. Otherwise, it need not be (see, for example, [10]) very reliable. The outgroup method is best suited to the case where the outgroup is relatively close, in genetic distance, to the sample.

The mutation rate is then estimated by comparing human and chimpanzee mtDNA: apparently human–chimpanzee divergence occurred about 4 to 6 million years ago [14], although there are estimates suggesting that the divergence occurred as far back as 9 million years ago [34]. As we remarked earlier, the average divergence between the humans in the sample of [40] and the chimpanzee was observed to be 15·1%. However, as this is over a considerable amount of time, there is the possibility of multiple substitutions at some sites, in which case the number of observed differences is an underestimate for the number of substitutions that actually occurred. To correct for multiple substitutions, one could naively consider that substitutions at different sites are independent and equally likely, leading one to an unbiased estimate for the divergence rate as 16·2% per N_{hc} million years, where N_{hc} is the number of millions of years ago that human–chimpanzee divergence occurred; or equivalently, $16{\cdot}2/N_{hc}$% per million years. If one supposes that substitution rates along the segment are variable, then the corrected estimate will be higher. The effect of interdependence between substitution events at different sites is not absolute. Note that if the naive estimate is valid, then the correction factor is small, and can certainly be ignored over the post-Eve period. In other words, there is no need to correct the observed inter-human divergences.

Vigilant et al. [40] take a different approach. There are basically two kinds of substitutions that can occur: *transitions* and *transversions*. Vigilant et al. observe that on the constructed tree, the ratio of transitions to transversions is 15:1. Then, assuming that multiple transversions do not occur, this leads to a corrected estimate of the divergence rate of 11·5–17·3% per million years, given that human–chimpanzee divergence occurred between 4 and 6 million years ago.

We are therefore left with a dilemma: either the naive estimate is hopeless, or the constructed tree is wrong (or both!). There is clearly a need for more research in this area, especially since the divergence rate is a key parameter for estimating the age of Eve. Wakely [42] has made some progress in modelling substitution rate variability along the segment, and gives a summary of the recent work on this topic. Note that this entire discussion rests upon the estimated time of human–chimpanzee divergence, and the assumption that the rate of divergence has not changed since then.

Finally, Vigilant et al. [40] consider all pairs of individuals whose most recent common ancestor, according to the constructed tree, is at the root of the tree, and calculate the average divergence between such pairs. Apparently, Vigilant et al.

feel it is not necessary to correct these observed divergences for multiple substitutions. This average is then compared with their estimated rate of divergence, leading to an estimate for the age of Eve between 166 000 and 249 000 years ago. The naive estimate for the rate of divergence would lead to an estimate between 718 000 and 1 060 500 years ago, or as much as 1 650 000 years if human–chimpanzee divergence occurred 9 million years ago.

In this section we will demonstrate that it is not necessary to construct a tree in order to make inferential statements about the *age* of Eve. Instead we will model the genealogy via branching processes. For similar work in this vein, see [19], [17].

3.2. *An alternative approach to dating the ancestor.* It is thought that mtDNA is inherited primarily from the mother. This assumption allows us to restrict our attention to single-sex populations, and so we are not forced to make questionable assumptions about the mating behaviour of people. According to the molecular clock hypothesis, substitutions occur randomly along lines of descent at a constant rate. Neutrality is assumed; that is, the occurrence of substitutions along a particular line of descent is independent of the family tree structure and geographical location of individuals, and that substitutions along distinct lines occur independently of each other. The divergence rate is very small, so over the time period we are considering here (the post-Eve period) we can assume that each substitution produces a new type, that is, reverse substitutions do not occur. Thus, if the most recent common ancestor of two individuals died s million years ago, the number of differences between their mtDNA types will be approximately Poisson with mean $2us$, where u is the substitution rate (in units of number of substitutions per million years). Now suppose two individuals are sampled randomly from the current population, and δ denotes the rate of divergence (in units of percentage divergence per million years). Note that if l denotes the sequence length, then $\delta = 2u/l$. If we have a model for the genealogical structure of the population, then the expected amount of divergence between the mtDNA sequences of the two individuals will be equal to the expected time back to the common ancestor of the two individuals (under our model, in units of millions of years), multiplied by the divergence rate, δ.

We will assume that the (effective) female population size follows a Markov branching process Z_T with mean offspring $1 + \alpha/T$, where $T = T_a/\lambda$; T_a is the time to our most recent common ancestor, λ is the mean effective lifetime (or *generation time*) and $\alpha \in \mathbb{R}$ is our 'growth' parameter.

To get an indication of how fast the population might have been growing, suppose the estimate of 200 000 years were correct. Then a straightforward moment calculation based on this model with offspring variance $\sigma^2 = 2$, mean (effective) lifetime 25 years and current (effective) female population size 1 billion, yields the rough estimate $\hat{\alpha} = 13{\cdot}7$ (cf. Equation (7) below). (Note that although the above chosen parameter values seem somewhat arbitrary, the estimate $\hat{\alpha}$ is quite

insensitive to apparently large adjustments in these values, and remains in the 'slightly supercritical' framework for quite a wide range.)

If we start time at the death of Eve, then in the notation of Section 1, $N_T(0) = 2$. (Eve, by definition, had at least two daughters with descendants alive today, and Theorem 2.2 tells us that three such daughters is extremely unlikely: $N_T(0-) = 1$ and $N_T(0) \geqq 2$ together imply that $N_T(0) = 2$ with high probability when T is large.) Then $Z_T(T)$ is the current (effective) female population size.

