J. Math. Biol. 46, 241–264 (2003)
Digital Object Identifier (DOI):
10.1007/s00285-002-0180-8

Mathematical Biology

Carsten Wiuf

Inferring population history from genealogical trees

Received: 18 January 2001 / Revised version: 4 September 2002 /
Published online: 18 December 2002 – © Springer-Verlag 2002

Abstract. Inference about population history from DNA sequence data has become increasingly popular. For human populations, questions about whether a population has been expanding and when expansion began are often the focus of attention. For viral populations, questions about the epidemiological history of a virus, e.g., HIV-1 and Hepatitis C, are often of interest. In this paper I address the following question: Can population history be accurately inferred from single locus DNA data? An idealised world is considered in which the tree relating a sample of n non-recombining and selectively neutral DNA sequences is observed, rather than just the sequences themselves. This approach provides an upper limit to the information that possibly can be extracted from a sample. It is shown, based on Kingman's (1982a) coalescent process, that consistent estimation of parameters describing population history (e.g., a growth rate) cannot be achieved for increasing sample size, n. This is worse than often found for estimators of genetic parameters, e.g., the mutation rate typically converges at rate $\sqrt{\log(n)}$ under the assumption that all historical mutations can be observed in the sample. In addition, various results for the distribution of maximum likelihood estimators are presented.

1. Introduction

Methods and software for the statistical analysis of DNA sequences sampled from a population are today fairly well developed. Many such methods rely on novel advances in computational statistics and require many iterations of a procedure to, e.g., evaluate an estimate of a parameter or a density (Stephens and Donnelly 2000). The reason for this is to be found in the underlying stochastic structure of the data: DNA sequences sampled from a population are not independent observations, but highly correlated observations due to common ancestry. The complicated stochastic structure represents not just an obstacle in the analysis of data but also in assessing statistical properties of estimators and test statistics, e.g., asymptotic properties for increasing sample size or for increasing length of the DNA sequences.

The probability of obtaining an observed sample configuration, $s = (s_1, s_2, \ldots, s_n)$, can be written in the form

$$P(S = s) = \int_g P(S = s \mid G_n = x) P(G_n \in dx), \tag{1}$$

C. Wiuf: Department of Statistics, University of Oxford, Oxford, OX1 3TG, UK.
e-mail: wiuf@stats.ox.ac.uk

Key words or phrases: Coalescent process – Genealogy – Population history – Maximum likelihood inference

where x denotes the genealogy (or tree) relating the sampled sequences, s, and G_n and $S = (S_1, S_2, \ldots, S_n)$ denote the corresponding random variables (Figure 1). The probability $P(S = s \mid G_n = x)$ depends on the rate of mutation, whereas $P(G_n \in dx)$ depends on the population size and the demographic history of the population. Here and elsewhere it is assumed that sequences are non-recombining and all sequence types are selectively neutral. Because the common ancestry (the, in general, unknown outcome of the random tree G_n) induces correlations among the observed variables (the variables $S_1, \ldots, S_n$ are in general exchangeable), standard or classical statistical theory cannot be applied to evaluate procedures for inference.

One way to deal with this problem has been to assume that more information is available than actually is: For example, in a discussion of various estimators of the mutation rate, Felsenstein (1992) assumed that the random variable G_n was observed, rather than the variables $S_1, \ldots, S_n$, and that G_n was scaled in the expected number of substitutions per generation (rather than in generations or in real time, Figure 1). The genealogy, x, of the observed sample can be estimated consistently for increasing sequence length (Chang 1996), and Felsenstein's (1992) approach corresponds as such to an ideal world in which DNA sequences are of infinite length. The genealogy x is often called the *true* genealogy or tree of the sampled sequences to indicate that a genealogical relationship estimated from the sample, s, is at best an estimate and not likely to equal x precisely. (Some authors prefer 'reconstruction of the true genealogy' instead of 'estimation of the true genealogy', as is used here. I find the latter more correct than the former.) The issue of inferring mutation rate has in general been widely discussed in the literature (Felsenstein 1992; Fu and Li 1993; Klein et al. 1999, among others). Felsenstein (1992) showed

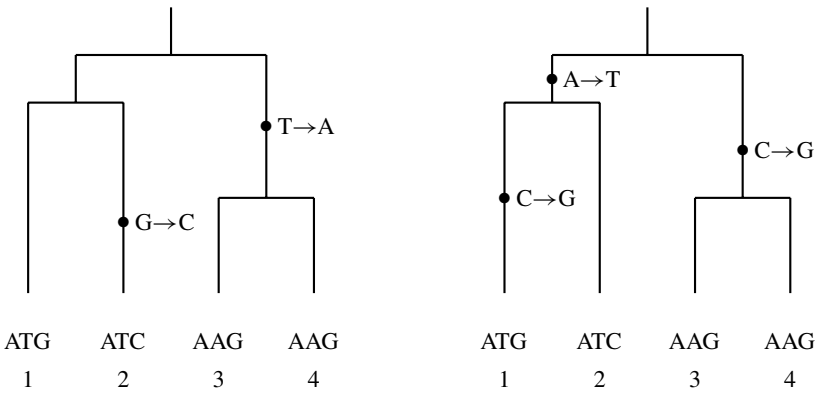

Fig. 1. Given an outcome (genealogy), x, of G_n, the sample configuration $s = (s_1, \ldots, s_n)$ can be obtained by mutations in many ways. Here, $n = 4$ and each sequence, labeled 1, 2, 3, and 4, has length three. In the first tree, the most recent common ancestor of the sample is of type ATG, in the second, of type AAC. Mutations are marked • and $y \to z$ means that z is substituted for y. Long sequences allow for accurate estimation of the genealogy x. Times between coalescence events are naturally measured in expected number of substitutions per generation.

that information about the tree allows a consistent estimator of the mutation rate that converges at rate $\sqrt{n}$ for increasing sample size. In contrast, sequence information alone allows a consistent estimator that converges at rate $\sqrt{\log(n)}$ (under an infinite-site assumption).

Demographic parameters are broadly of two kinds: 1) Parameters that describe variations in population size over time, and 2) Migration parameters that describe movements between populations. In equation (1), $P(S = s)$ depends on the mutation rate through the conditional probability $P(S = s \mid G_n = x)$, whereas $P(S = s)$ depends on demographic parameters only through $P(G_n \in dx)$. Thus, the accuracy in estimation of demographic parameters is not expected to exceed the accuracy in estimation of the mutation rate. In this paper, I focus on parameters that describe variations in population size of a large panmictic population. My approach is similar to Felsenstein's in that I assume the tree relating the sequences is observed, rather than the sequences. To be specific, I consider an observation of the random variable, $G_n^* = \psi_0 G_n$, where G_n follows Kingman's (1982a) coalescent and ψ_0 is a scaling factor that converts time in the coalescent to time in expected number of substitutions (ψ_0 is the mutation rate in Felsenstein 1992). The coalescent process is a robust approximation to the distribution of a genealogy sampled from a large panmictic population (Kingman 1982a, b, Donnelly and Tavaré 1995).

General asymptotic results for the convergence of the maximum likelihood estimators (mle) of ψ_0 and of demographic parameters are derived. In particular, a special case of the coalescent process, the coalescent with exponential growth (Slatkin and Hudson 1991), is studied in detail. This process has been of recent interest in the study of human genomic DNA sequences and of viral sequences.

2. Setting

The coalescent (Kingman 1982a, b) is adopted as a description of the genealogy of the population (or equivalently of samples from it) with time running from the present-day into the past. In the coalescent, time is measured such that one unit of time accounts for N generations in the real population of present effective population size N. The simplest case is a population of constant effective size. In this case, the distribution of the genealogy of a sample of size $n \geq 2$, is given by $n - 1$ independent exponential variables,

$$W_j \sim Exp\left(\frac{j(j-1)}{2}\right), \tag{2}$$

$2 \leq j \leq n$, where W_j denotes the time while there are j ancestral lineages.

If the population size is not constant but varies over time, Griffiths and Tavaré (1994) showed that the distribution of the genealogy could be obtained by a transformation of W_j, $2 \leq j \leq n$. Define the population intensity, $\lambda(t)$, $t \geq 0$, by

$$\lambda(t) = \lim_{N(0)\to\infty} \frac{N(\lfloor Nt \rfloor)}{N}, \tag{3}$$

(assuming the limit exists) where $N(\tau)$ denotes the population size at generation τ, $N(0) = N$ and $\lfloor x \rfloor$ denotes the integer part of x. Let T_{jn} be the time while

there are at least j ancestral lineages of the sample and let U_{jn} denote the similar variable under the assumption of constant population size, $U_{jn} = \sum_{i=j}^{n} W_i$. Then $T_n = (T_{2n}, T_{3n}, \ldots, T_{nn})$ defined by

$$U_{jn} = \int_0^{T_{jn}} \nu(t)\, dt, \tag{4}$$

$2 \leq j \leq n$, describes the genealogy of a sample of size n. The function $\nu(t)$ is the reciprocal of the population intensity, $\lambda(t) = \nu(t)^{-1}$. Note that, in general, the time between two coalescence events, $V_{jn} = T_{jn} - T_{j+1,n}$, depends on n; unlike W_j that is independent of n. The process T_n is a pure death process with time-dependent death rates, $j(j-1)\nu(t)/2$. The notation is briefly shown in Figure 2.

A number of regularity condition on ν is required. First, that ν is continuous and strictly positive for all $t > 0$. Second, that $\int_0^{\infty} \nu(t)dt = \infty$. These assumptions together with Equation (2) assure that multiple coalescence events cannot happen at the same time and that the entire population finds a most recent common ancestor (MRCA) in finite time. By definition of W_j, the variables $U_{j\infty} = \sum_{i=j}^{\infty} W_i$ and $T_{j\infty}$, defined by Equation (4) with $n = \infty$, are finite almost surely. (Here and elsewhere 'almost surely' is with respect to (wrt) the process U_n or U_∞.) Thus, the process U_n (and T_n) has a well-defined entrance boundary at $n = \infty$.

In the set-up of Felsenstein's (1992) $X_n = (X_{2n}, \ldots, X_{nn})$ is observed, rather than T_n, where $X_{jn} = \psi_0 T_{jn}$ and ψ_0 is an unknown scaling constant. Typically, ψ_0 is of the form $\psi_0 = N(0)u$, where u is the rate of mutation per generation.

In the following I consider the problem of estimating ν from an observation of T_n or of estimating ν and ψ_0 from an observation of X_n. All proofs are given in appendices.

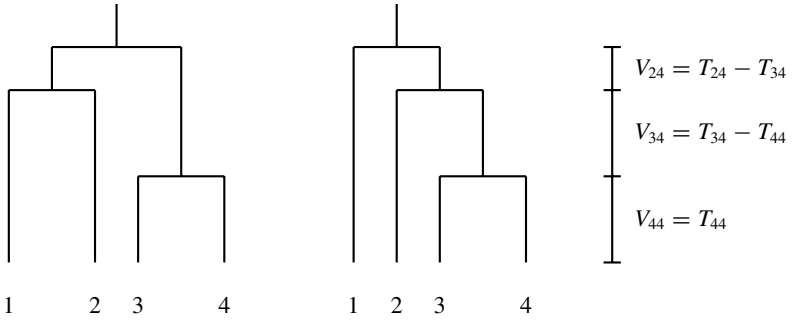

Fig. 2. Two examples of trees relating $n = 4$ sequences. A tree, G_n, is given by a topology (branching order) and a collection of $n-1$ variables, V_{jn}, $j = 2, \ldots, n$, denoting time between coalescence events. V_{jn} is the time while there are j ancestral lineages. Alternatively, the cumulative times, $T_{jn} = \sum_{i=j}^{n} V_{in}$, can be specified. All possible topologies are equally likely. In the examples, the two topologies have the same probability of occurring.