Using our approximation results, we can simultaneously estimate α and T, based on the observations $Z_T(T)$ and the average pairwise divergence in a random sample of n contemporary individuals $\bar{d}_n$. We will assume for the moment that the divergence rate δ is known. Denote by λ the mean effective lifetime of an individual. By Theorem 2.1,

$$\boldsymbol{E}(Z_T(T) \mid N_T(0) = 2) \simeq \frac{\sigma^2 T_a}{\lambda\alpha}(e^\alpha - 1). \tag{7}$$

We will assume that the sample size is such that

$$\boldsymbol{P}(Z_T(T) \geqq n \mid N_T(0) = 2) \simeq 1. \tag{8}$$

To see that this is not such an unreasonable assumption, suppose $T = 200\,000$, $\alpha = 12$, $\sigma^2 = 2$ and $\lambda = 25$. Then, by Theorem 2.1,

$$\boldsymbol{P}(Z_T(T) \geqq n \mid N_T(0) = 2) \simeq \boldsymbol{P}(X + Y \geqq 4(10^{-9})n), \tag{9}$$

where X and Y are independent exponential random variables with means equal to 1. In this case, the assumption (8) is certainly valid for $n < 10^8$. So we also have, by Theorem 2.3,

$$\begin{aligned} \boldsymbol{E}(\bar{d}_n \mid N_T(0) = 2, Z_T(T) \geqq n) &\simeq \delta\lambda \boldsymbol{E}(T - D_T \mid N_T(0) = 2) \\ &= \delta T_a\left[1 - \int_0^1 \boldsymbol{P}(D_T > rT \mid N_T(0) = 2)\, dr\right] \\ &\simeq \delta T_a \gamma(\alpha), \end{aligned} \tag{10}$$

where

$$\gamma(\alpha) = 1 - 2\int_0^1 \frac{q_r}{(1-q_r)^3}[1 - q_r^2 + 2q_r \log q_r]\, dr, \tag{11}$$

and

$$q_r = \frac{e^{-r\alpha} - e^{-\alpha}}{1 - e^{-\alpha}}. \tag{12}$$

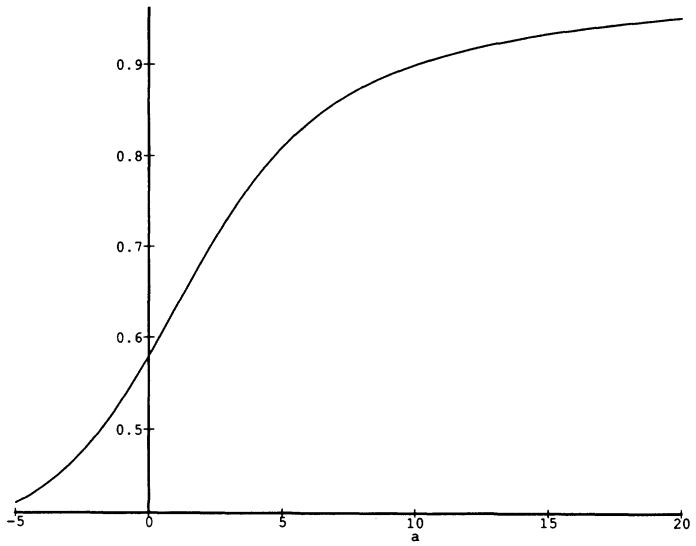

Figure 1. Plot of $\gamma(\alpha)$

One can simplify (11) to get

$$\gamma(\alpha) = 1 - 2\alpha^{-1}\int_0^1 \frac{u}{(1-u)^3(u+\kappa)}[1-u^2+2u\log u]\,du, \tag{13}$$

where

$$\kappa = \frac{e^{-\alpha}}{1-e^{-\alpha}}. \tag{14}$$

Note that $\gamma(\alpha)$ is positive and increasing in α, $\frac{1}{3} < \gamma(\alpha) < 1$, and $\gamma(\alpha) \uparrow 1$ as $\alpha \to \infty$. A plot of $\gamma(\alpha)$ is shown in Figure 1, for $-5 \leqq \alpha \leqq 15$.

For the simplest moment based estimates, assuming that δ, σ^2 and λ are known, just set

$$Z_T(T) = \frac{\sigma^2 \hat{T}_a}{\lambda\hat{\alpha}}(e^{\hat{\alpha}} - 1), \tag{15}$$

$$\hat{T}_a = \frac{\bar{d}_n}{\delta\gamma(\hat{\alpha})}; \tag{16}$$

and solve for $(\hat{\alpha}, \hat{T})$. Although σ^2 is unknown, when α is sufficiently large the actual value (within reason) will not affect the estimates considerably. (This is due to the dominating exponential term in Equation (15).) The same is true for λ.

Note that in theory this approach assumes that α is small relative to T. However, Remark (3) of Section 2 tells us that a large value of α corresponds to the

(significantly) supercritical case, and the estimate of T obtained from (16) will still make sense.

3.3. *Performance of the estimators.* Suppose we have a sample of n individuals, chosen at random from the current population: as before we will assume that

$$\boldsymbol{P}(Z_T(T) \geqq n \mid N_T(0) = 2) \simeq 1.$$

From now on, we will implicitly assume the conditioning $\{N_T(0) = 2,\ Z_T(T) \geqq n\}$. We begin by calculating the variance of the estimator

$$\hat{T}_a^n = \frac{\bar{d}_n}{\delta\gamma(\alpha)}, \tag{17}$$

where δ and α are assumed known. Denote by d_{ij} the divergence between individuals i and j, and by T_{ij} the time back to their most recent common ancestor. Then

$$\begin{aligned} \operatorname{var} \bar{d}_n &= \frac{1}{n^2(n-1)^2}\left\{2\sum_{i\neq j} \operatorname{var} d_{ij} + \sum_{|\{i,j,k,l\}|\geqq 3}^{i\neq j, k\neq l} \operatorname{cov}(d_{ij}, d_{kl})\right\} \\ &= a_n \operatorname{cov}(d_{12}, d_{34}) + b_n \operatorname{cov}(d_{12}, d_{23}) + c_n \operatorname{var} d_{12}, \end{aligned} \tag{18}$$

where

$$a_n = \frac{(n-2)(n-3)}{n(n-1)}, \qquad b_n = \frac{4(n-2)}{n(n-1)},$$

and

$$c_n = \frac{2}{n(n-1)}.$$

Note that $a_n + b_n + c_n = 1$. It follows immediately from (18) that for n sufficiently large,

$$\operatorname{var} \hat{T}_a^n \simeq \frac{\operatorname{cov}(d_{12}, d_{34})}{\delta^2\gamma(\alpha)^2}. \tag{19}$$

In particular, ignoring small order terms, (19) gives us a lower bound for $\operatorname{var} \hat{T}_a^n$. In other words, this is the best we can do if we base our inference solely on the average pairwise divergence in the sample. It is open to question how much room there is for improvement. In fact, (19) would give us an exact lower bound if

$$\operatorname{cov}(d_{12}, d_{23}) \geqq \operatorname{cov}(d_{12}, d_{34}), \tag{20}$$

since we also have by the correlation (or Cauchy–Schwartz) inequality that

$$\text{var } d_{12} \geqq \text{cov}\,(d_{12}, d_{34}). \tag{21}$$

Although the statement (20) might seem intuitively obvious at first sight, it is actually far from being clear, as we shall see later. A trivial upper bound for var $\bar{d}_n$ is var d_{12}, again by the correlation inequality.

When the parameters δ and α are not known, our estimate

$$\hat{T}_a^n = \frac{\bar{d}_n}{\hat{\delta}\gamma(\hat{\alpha})}, \tag{22}$$

is more variable. The major contributor to this extra variance is the variance due to our uncertainty about the rate of divergence δ: this has a multiplicative effect. On the other hand, in the range of possible parameter values we are dealing with in this problem, the variability of the other parameter estimates has only a small effect on the variance of $\hat{T}_a^n$.

We will now attempt to make the quantities that appear in (18) more explicit. To begin with,

$$\begin{aligned} \text{var } d_{12} &= \boldsymbol{E} \text{ var } (d_{12} \mid T_{12}) + \text{var } \boldsymbol{E}(d_{12} \mid T_{12}) \\ &= \delta \boldsymbol{E} T_{12} + \delta^2 \text{ var } T_{12} \\ &= \delta T_a \gamma(\alpha) + \delta^2 T_a^2 \rho(\alpha), \end{aligned} \tag{23}$$

where, in the notation of Sections 2 and 3.2,

$$\begin{aligned} \rho(\alpha) &= \text{var } (D_T/T) \\ &= 2\int_0^1 r\boldsymbol{P}(D_T > rT)\, dr - \left[\int_0^1 \boldsymbol{P}(D_T > rT)\, dr\right]^2 \\ &\simeq 4\int_0^1 \frac{rq_r}{(1-q_r)^3}[1 - q_r^2 + 2q_r \log q_r]\, dr - [1 - \gamma(\alpha)]^2. \end{aligned} \tag{24}$$

This gives us an exact upper bound:

$$\text{var } \hat{T}_a^n \leqq \frac{T_a}{\delta\gamma(\alpha)} + \frac{T_a^2 \rho(\alpha)}{\gamma(\alpha)^2}. \tag{25}$$

Note that $\rho(\alpha) \to 0$ as $\alpha \to \infty$, so that the upper bound tends to T_a/δ. But this is not sharp: we will see later that in fact var $\hat{T}_a^n \simeq 0$, when n is large and $\alpha \to \infty$.

To calculate cov (d_{12}, d_{23}), recall that

$$\boldsymbol{E} d_{12} = \delta \boldsymbol{E} T_{12} = \delta T_a \gamma(\alpha), \tag{26}$$

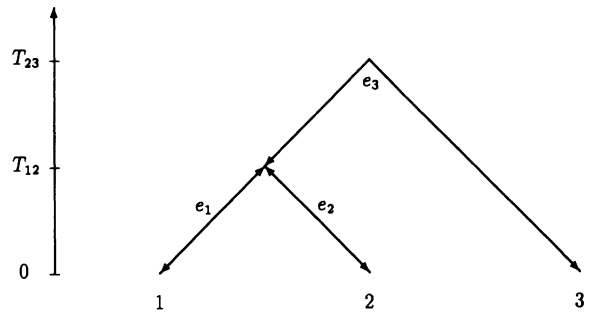

Figure 2. Typical genealogy relating three individuals. The components of the tree are labelled with their associated divergences e_1, e_2 and e_3

and set

$$e_1 = \tfrac{1}{2}(d_{12} + d_{13} - d_{23}), \tag{27}$$

$$e_2 = \tfrac{1}{2}(d_{12} + d_{23} - d_{13}), \tag{28}$$

$$e_3 = \tfrac{1}{2}(d_{13} + d_{23} - d_{12}). \tag{29}$$

Observe (see Figure 2) that e_1, e_2 and e_3 are conditionally independent given T_{12} and T_{23}. Now on $\{T_{12} < T_{23}\}$,