3. General results

The density (wrt Lebesgue measure) of $T_n = (T_{2n}, T_{3n}, \ldots, T_{nn})$ is given by

$$f_n(\underline{t}, \nu) = \frac{n!\,(n-1)!}{2^{n-1}} \prod_{j=2}^{n} \nu(t_j) \exp\left\{ -\sum_{j=2}^{n} \frac{j(j-1)}{2} \int_{t_{j+1}}^{t_j} \nu(t)dt \right\}$$

$$= \frac{n!\,(n-1)!}{2^{n-1}} \prod_{j=2}^{n} \nu(t_j) \exp\left\{ -\sum_{j=2}^{n} (j-1) \int_{0}^{t_j} \nu(t)dt \right\}, \qquad (5)$$

for $n \geq 2$, $\underline{t} = (t_2, t_3, \ldots, t_n)$, $t_2 > t_3 > \ldots > t_n > t_{n+1} = 0$. For convenience, the density is indexed by ν. In practice, ν will be a function depending on a vector of unknown parameters, $\underline{\alpha} = (\alpha_1, \ldots, \alpha_d)$, with $\underline{\alpha} \in A$. Let $f_n(\underline{t}, 1)$ denote the density of the standard coalescent process, $U_n = (U_{2n}, \ldots, U_{nn})$, given by Equation (5) with $\nu(t) \equiv 1$.

Sampling increases the number of branches in the genealogy and eventually only branches near the present time are included (Figure 3). The latter follows readily from Equations (3) and (4) as $U_{jn} \approx 0$ implies $T_{jn} \approx 0$. As a consequence dense observations are only obtained in the vicinity of $t = 0$ and there cannot exist a consistent non-parametric estimate of $\nu(t)$, $t \geq 0$. If ν belongs to a parameterized family, $\nu(t) = \nu(t; \underline{\alpha})$, $\underline{\alpha} \in A$, the existence of a consistent estimator of $\underline{\alpha}$ depends on the behaviour of the process near zero. This is unfortunate as recent 'trends' in variation of the population size might not be related to variations in the past. As an example consider the case of logistic growth,

$$\nu(t; \beta, c) = \frac{1 + c\, e^{\beta t}}{1 + c}, \qquad (6)$$

(see Pybus et al. 2000 for further explanation and an application to viral data). The parameter β is a growth rate, whereas c determines the onset of growth in the past.

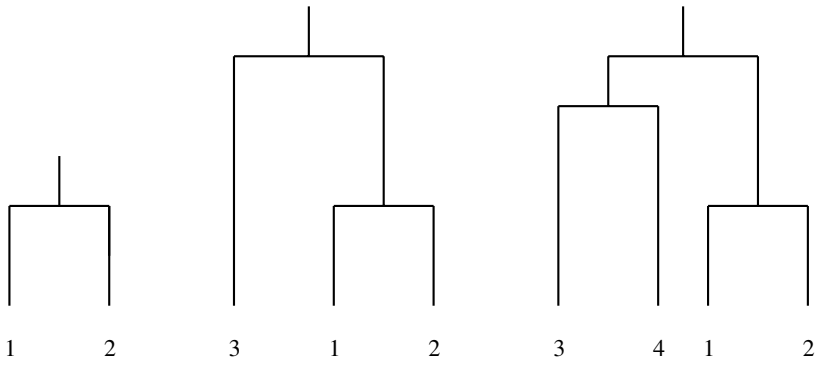

Fig. 3. The genealogy is built up by adding branches to the tree. As soon as the MRCA of the entire population is sampled one cannot hope to obtain sample points further back in time than the MRCA. Eventually only branches near the tips are included.

If t is small so that $\nu(t;\beta,c) \approx 1$, then it is likely that observations nearer zero than t do not improve an estimate of β considerably.

Below, it will be shown that this is the rule rather than the exception: Consistent estimation of $\underline{\alpha}$ fails under very mild assumptions about the population history. To prove this, let $O(t)$ denote a function that is bounded by Ct for all $0 \leq t < D$ and some finite constants C and D.

Theorem 3.1. *Assume $\nu(t) = 1 + at + O(t^2)$ for some finite constant a. Let $g_n(\underline{t},\nu) = f_n(\underline{t},\nu)/f_n(\underline{t},1)$ be the density of T_n wrt U_n. Then*

$$\lim_{n\to\infty} g_n(U_n,\nu)$$
$$= \lim_{n\to\infty} \prod_{j=2}^{n} \nu(U_{jn}) \exp\left\{ -\sum_{j=2}^{n}(j-1)\int_0^{U_{jn}} \nu(t)dt + \sum_{j=2}^{n}(j-1)U_{jn} \right\},$$

with $U_{jn} = \sum_{i=j}^{n} W_i$, exists almost surely and $0 < \lim_n g_n(U_n,\nu) < \infty$ almost surely.

Theorem 3.1 has the immediate consequence that there cannot be a consistent estimator for $\underline{\alpha}$ as $n \to \infty$. It also implies that a test for a hypothesis, $H_1 : \underline{\alpha} \in A_1 \subset A$, against the alternative, $H_2 : \underline{\alpha} \in A \backslash A_1$, cannot obtain power one as $n \to \infty$. In practice, however, these consequences might not be serious, e.g., the power of a test might practically be one. The assumption, $\nu(t) = 1 + at + O(t^2)$, is fulfilled for the logistic growth model, the model of exponential growth (to be introduced in the next section), as well as other models proposed by Pybus et al. (2000). It mainly rules out models where the population size 'explodes' at the present time. For instance consider $\nu(t) = 1 + \sqrt{t}$, that has $\lambda'(0) = -\infty$.

Next, consider $X_n = \psi_0 T_n$ for a fixed ν. Let $P(\cdot,\psi_0,\nu)$ be the probability distribution of X_n. The family $\mathcal{P} = \{P(\cdot,\psi_0,\nu) \mid \psi_0 > 0\}$ constitutes a scale model or a transformation model in the sense of Barndorff-Nielsen et al. (1989). This is to say, $\mathcal{P}$ is generated from $P(\cdot,1,\nu)$ by the group of transformations $x \mapsto \psi_0 x$, $\psi_0 > 0$. If ν is parameterized by $\underline{\alpha} \in A$ then $\{P(\cdot,\psi_0,\nu(\cdot\,;\underline{\alpha})) \mid \psi_0 > 0, \underline{\alpha} \in A\}$ constitutes a composite transformation model, i.e., for each $\underline{\alpha} \in A$, the family $\{P(\cdot,\psi_0,\nu(\cdot\,;\underline{\alpha})) \mid \psi_0 > 0,\}$ is a transformation model. This implies that the distribution of the mles of ψ_0 and $\underline{\alpha}$ have certain nice properties. Before turning to these matters another estimator, $\hat{\phi}_n$, of ψ_0 is introduced. Define $\hat{\phi}_n$ by

$$\hat{\phi}_n = \frac{1}{n-1}\sum_{j=2}^{n}(j-1)X_{jn}. \tag{7}$$

If the population has constant size (i.e., $\nu(t) \equiv 1$), Felsenstein (1992) showed that $\hat{\phi}_n$ is an unbiased estimator of ψ_0 and established convergence to a normal distribution. In the general setting a very similar result holds.