$$\begin{aligned} \boldsymbol{E}(d_{12}d_{23} \mid T_{12}, T_{23}) &= \boldsymbol{E}[(e_1 + e_2)(e_2 + e_3) \mid T_{12}, T_{23}] \\ &= \boldsymbol{E}(e_1 \mid T_{12}, T_{23})E(e_2 \mid T_{12}, T_{23}) + \boldsymbol{E}(e_2^2 \mid T_{12}, T_{23}) \\ &\quad + \boldsymbol{E}(e_1 \mid T_{12}, T_{23})E(e_3 \mid T_{12}, T_{23}) \\ &\quad + \boldsymbol{E}(e_2 \mid T_{12}, T_{23})\boldsymbol{E}(e_3 \mid T_{12}, T_{23}) \\ &= \frac{\delta^2}{4} T_{12}^2 + \frac{\delta^2}{4} T_{12}^2 + \frac{\delta}{2} T_{12} + 2\frac{\delta^2}{4} T_{12}^2(2T_{23} - T_{12}) \\ &= \tfrac{1}{2}\delta T_{12} + \delta^2 T_{12} T_{23}. \end{aligned} \tag{30}$$

Applying the obvious symmetry, and taking expections, we get

$$\boldsymbol{E}d_{12}d_{23} = \tfrac{1}{2}\delta \boldsymbol{E}(T_{12} \wedge T_{23}) + \sigma^2 \boldsymbol{E}T_{12}T_{23}. \tag{31}$$

Thus,

$$\operatorname{cov}(d_{12}, d_{23}) = \tfrac{1}{2}\delta \boldsymbol{E}(T_{12} \wedge T_{23}) + \delta^2 \boldsymbol{E}T_{12}T_{23} - \delta^2 T^2 \gamma(\alpha)^2. \tag{32}$$

To simplify the key quantity $\operatorname{cov}(d_{12}, d_{34})$ we need to introduce some notation.

The set $\{T_{ij}: i, j = 1, 2, 3, 4\}$ almost surely consists of exactly three elements, which we rank in ascending order and denote by $0 < \tau_1 < \tau_2 < \tau_3$. For each $\pi \in S_4$, the permutation group on four elements, define events

$$A_\pi = \{\tau_1 = T_{\pi(1),\pi(2)};\ \tau_2 = T_{\pi(1),\pi(3)};\ \tau_3 = T_{\pi(1),\pi(4)}\}, \tag{33}$$

$$B_\pi = \{\tau_1 = T_{\pi(1),\pi(2)};\ \tau_2 = T_{\pi(3),\pi(4)};\ \tau_3 = T_{\pi(1),\pi(3)}\}, \tag{34}$$

and set

$$A = \bigcup_{\pi \in S_4} A_\pi, \qquad B = \bigcup_{\pi \in S_4} B_\pi. \tag{35}$$

Note that $\boldsymbol{P}(A \cup B) = 1$, $\boldsymbol{P}(A \cap B) = 0$, and by symmetry,

$$\boldsymbol{P}(A_\pi \mid A) = \boldsymbol{P}(B_\pi \mid B) = \tfrac{1}{24}. \tag{36}$$

The events $\{A_\pi, B_\pi: \pi \in S_4\}$ represent the 48 possible tree topologies relating our four chosen individuals.

Write $\boldsymbol{E}_c$ for conditional expectations given $\sigma(\tau_1, \tau_2, \tau_3, A)$. Let

$$C = \{T_{12} \wedge T_{34} = \tau_1\}, \tag{37}$$

and observe that on C,

$$\boldsymbol{E}_c d_{12} d_{34} = \frac{\delta^2}{4} T_{12} T_{34}, \tag{38}$$

almost surely. On $A \cap C^c$, the total divergence on the tree can be written (see Figure 3) as the sum of five independent components $\{e_i\}_1^5$ such that

$$\boldsymbol{E}_c e_1 = \boldsymbol{E}_c e_2 = \tfrac{1}{2}\delta\tau_1 \tag{39}$$

$$\boldsymbol{E}_c e_3 = \tfrac{1}{2}\delta(\tau_2 - \tau_1) \tag{40}$$

$$\boldsymbol{E}_c e_4 = \tfrac{1}{2}\delta\tau_2 \tag{41}$$

$$\boldsymbol{E}_c e_5 = \tfrac{1}{2}\delta(2\tau_3 - \tau_2) \tag{42}$$

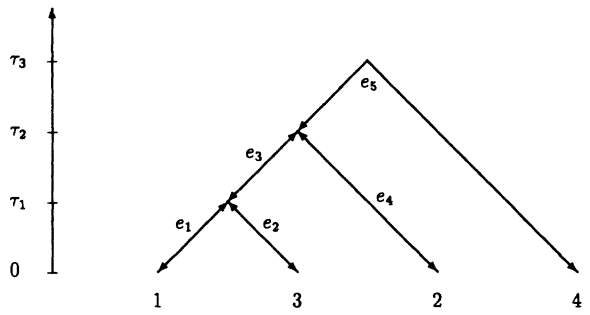

Figure 3. Typical genealogy on $A \cap C^c$ relating four individuals. The components of the tree are labelled with their associated divergences $e_1, \cdots, e_5$

almost surely, and

$$d_{12}d_{34} = (e_1 + e_3 + e_4)(e_2 + e_3 + e_5). \tag{43}$$

It follows that on $A \cap C^c$,

$$\begin{aligned} \boldsymbol{E}_c d_{12}d_{34} &= \boldsymbol{E}_c(e_1 + e_3 + e_4)(e_2 + e_3 + e_5) \\ &= \tfrac{1}{2}\delta(\tau_2 - \tau_1) + \tfrac{1}{2}\delta^2(\tau_2 - \tau_1)^2 \\ &\quad + \frac{\delta^2}{4}[2\tau_3(\tau_1 + \tau_2) + 2\tau_2(\tau_1 - \tau_2 + 2\tau_3)] \\ &= \tfrac{1}{2}\delta(\tau_2 - \tau_1) + \delta^2\tau_2\tau_3 \\ &= \tfrac{1}{2}\delta(T_{12} \wedge T_{34} - \tau_1) + \delta^2 T_{12}T_{34} \end{aligned} \tag{44}$$

almost surely. Similarly, on $B \cap C^c$,

$$\begin{aligned} \boldsymbol{E}_c d_{12}d_{34} &= \tfrac{1}{2}\delta(2\tau_3 - \tau_2 - \tau_1) + \delta^2\tau_3^2 \\ &= \tfrac{1}{2}\delta(T_{12} \wedge T_{34} - \tau_1) + \tfrac{1}{2}\delta(T_{12} \wedge T_{34} - \tau_2)^+ + \delta^2 T_{12}T_{34} \end{aligned} \tag{45}$$

almost surely. Combining (38), (44) and (45) we get

$$\boldsymbol{E}_c d_{12}d_{34} = \delta^2 \boldsymbol{E}_c T_{12}T_{34} + \tfrac{1}{2}\delta \boldsymbol{E}_c(T_{12} \wedge T_{34} - \tau_1) + \tfrac{1}{2}\delta \boldsymbol{E}_c(T_{12} \wedge T_{34} - \tau_2)^+, \tag{46}$$

almost surely. Taking expectations,

$$\boldsymbol{E} d_{12}d_{34} = \delta^2 \boldsymbol{E} T_{12}T_{34} + \tfrac{1}{2}\delta \boldsymbol{E}(T_{12} \wedge T_{34} - \tau_1) + \tfrac{1}{2}\delta \boldsymbol{E}(T_{12} \wedge T_{34} - \tau_2)^+, \tag{47}$$

and so

$$\begin{aligned} \operatorname{cov}(d_{12}, d_{34}) &= \delta^2 \boldsymbol{E} T_{12}T_{34} + \tfrac{1}{2}\delta \boldsymbol{E}(T_{12} \wedge T_{34} - \tau_1) \\ &\quad + \tfrac{1}{2}\delta \boldsymbol{E}(T_{12} \wedge T_{34} - \tau_2)^+ - \delta^2 T^2 \gamma(\alpha)^2. \end{aligned} \tag{48}$$

Unfortunately, this is as far as we have been able to go in computing $\operatorname{cov}(d_{12}, d_{34})$. In principle, it should be possible to calculate the unknowns in the expression (48) using Theorem 2.2, although I do not presently see how. It would also be helpful to know the distributions of the branch points (τ_i), and perhaps $\boldsymbol{P}(A)$. These are topics for future research.

However, we can argue that when α is large, $\operatorname{cov}(d_{12}, d_{34})$ is small. We observed earlier (cf. Equation (11)) that as $\alpha \to \infty$, $\boldsymbol{E}T_{12}/T_a \simeq \gamma(\alpha) \to 1$. This implies that for each i, j, T_{ij} converges in probability to T_a, so by (48) we have $\operatorname{cov}(d_{12}, d_{34}) \to 0$, as required. This means that when α and n are sufficiently large we can do well in estimating T, with the only significant source of variability being that of our estimate for the rate of divergence. For example, suppose $\alpha = 12$. Then $\boldsymbol{E}T_{12}/T_a \simeq \gamma(12) \simeq 0{\cdot}917$,

and we would expect var $\hat{T}_a^n$ to be relatively low, given a reasonably large sample and a good estimate for the rate of divergence.

Finally, we compare the expressions (32) and (48), and question our 'conjecture' (20). We can show, as one would expect, that $\boldsymbol{E}T_{12}T_{23} \geqq \boldsymbol{E}T_{12}T_{34}$. However, in attempting to establish (20), the difficulty arises that d_{12} and d_{34} pick up a high level of correlation on $B \cap C^c$ (cf. Equation (45)). To see this more clearly, note that by (36),

$$\boldsymbol{P}(C \mid A) = \boldsymbol{P}(C \mid B) = \tfrac{1}{3}, \tag{49}$$

and so

$$\boldsymbol{E}_c d_{12}d_{34} = \delta^2 \boldsymbol{E}_c T_{12}T_{34} + \tfrac{1}{3}\delta(\tau_2 - \tau_1)1_A + \tfrac{1}{3}\delta(2\tau_3 - \tau_1 - \tau_2)1_B, \tag{50}$$

almost surely. On the other hand, by (30), we can condition on the events $\{A_\pi, B_\pi : \pi \in S_4\}$ to get

$$\begin{aligned}\boldsymbol{E}_c d_{12}d_{23} &= \delta^2 \boldsymbol{E}_c T_{12}T_{23} + \tfrac{1}{2}\delta \boldsymbol{E}_c(T_{12} \wedge T_{23}) \\ &= \delta^2 \boldsymbol{E}_c T_{12}T_{34} + \tfrac{1}{2}\delta(\tfrac{1}{3}\tau_1 + \tfrac{5}{12}\tau_2 + \tfrac{1}{4}\tau_3)1_A + \tfrac{1}{6}\delta(\tau_1 + \tau_2 + \tau_3)1_B,\end{aligned}$$

almost surely. The problem is that on $B \cap \{\tau_3 < \tau_1 + \tau_2\}$,

$$\boldsymbol{E}_c d_{12}d_{23} - \delta^2 \boldsymbol{E}_c T_{12}T_{23} < \boldsymbol{E}_c d_{12}d_{34} - \delta^2 \boldsymbol{E}_c T_{12}T_{34}, \tag{51}$$

almost surely. One might surmise that there is something to be gained in the difference $\boldsymbol{E}_c T_{12}T_{23} - \boldsymbol{E}_c T_{12}T_{34}$, but this is not the case, as we shall see in the proof of the following 'consolation' lemma.