Theorem 3.2. *Assume as in Theorem 3.1. The distribution of $\hat{\phi}_n/\psi_0$ does not depend on ψ_0. Further, under the assumptions of Theorem 3.1, $\hat{\phi}_n$ converges almost surely to ψ_0 for $n \to \infty$, and*

$$\sqrt{n}\,(\hat{\phi}_n - \psi_0) \rightarrow N(0, \psi_0^2) \tag{8}$$

in distribution.

Thus, it is always possible to estimate ψ_0 consistently as $n \rightarrow \infty$. In equation (8), $N(\mu, \sigma^2)$ denotes a normal distributed variable with mean μ and variance σ^2. The next theorem is a consequence of general properties of composite transformation models.

Theorem 3.3. *Let* $(\hat{\psi}_n, \underline{\hat{\alpha}}_n)$ *denote the mle of* $(\psi_0, \underline{\alpha})$*, if it exists, and otherwise let* $(\hat{\psi}_n, \underline{\hat{\alpha}}_n) = (0, 0)$*. Then the distribution of* $(\underline{\hat{\alpha}}_n, \hat{\psi}_n/\psi_0)$ *does not depend on* ψ_0*,* $\hat{\psi}_n$ *is invariant, i.e.,* $\hat{\psi}_n(cX_n) = c\hat{\psi}_n(X_n)$ *for all* $c > 0$*, and* $\underline{\hat{\alpha}}_n$ *is equivariant, i.e.,* $\underline{\hat{\alpha}}_n(cX_n) = \underline{\hat{\alpha}}_n(X_n)$.

The results about $\hat{\psi}_n$ in Theorem 3.3 are also true if $\underline{\alpha}$ is known and $\hat{\psi}_n$ is the profile mle of ψ_0. Whereas $\hat{\phi}_n$ is always defined, $\hat{\psi}_n$ might not exist and if it exists, it might not be unique. When $\hat{\psi}_n$, $n \geq 2$, exists the asymptotic difference between $\hat{\psi}_n$ and $\hat{\phi}_n$ can in special cases be established; e.g., the difference is of order $\log(n)/n$ in the example given in Theorem 5.3.

The fact that $\underline{\hat{\alpha}}_n$ has distribution independent of ψ_0 is of importance: Assume X_n is estimated from DNA sequence data. If X_n is estimated without error, the accuracy of the mle of $\underline{\alpha}$ does not depend on the true scale ψ_0. In practice, however, the accuracy in the estimation of X_n depends on ψ_0. A small ψ_0 indicates little variation in the sample, and many sites are required to ensure reliable estimation of the tree. On the other hand, also very high variation reduces the accuracy in the estimation of X_n. With a high mutation rate all branches tend to be statistically identical.

As an example, consider a two-state Jukes-Cantor model (Jukes and Cantor 1969) and a sample of size two. Let the mutation rate per site per time unit be ψ_0 and let the two sequences be separated by an ancestor time T_2 ago. The chance that the two sequences differ in a particular site is $p = 1/2 - 1/2\exp(-4\psi_0 T_2)$ (Jukes and Cantor 1969). Further, let p be estimated by $\hat{p} = \min\{\sum_j Y_j/k, 1/2\}$, where Y_j is one if the two sequences differ in site j, zero otherwise, and k is the number of sites. Then the expectation and the variance of $S_2 = -\log(1-2\hat{p}) = 4(\widehat{\psi_0 T_2})$ are approximately given by $E(S_2) \approx -\log(1-2p)$ and $Var(S_2) \approx 4p(1-p)/[k\,(1-2p)^2]$ for large k. The variable S_2 is an approximation of $X_2 = \psi_0 T_2$; in particular, $S_2 = T_2$ for $k = \infty$. Figure 4 shows the ratio of the approximative expectation to the approximative standard deviation (sd) of S_2, assuming $k = 1$, for p between 0 and 1/2. The ratio obtains it maximum for $p \approx 0.28$ in which case $4\psi_0 T_2$ is about 0.40. If p is close to 1/2, $\hat{p}$ is often 1/2 and $S_2 = \infty$.

4. Exponential growth with known scale

The coalescent with exponential growth was introduced by Slatkin and Hudson in 1991 and subsequently discussed by Griffiths and Tavaré (1994). Prior to these papers, both Chakraborty (1977) and Kingman (1982b) had discussed similar models. Assume the population size has been increasing exponentially at a constant rate, $\beta \geq 0$, for a long time up till its present size. The population intensity is

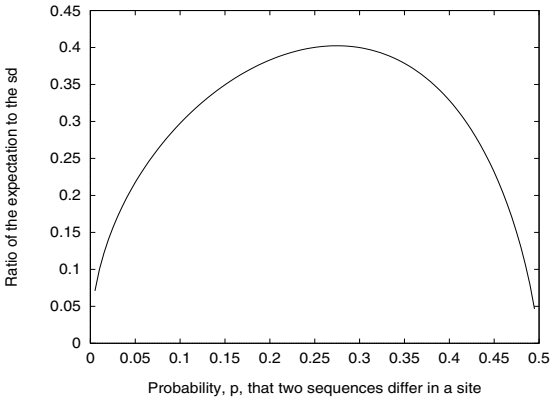

Fig. 4. The figure shows the ratio of the expectation to the sd of S_2 as a function of p, the probability that two sequences differ in a site. If $p = 1/2$ the two sequences are independent. If $p = 0$, the two sequences are identical.