Lemma 3.1. $ET_{12}T_{23} \geqq ET_{12}T_{34}$.

Proof. In the above notation, we have by (49),

$$\boldsymbol{E}_c T_{12}T_{34} = (\tfrac{1}{3}\tau_1\tau_3 + \tfrac{2}{3}\tau_2\tau_3)1_A + (\tfrac{1}{3}\tau_1\tau_2 + \tfrac{2}{3}\tau_3^2)1_B,$$

almost surely. Similarly, by conditioning on the events $\{A_\pi, B_\pi : \pi \in S_4\}$, we get

$$\boldsymbol{E}_c T_{12}T_{23} = (\tfrac{1}{6}\tau_1\tau_2 + \tfrac{1}{6}\tau_1\tau_3 + \tfrac{1}{3}\tau_2\tau_3 + \tfrac{1}{12}\tau_2^2 + \tfrac{1}{4}\tau_3^2)1_A + (\tfrac{1}{3}\tau_1\tau_3 + \tfrac{2}{3}\tau_3^2)1_B,$$

almost surely. It follows that

$$\begin{aligned}\boldsymbol{E}_c T_{12}T_{23} - \boldsymbol{E}_c T_{12}T_{34} = (\tfrac{1}{6}\tau_1\tau_2 - \tfrac{1}{6}\tau_1\tau_3 - \tfrac{1}{3}\tau_2\tau_3 + \tfrac{1}{2}\tau_2^2 + \tfrac{1}{4}\tau_3^2)1_A \\ + \tfrac{1}{3}\tau_1(\tau_3 - \tau_2)1_B,\end{aligned}$$

almost surely. Clearly, the second term is almost surely non-negative, since $\tau_3 > \tau_2$ almost surely. To show likewise for the first term, consider the function

$$f(r, s, t) = \tfrac{1}{6}rs - \tfrac{1}{6}rt - \tfrac{1}{3}st + \tfrac{1}{2}s^2 + \tfrac{1}{4}t^2.$$

It suffices to show that f is non-negative on $0 \leqq r \leqq s \leqq t$. First we observe that $f(r, s, t) = 0$, if $s = t$. Moreover,

$$\frac{\partial}{\partial t} f(r, s, t) = \tfrac{1}{2}t - \tfrac{1}{6}r - \tfrac{1}{3}s \geqq 0,$$

on $0 \leqq r \leqq s \leqq t$, and so we are done.

3.4. *Some numbers at last.* We would now like to apply our method to some data: but where does one find a *random* sample of individuals? Strictly speaking this is simply not available, as yet. However, we will do our best with what we have.

Of the 189 individuals considered by Vigilant et al. [40], we have hand-picked a somewhat representative sub-sample of 19, without being deliberately biased in any way. The larger the sub-sample, the less representative it becomes; the smaller it is, the less useful it becomes. Our sample consists of 6 Asians, 1 Native Australian, 1 Papua New Guinean, 6 Europeans and 5 Africans.

A histogram of the 171 pairwise divergences in this sample is shown in Figure 4. The average divergence was found to be 2·8%.

In June 1992, according to the *Population Reference Bureau Estimates,* the human population size was approximately 5·412 billion. This gives us about 1 billion as a rough estimate for the current effective female population size, assuming that about half the population is female, and that the current female population represents

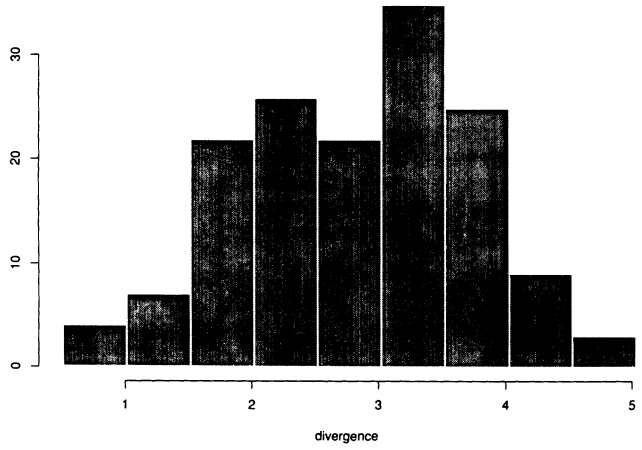

Figure 4. Pairwise divergences among sample of 19 individuals

TABLE 1
Estimates for α and T_a

$\lambda Z_T(T)/\sigma^2$	δ	$\hat{\alpha}$	$\hat{T}_a$
12·5 billion	1·8	11·33	1 706 103
	2·7	11·77	1 113 265
	4	12·21	762 437
5 billion	1·8	10·31	1 722 531
	2·7	10·76	1 143 245
	4	11·2	768 607
30 billion	1·8	12·29	1 693 286
	2·7	12·94	1 125 364
	4	13·17	757 516

approximately 2·7 generations. We will soon see that our estimates are quite insensitive to variations in this figure, so we need not be very exact.

Note that the estimates $\hat{\alpha}$ and $\hat{T}_a$ are functions of $\lambda Z_T(T)/\sigma^2$ and δ; these are shown in Table 1, for various different values of $\lambda Z_T(T)/\sigma^2$ and δ. If $Z_T(T) = 1$ billion, $\sigma^2 = 2$ and $\lambda = 25$, then $\lambda Z_T(T)/\sigma^2 = 12{\cdot}5$ billion. Although these choices seem somewhat arbitrary, we can see from Table 1 that any kind of realistic deviations from these values will have little or no effect on the estimates. The most important parameter is δ, the rate of divergence.

4. Concluding remarks

To derive our estimates for the growth rate, α, and the age of Eve, T_a, we simply calculated the expected current population size and the expected average pairwise divergence in a sample of contemporary individuals, and assumed the other parameters were known. We are therefore not fully utilising the information contained in the sample. It might be helpful to know more about the joint distribution of the pairwise divergences (d_{ij}), or the joint distribution of the respective frequencies of distinct types, in a finite sample. The latter would be analogous to *Ewens' sampling formula* for the infinite-alleles Wright–Fisher model for neutral evolution. Ewens' sampling formula is not applicable to the Eve problem because it is based on the assumption that the population size is constant over time. Although a sampling formula for branching processes is a desirable goal, it might be the case here that an explicit description for the distribution of the pairwise divergences would be more useful, for the following reason. Typically, the frequencies of each distinct type in a sample are very low, in which case such a sampling formula would not be very powerful. For our sample of 19, there are 19 distinct types! This could change however, with the availability of larger 'random' samples.

In particular, it may be possible to estimate α, T_a and δ simultaneously, without

having to rely on human–chimpanzee comparisons, thus avoiding the assumption that the rate of divergence has been constant ever since the human and chimpanzee lines diverged.

Taib [37] has made some progress in this direction by obtaining an expression for the asymptotic proportion of alleles (types) with exactly j representatives in the population, for a supercritical branching process with neutral mutations. Unfortunately, this result is not directly applicable here.

There are many directions one could take on this quest. An obvious starting point would be to say something useful about the relative proportions of distinct types in the current population using our model. There are results about random partitions and stickbreaking schemes that may be useful in this regard, in particular those due to Pitman [29], [28]. For example, if one could prove that the random partition of types has a stickbreaking structure, then according to a theorem of Pitman ([29], Theorem 2) the distribution of the random partition must belong to a simple two-parameter family, where explicit sampling formulas are known [28].

It is also convenient to approximate the process by a Dawson–Watanabe superprocess with branching mechanism $\phi(z) = \alpha z - \sigma^2 z^2/2$, conditioned to be currently non-extinct, where the spatial motion is a jump process on the space of all possible types (see, for example, [8], Example 10.4.4, for a description of this kind of spatial motion). Facts about the distribution of the random partition of types can then be described in terms of solutions to the corresponding partial differential equations.

An alternative approach would be to apply the results of Etheridge and March [7] and Perkins [26] on the relationship between Dawson–Watanabe superprocesses and (inhomogeneous) Fleming–Viot processes. Ewens' sampling formula is valid for the homogeneous Fleming–Viot process.

Finally, we remark that using the coalescent model (an approximation for the genealogy in the Wright–Fisher and Fleming–Viot models, where the population size is held constant) Lundstrom et al. [24] have derived simultaneous estimates for the branch times (and in particular, the time of common ancestry) and substitution rates along a nucleotide sequence, based on molecular data collected from a finite sample of contemporary individuals. This is the most generally applicable approach available up to now, but unfortunately relies on the assumption that the total population size has been roughly constant over time.

Acknowledgements

This work was carried out as part of my Ph.D. thesis at the University of California at Berkeley under the direction of Steve Evans, to whom I am indebted. Thanks also to Terry Speed who gave me a lot of help on this problem, to David Aldous, Alison Etheridge, Jim Pitman and Joachim Rebholz for many valuable discussions, and to the referee for helpful comments and suggestions.

References

[1] ATHREYA, K. AND NEY, P. (1972) *Branching Processes.* Springer-Verlag, New York.

[2] BÜHLER, W. (1972) The distribution of generations and other aspects of the family structure of branching processes. In *Proc. 6th Berkeley Symp. Math. Statist. Prob.* **3,** 463–480.

[3] DAWSON, D. AND PERKINS, E. (1991) Historical processes. *Mem. Amer. Math. Soc.* **93**(454).

[4] DONNELLY, P. (1986) Partition structures, Polya urns, the Ewens' sampling formula, and the ages of alleles. *Theoret. Popn. Biol.* **30,** 271–288.

[5] DURRETT, R. (1978) The genealogy of critical branching processes. *Stoch. Proc. Appl.* **8,** 101–116.

[6] ETHERIDGE, A. M. (1992) Conditioned superprocesses and a semilinear heat equation. Preprint.

[7] ETHERIDGE, A. M. AND MARCH, P. (1991) A note on superprocesses. *Prob. Theory Rel. Fields* **89,** 141–147.

[8] ETHIER, S. AND KURTZ, T. (1986) *Markov Processes: Characterisation and Convergence.* Wiley, New York.

[9] FELLER, W. (1951) Diffusion processes in genetics. In *Proc. 2nd Berkeley Symp. Math. Statist. Prob.,* 227–246.

[10] FELSENSTEIN, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27,** 401–410.