$\lambda(t) = \exp(-\beta t)$. Note that $\nu(t) = 1 + \beta t + O(t^2)$ so that the results from the preceding section applies. Since time is measured in units of $N(0)$ generations, β has the form $\beta = N(0)b$, where b denotes the growth rate per generation. The relation between T_n and U_n ($\beta = 0$) is given by

$$T_{jn} = \frac{1}{\beta}\log(1 + \beta U_{jn}), \tag{9}$$

so that $U_{jn} \geq T_{jn}$ almost surely. In this section, the scale ψ_0 is assumed known ($\psi_0 = 1$ without loss of generality). The density of T_n is,

$$f_n(\underline{t}, \beta) = \frac{n!\,(n-1)!}{2^{n-1}} \exp\left\{ \beta \sum_{j=2}^{n} t_j - \frac{1}{\beta}\sum_{j=2}^{n}(j-1)\left[e^{\beta t_j} - 1\right] \right\} \tag{10}$$

for $t_2 > \ldots > t_n > 0$. Define $\delta_n(T_n)$ by

$$\delta_n(T_n) = \sum_{j=2}^{n} T_{jn} - \frac{1}{2}\sum_{j=2}^{n}(j-1)T_{jn}^2, \tag{11}$$

and assume the true value of β is $\beta_0 \geq 0$. Let 'iff' be short for 'if and only if'.

Theorem 4.1. *The mle $\hat{\beta}_n$ of β_0 exists and is unique almost surely for all n. It fulfills the relations*

$$\hat{\beta}_n = 0 \quad \textit{iff} \quad \delta_n(T_n) \leq 0, \tag{12}$$

and

$$\hat{\beta}_n > 0 \quad \textit{iff} \quad \delta_n(T_n) > 0. \tag{13}$$

Both $P(\hat{\beta}_n > 0)$ and $P(\hat{\beta}_n = 0)$ are positive for all n and β_0, and $P(\hat{\beta}_n > 0) \to 1$ as $\beta_0 \to \infty$.

Equation (11) and Theorem 4.1 can be used to construct an alternative to the log-likelihood test for testing $\beta_0 = 0$ against $\beta_0 > 0$. Define $\Delta_n(T_n)$ by

$$\Delta_n(T_n) \;=\; \frac{2\sum_{j=2}^{n} T_{jn}}{\sum_{j=2}^{n}(j-1)T_{jn}^2}. \tag{14}$$

The mean of $\Delta_n(T_n)$ is close to one for $\beta_0 = 0$ (in fact, the ratio of the mean of the numerator to that of the denominator is one), the mean increases towards infinity for $\beta_0 \to \infty$, and large deviations from one are therefore indicative of expansion. The test will not be further pursued here, but discussed in the next section in connection with a similar test for the case where the scale is unknown.

The asymptotic behaviour of $\hat{\beta}_n$ for $n \to \infty$ is captured in the next theorem.

Theorem 4.2. *The variables $\hat{\beta}_n$, $n \geq 2$, converge in distribution to a non-degenerate variable $\hat{\beta}_\infty$,*

$$\hat{\beta}_n \;\to\; \hat{\beta}_\infty \tag{15}$$

for $n \to \infty$. Both $P(\hat{\beta}_\infty > 0)$ and $P(\hat{\beta}_\infty = 0)$ are positive for all β_0.

For large β_0 the following result is true. Let $\hat{\rho}_n = \hat{\beta}_n/\beta_0$.

Theorem 4.3. *For $\beta_0 \to \infty$,*

$$\log(\beta_0)\,(\hat{\rho}_n - 1) \;\to\; -\log(Y_{n-1}) \tag{16}$$

in distribution, where Y_n denotes a Gamma distributed variable, $Y_n \sim \Gamma(n, n)$. In consequence, $\hat{\rho}_n \to 1$ in distribution and $\liminf_{\beta_0\to\infty} E(\hat{\rho}_n) \;\geq\; 1$.

Simulation results (not shown) suggest that the convergence in β_0 is very slow, as might be anticipated by the scaling of order $\log(\beta_0)$. Therefore it is not practically feasible to approximate the distribution of $\hat{\rho}_n$ with (16). In contrast, the convergence of $\hat{\rho}_n$ towards one is fast, indicating that $\hat{\beta}_n$ provides a reliable estimate of β_0 for large β_0, say $\beta_0 > 100$ for $n > 50$. In addition, Theorem 4.3 establishes that $\hat{\beta}_n$ is positively biased for large β_0. This is also true for $\beta_0 = 0$ because $\hat{\beta}_n \geq 0$ and $P(\hat{\beta}_n > 0) > 0$ for all n (according to Theorem 4.1). Simulation results (not shown) indicate a positive bias for all β_0.

5. Exponential growth with unknown scale

Consider the variables $X_n = \psi T_n$. The relation between $Z_n = \psi U_n$ and X_n is found from (9),

$$X_{jn} \;=\; \frac{1}{\xi}\,\log(1 + \xi Z_{jn}), \tag{17}$$

with $\xi = \beta/\psi$. If $\beta = N(0)b$ and $\psi = N(0)u$, ξ is independent of $N(0)$ and equals b/u. The density of X_n is given by

$$f_n(\underline{x}, \xi, \psi) \;=\; \frac{n!\,(n-1)!}{2^{n-1}\,\psi^{n-1}} \exp\left\{ \xi \sum_{j=2}^{n} x_j \;-\; \frac{1}{\psi\xi} \sum_{j=2}^{n} (j-1)\left[e^{\xi x_j} \;-\; 1\right] \right\} \tag{18}$$

for $x_2 > \ldots > x_n > 0$, $\xi \geq 0$, and $\psi > 0$. Assume the true value of (ξ, ψ) is (ξ_0, ψ_0) and let

$$\gamma_n(X_n) = \sum_{j=2}^{n} X_{jn} - \frac{(n-1)\sum_{j=2}^{n}(j-1)X_{jn}^2}{2\sum_{j=2}^{n}(j-1)X_{jn}}. \tag{19}$$

Note the analogy between $\delta_n(T_n)$ in Equation (11) and $\gamma_n(X_n)$: The last term in (11) is divided by $\hat{\phi}_n$ to compensate the difference in scaling in X_{jn} and X_{jn}^2.