[11] FLEISCHMANN, K. AND SIEGMUND-SCHULTZE, R. (1977) The genealogy of critical branching processes. *Math. Nachr.* **79,** 233–241.

[12] HARRIS, T. E. (1963) *The Theory of Branching Processes.* Springer-Verlag, Berlin.

[13] HASEGAWA, M. AND HORAI, S. (1990) Time of the deepest root for polymorphism in human mitochondrial DNA. *J. Mol. Evol.* **32,** 37–42.

[14] HASEGAWA, M., KISHINO, H., HAYASAKA, K. AND HORAI, S. (1990) Mitochondrial DNA evolution in primates; transition rate has been extremely low in lemur. *J. Mol. Evol.* **31,** 113.

[15] HOPPE, F. M. (1986) Size-biased filtering of Poisson–Dirichlet samples with an application to partition structures in genetics. *J. Appl. Prob.* **23,** 1008–1012.

[16] JAGERS, P. (1971) Diffusion approximations for branching processes. *Ann. Math. Statist.* **42,** 2074–2078.

[17] JAGERS, P. (1991) The growth and stabilisation of populations. *Statist. Sci.* **6,** 269–283.

[18] JAGERS, P. (1985) *Branching Processes with Biological Applications.* Wiley, Chichester.

[19] JAGERS, P., NERMAN, O. AND TAIB, Z. (1991) When did Joe's great . . . grandfather live? Or on the timescale of evolution. In *Selected Proceedings of the Sheffield Symposium on Applied Probability,* ed. I. V. Basawa and R. L. Taylor. IMS Lecture Notes—Monograph Series **18.**

[20] JIŘINA, M. (1969) On Feller's branching diffusion process. *Časopis. Pěst. Mat.* **94,** 84–90.

[21] KNIGHT, F. (1981) Essentials of Brownian motion and diffusion. *Amer. Math. Soc., Math. Surveys* **218.**

[22] KRISHTALKA, L. (1989) *Dinosaur Plots and Other Intrigues in Natural History.* Avon Books, New York.

[23] LINDVALL, T. Limit theorems for some functionals of certain Galton–Watson branching processes. *Adv. Appl. Prob.* **6,** 309–321.

[24] LUNDSTROM, R., TAVARÉ, S. AND WARD, R. H. (1992) Estimating substitution rates from molecular data using the coalescent. *Proc. Natl. Acad. Sci. USA* **89,** 5961–5965.

[25] O'CONNELL, N. (1993) Yule process approximation for the skeleton of a branching processes. *J. Appl. Prob.* **30,** 725–729.

[26] PERKINS, E. A. (1992) Conditional Dawson–Watanabe processes and Fleming–Viot processes. Preprint.

[27] PERMAN, M., PITMAN, J. W. AND YOR, M. (1991) Size-biased sampling of Poisson point processes and excursions. *Prob. Theory Rel. Fields.*

[28] PITMAN, J. W. (1992) A two-parameter version of Ewens' sampling formula. Preprint.

[29] PITMAN, J. W. (1993) Random discrete distributions invariant under size-biased permutation. *J. Appl. Prob.* To appear.

[30] PITMAN, J. W. AND YOR, M. (1982) A decomposition of Bessel bridges. *Z. Wahrscheinlichkeitsth.* **59,** 425–457.

[31] PITMAN, J. W. AND YOR, M. (1990) Arcsine laws and interval partitions derived from a stable subordinator. Technical Report 189, University of California, Berkeley.

[32] SAGITOV, S. (1989) The limit behaviour of reduced critical branching processes. *Soviet Math. Dokl.* **38,** 488–491.

[33] SAWYER, S. (1976) Branching diffusion processes in population genetics. *Adv. Appl. Prob.* **8,** 659–689.

[34] SPUHLER, J. N. (1988) Evolution of mitochondrial DNA in monkeys, apes and humans. *Yearb. Phys. Anthropol.* **31,** 15–48.

[35] STONEKING, L., SHERRY, S. T., REDD, A. J. AND VIGILANT, L. (1992) New approaches to dating suggest a recent age for the human mtDNA ancestor. Preprint.

[36] TAIB, Z. Labelled Branching Processes with Applications to Neutral Evolution Theory. PhD thesis, Chalmers University of Technology, Sweden.

[37] TAIB, Z. (1991) The most frequent alleles in a branching process with neutral mutations. Preprint.

[38] TAVARÉ, S. (1984) Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoret. Popn Biol.* **26,** 119–164.

[39] THORNE, A. G. AND WOLPOFF, M. H. (1992) The multiregional evolution of humans. *Sci. Amer.* April, 76–83.

[40] VIGILANT, L. PENNINGTON, R., HARPENDING, H., KOCHER, T. D. AND WILSON, A. (1989) Mitochondrial DNA sequences in single hairs from a southern African population. *Proc. Natl. Acad. Sci. USA* **86,** 9350–9354.

[41] VIGILANT, L., STONEKING, L., HARPENDING, H., HAWKES, K. AND WILSON, A. (1991) African populations and the evolution of human mitochondrial DNA. *Science* **253,** 1503–1507.

[42] WAKELY, J. (1993) Substitute rate variation among sites in hypervariable region 1 of human mitochondrial DNA. Preprint.

[43] WILSON, A. C. AND CANN, R. L. (1992) Recent African genesis of humans. *Sci. Amer.* April, 68–73.

[44] ZUBKOV, A. (1975) Limiting distributions of the distance to the closest common ancestor. *Theory Prob. Appl.* **20,** 602–612.