Theorem 5.1. *The mle $(\hat{\xi}_n, \hat{\psi}_n)$ of (ξ_0, ψ_0) exists and is unique almost surely for all n. It fulfills the relations*

$$\hat{\xi}_n = 0 \quad \textit{iff} \quad \hat{\psi}_n = \frac{1}{n-1}\sum_{j=2}^{n}(j-1)X_{jn} \quad \textit{iff} \quad \gamma_n(X_n) \leq 0, \tag{20}$$

and

$$\hat{\xi}_n > 0 \quad \textit{iff} \quad \hat{\psi}_n > \frac{1}{n-1}\sum_{j=2}^{n}(j-1)X_{jn} \quad \textit{iff} \quad \gamma_n(X_n) > 0. \tag{21}$$

If $\gamma_n(X_n) > 0$, then $\hat{\xi}_n$ is the unique solution to $g_n(\xi) = 0$, where

$$g_n(\xi) = \sum_{j=2}^{n} X_{jn} + \frac{1}{\hat{\psi}_\xi \xi^2}\sum_{j=2}^{n}(j-1)\left\{1 - \xi X_{jn}\right\}e^{\xi X_{jn}} - \frac{n(n-1)}{2\hat{\psi}_\xi \xi^2},$$

and

$$\hat{\psi}_\xi = \frac{1}{n-1}\sum_{j=2}^{n}(j-1)(e^{\xi X_{jn}} - 1)\frac{1}{\xi}.$$

If the tree is perfectly star-shaped, $\hat{\xi}_n = \infty$ and $\hat{\psi}_n = \infty$, and otherwise both estimators are finite. Further, $\hat{\psi}_n(cX_n) = c\,\hat{\psi}_n(X_n)$ and $\hat{\xi}_n(cX_n) = \hat{\xi}_n(X_n)/c$ for all $c > 0$.

It is worth remarking that $(\hat{\xi}_n, \hat{\psi}_n)$ exists whenever $X_{2n} \geq X_{3n} \geq \ldots \geq X_{nn} \geq 0$ and $X_{2n} > 0$. This assures that $(\hat{\xi}_n, \hat{\psi}_n)$ is well-defined even if some branches have length zero. This might be the case if X_n is estimated from sequence data. Another interesting fact is that $\gamma_3(X_3) = 3X_{23}X_{33}/(X_{23} + 2X_{33}) > 0$, implying that $\hat{\xi}_3$ is always positive (compare this to Theorem 4.1). Similarly, $\gamma_2(X_2) = X_{22}/2 > 0$ and $\hat{\xi}_2 = \infty$, because a tree based on two sequences is necessarily star-shaped.

Corollary 5.1. *The distribution of $(\psi_0\hat{\xi}_n, \hat{\psi}_n/\psi_0)$ depends on (ξ_0, ψ_0) only through $\beta_0 = \xi_0\psi_0$.*

Note that the mle of β_0 is $\hat{\xi}_n\hat{\psi}_n$. Its distribution depends on β_0 only (this also follows directly from Theorem 3.3).

Theorem 5.2. *The variables $\hat{\xi}_n$ and $\hat{\psi}_n$, $n \geq 2$, converge in distribution, in fact*

$$\hat{\xi}_n \;\rightarrow\; \hat{\xi}_\infty, \quad \textit{and} \quad \sqrt{n}\,(\hat{\psi}_n - \psi_0) \;\rightarrow\; N(0, \psi_0^2) \tag{22}$$

for $n \rightarrow \infty$, with $\hat{\xi}_\infty = \hat{\beta}_\infty/\psi_0$ and $\hat{\beta}_\infty$ defined as in Theorem 4.2.

There is not an analogue of Theorem 4.3 when ψ_0 is unknown because $\hat{\psi}_n$ does not converge to ψ_0 for $\beta_0 \rightarrow \infty$. The next theorem establishes the asymptotic order of convergence of the difference between $\hat{\psi}_n$ and $\hat{\phi}_n$.

Theorem 5.3. *The variables $\hat{\psi}_n - \hat{\phi}_n$, $n \geq 2$, converge in distribution*

$$\frac{n}{\log(n)}\,(\hat{\psi}_n - \hat{\phi}_n) \;\rightarrow\; \psi_0^2\,\hat{\xi}_\infty \;=\; \psi_0\,\hat{\beta}_\infty. \tag{23}$$

for $n \rightarrow \infty$.

The variable, $\hat{\phi}_n$, seems to be biased downwards at least for $n < 100$ and $\beta_0 > 10$ (simulation results not shown). In contrast, $\hat{\psi}_n$ is biased upwards (simulation results not shown). Pybus et al. (2000) provide extensive simulation results for $\hat{\psi}_n$ and $\hat{\xi}_n$ for various parameter values.

Equation (19) and Theorem 5.1 can be used to construct an alternative to the log-likelihood test for testing $\xi_0 = 0$ against $\xi_0 > 0$. Define $\Gamma_n(X_n)$ by

$$\Gamma_n(X_n) \;=\; \frac{2\left(\sum_{j=2}^n X_{jn}\right)\left(\sum_{j=2}^n (j-1)X_{jn}\right)}{(n-1)\,\sum_{j=2}^n (j-1)X_{jn}^2}. \tag{24}$$

The mean of $\Gamma_n(X_n)$ is close to one for $\xi_0 = 0$ and increases towards two as $\xi_0 \rightarrow \infty$ for ψ_0 fixed. Table 1 shows the power of $\Gamma_n(X_n)$ compared to the power of the log-likelihood test for various values of β_0 and n (assuming $\psi_0 = 1$). Note that the distribution of $\Gamma_n(X_n)$ as well as the distribution of the log-likelihood test depend on (ξ_0, ψ_0) only through β_0.

Table 1. Shown is the power (in percentage) of $\Delta_n(T_n)$, $\Gamma(X_n)$, and the log-likelihood ratio, $L_n(X_n) = -2\log(Q_n)$ for small values of $\beta_0 = \xi_0\psi_0$ and various values of the sample size, n. The null hypothesis is $H_T : \beta_0 = 0$ for $\Delta_n(T_n)$, and $H_X : \beta_0 = 0\,(\xi_0 = 0)$ and $\psi_0 > 0$ for the two other statistics. The loss in power can be substantial when ψ_0 is not known, but estimated from data; compare $\Delta_n(T_n)$ with $\Gamma(X_n)$. For small n, the difference in power between $\Gamma(X_n)$ and $L_n(X_n)$ is noteworthy. This difference disappears for large n. 10^6 simulations were performed to obtain the null distribution ($\beta_0 = 0$) for each n and 10^5 simulations for each combination of $\beta_0 > 0$ and n.

The Power of $\Delta_n(T_n)$, $\Gamma_n(X_n)$, and $-2\log(Q_n)$

	$n = 10$			$n = 20$			$n = 100$		
β_0	$\Delta_n(T_n)$	$\Gamma_n(X_n)$	$L_n(X_n)$	$\Delta_n(T_n)$	$\Gamma_n(X_n)$	$L_n(X_n)$	$\Delta_n(T_n)$	$\Gamma_n(X_n)$	$L_n(X_n)$
1	13	12	31	19	14	30	31	27	39
5	89	41	74	98	62	88	100	97	99
10	100	61	90	100	85	98	100	100	100

Table 2. Shown is the power (in percentage) of $\Gamma(X_n)$ for two models, the logistic growth model, $\nu(t) = (1 + ce^{\beta t})/(1 + c)$, and a model of periodic varying population size, $\nu(t) = 1 + c\sin(dt)$, $|c| < 1$. Sample size is $n = 10$. The power of the logistic growth model is very similar to the power of the exponential growth model, whereas the power of the other model is extremely poor. It shows that $\Gamma_n(X_n)$ should only be used if recent growth is anticipated. A similar conclusion holds for the log-likelihood test. 10^5 simulations were performed for each combination of parameters.

The Power of $\Gamma_n(X_n)$ for Various Models

Logistic Growth				Periodic Variation			
β, c	1	10	100	c, d	1	3	10
1	9	11	13	−0.5	3	2	10
10	55	60	61	0.5	7	9	3

The statistics $\Gamma_n(X_n)$ might in general be used as a testor of the hypothesis H_1 : *The population has been of constant size* against the alternative H_2 : *The population size has varied over time*. Table 2 shows the power of $\Gamma_n(X_n)$ for two models, the logistic growth model discussed in Section 3, and a model with periodic varying population size. As in the exponential growth model, the distribution of $\Gamma_n(X_n)$ depends on $\underline{\alpha}$ only, not ψ_0.

6. Discussion

In this paper, I have discussed inference about population history based on an observation of a tree, relating n individuals and drawn from Kingman's coalescent. As mentioned in the introduction, the setting I have adopted corresponds to an ideal situation in which sequences of infinite length are available. However, in practice this is not the case. The variation in nuclear sequences are in general so low that not even the topology can accurately be estimated, let alone the length of individual branches. If this is so, it is natural to estimate demographic parameters based on data from unlinked loci, because all loci share the same demographic history and each locus represents an independent draw from the underlying genealogical process. This approach, however, runs into other difficulties. Different loci are likely to have different mutation rates and it can be hard to argue that all sampled loci evolve under neutrality. One way to circumvent such problems could be to adopt a fully Bayesian approach, assuming prior probabilities on mutation rates, selection coefficients and demographic parameters. Polanski et al. (1998) take a different approach and develop a non-parametric method for inferring past population sizes using a Laplace transform idea.

Other DNA sequence types, e.g., virus sequences, show much higher variation and allow in principle for better estimation of the underlying tree than do nuclear sequences. In principle only, because the mutation mechanism in viruses is often extremely complex and difficult to model, selection is an active player, and only few viruses are known to be non-recombining. Schierup and Hein (2000) showed that if recombination is ignored in analysis of recombining sequences growth is likely to be inferred when in fact there has been no growth at all. In contrast to nuclear sequences one cannot sample unlinked viral loci, because viral genomes

are relatively short. If the virus does not recombine (e.g., Hepatitis C) there is only one locus, if it does, there might be a few, highly correlated loci.

It was argued that no consistent estimator for the growth rate exists (or, indeed, for most other parameter describing population history). This disappointing result reiterates what has been stressed by other authors: In the coalescent model information accumulates slowly for increasing sample size, often at rate $\log(n)$, and in the present case at rate 1. Joyce (1994) found a similar result: He studied the infinite-allele model with selection and showed that consistent estimation of selection coefficients cannot be obtained from population frequencies of alleles.

7. Appendices

Moments of variables that are used in proofs are listed in Appendix 7.4. Let W_j, $j \geq 2$, be a series of independent exponential variables, $W_j \sim Exp(j(j-1)/2)$. Define U_{jn} by $U_{jn} = \sum_{i=j}^{n} W_i$ and T_{jn} by

$$U_{jn} = \int_0^{T_{jn}} \nu(t)\, dt. \tag{25}$$

Then $U_n = (U_{2n}, \ldots, U_{nn})$ and $T_n = (T_{2n}, \ldots, T_{nn})$ are defined on the same probability space, and U_{jn}, $2 \leq j \leq n$, and T_{jn}, $2 \leq j \leq n$, fulfill useful inequalities: In particular, $U_{j+1,n} < U_{jn} < U_{j,n+1} < U_{j\infty}$ almost surely, and $T_{j+1,n} < T_{jn} < T_{j,n+1} < T_{j\infty}$ almost surely. Here $U_{j\infty} = \sum_{i=j}^{\infty} W_i$ and $T_{j\infty}$ is given by Equation (25) with $n = \infty$. The variables $U_{j\infty}$, $j \geq 2$, and $T_{j\infty}$, $j \geq 2$, are finite almost surely according to Appendix 7.4 and Equation (25). Further, $U_{jn} \to U_{j\infty}$ and $T_{jn} \to T_{j\infty}$ almost surely for $n \to \infty$.

This way of defining U_n (and T_n) is convenient for proving the results in the previous sections, but it does not reflect the way sequences are sampled from a population (compare Figure 3).

7.1. Appendix: General results

In this section proofs of the theorems given in Section 3 are derived.

Proof of Theorem 3.1. The form of the density $g_n(\underline{t}, \nu)$ follows readily from Equation (5). Define $h_n(\underline{t}, \nu)$ by $g_n(\underline{t}, \nu) = \exp(h_n(\underline{t}, \nu))$. If $h_n(U_n, \nu), n \geq 2$, converges almost surely, then so does $g_n(U_n, \nu)$. Using $\nu(t) = 1 + at + O(t^2)$, $h_n(U_n, \nu)$ can be rewritten

$$h_n(U_n, \nu) = \sum_{j=2}^{n} \log(\nu(U_{jn})) - \sum_{j=2}^{n} (j-1) \int_0^{U_{jn}} \nu(t) dt + \sum_{j=2}^{n} (j-1) U_{jn}$$

$$= a \sum_{j=2}^{n} U_{jn} - \frac{a}{2} \sum_{j=2}^{n} (j-1)\, U_{jn}^2 + \sum_{j=2}^{n} O[\, U_{jn}^2 + (j-1) U_{jn}^3 \,],$$

for some function $O(y)$. For each possible outcome of U_{jn}, $O[\, U_{jn}^2 + (j-1)U_{jn}^3 \,]$ is eventually dominated by

$$R_{jn} = C\,(\, U_{jn}^2 + (j-1)U_{jn}^3 \,) \leq C\,(\, U_{j\infty}^2 + (j-1)U_{j\infty}^3 \,) = R_j$$

for $n \geq j > J$ and some J and C. Summed over j the right side is finite, $\sum_{j=2}^{\infty} R_j < \infty$ (Appendix 7.4). Thus, further using that ν is continuous,

$$\sum_{j=2}^{n} O[\, U_{jn}^2 + (j-1)U_{jn}^3 \,] \;=\; \sum_{j=2}^{J} O[\, U_{jn}^2 + (j-1)U_{jn}^3 \,] \;+\; \sum_{j=J+1}^{n} R_{jn}$$

is convergent almost surely as $n \to \infty$. The convergence of $h_n(U_n, \nu)$ follows now from Corollary A1 and Lemma A1. □

Lemma A1. *Define $\delta_n(U_n)$ and K_n by*

$$\delta_n(U_n) = \sum_{j=2}^{n} U_{jn} - \frac{1}{2}\sum_{j=2}^{n}(j-1)U_{jn}^2, \tag{26}$$

and

$$K_n \;=\; \frac{2}{n}\sum_{j=2}^{n}(j-1)U_{jn} \;-\; \frac{2}{n}(n-1) \;=\; \frac{2}{n}\sum_{j=2}^{n}\left\{\frac{j(j-1)}{2}W_j - 1\right\}. \tag{27}$$

The series $\delta_n(U_n) - K_n$, $n \geq 2$, is a martingale with expectation $E(\delta_n(U_n) - K_n) = E(\delta_n(U_n)) = 0$ and filter $\mathcal{F}_n = \sigma(W_2, \ldots, W_n)$. Further,

$$\sup_{n\geq 2} E\{\,(\delta_n(U_n) - K_n)^2\,\} \;<\; \infty. \tag{28}$$

Proof of Lemma A1. Note that

$$\delta_n(U_n) \;=\; M_n - N_n, \tag{29}$$

where

$$M_n \;=\; \sum_{j=2}^{n} U_{jn} \;-\; 2\sum_{j=2}^{n}\frac{1}{j} \;=\; \sum_{j=2}^{n}(j-1)W_j \;-\; 2\sum_{j=2}^{n}\frac{1}{j}, \tag{30}$$

and

$$N_n \;=\; \frac{1}{2}\sum_{j=2}^{n}(j-1)U_{jn}^2 \;-\; 2\sum_{j=2}^{n}\frac{1}{j} \;=\; S_n \;-\; 2\sum_{j=2}^{n}\frac{1}{j}. \tag{31}$$

Both M_n, $n \geq 2$, and $N_n + K_n$, $n \geq 2$, are martingales with filter $\mathcal{F}_n = \sigma(W_2, \ldots, W_n)$ and $E(M_n) = E(N_n) = E(K_n) = 0$, because $E(N_{n+1}|\mathcal{F}_n) = N_n + K_n/(n+1)$ and $E(K_{n+1}|\mathcal{F}_n) = nK_n/(n+1)$. Thus, $\delta_n(U_n) - K_n$, $n \geq 2$, is a martingale with the desired expectation and filter. To prove (28) note that

$$E(M_n^2) \;=\; Var(M_n) \;=\; \sum_{j=2}^{n}(j-1)^2 Var(W_j) \;=\; 4\sum_{j=2}^{n}\frac{1}{j^2} \;<\; \frac{2}{3}\pi^2 \;<\; \infty, \tag{32}$$

and

$$E(K_n^2) \;=\; Var(K_n) \;=\; \frac{4(n-1)}{n^2} \;\leq\; 2 \;<\; \infty. \tag{33}$$

If also

$$E(N_n^2) \;=\; Var(N_n) \;<\; c_1 \;<\; \infty, \tag{34}$$

for some constant c_1 then (28) holds by combining (32), (33), and (34), and the proof will be completed. Using $U_{in} = U_{i,j-1} + U_{jn}$ for $2 \le i < j \le n$, it follows that

$$\begin{aligned} 4S_n^2 = \left(\sum_{j=2}^{n} (j-1) U_{jn}^2 \right)^2 \;&=\; \sum_{j=2}^{n} (j-1)^3 U_{jn}^4 \\ &+ \sum_{2 \le i < j \le n} 4\,(i-1)(j-1) U_{jn}^3 U_{i,j-1} \\ &+ \sum_{2 \le i < j \le n} 2\,(i-1)(j-1) U_{jn}^2 U_{i,j-1}^2. \end{aligned} \tag{35}$$

Taking the expectation of N_n^2 using (35), Appendix 7.4, and the independence of $U_{i,j-1}$ and U_{jn} it is found that

$$\begin{aligned} E(N_n^2) \;&= E(S_n^2) \;-\; (E(S_n))^2 \;<\; 4 \sum_{j=2}^{n} \frac{1}{j} \;+\; E\left\{ \sum_{j=2}^{n} (j-1)(j-2) U_{jn}^3 \right\} \\ &+ 2E\left\{ \sum_{j=2}^{n} (j-1) U_{jn}^2 \cdot \left(\sum_{k=2}^{j-1} \frac{1}{k} \right) \right\} - \;4 \left(\sum_{j=2}^{n} \frac{1}{j} \right)^2 + c_2 \\ &< \; 8 \sum_{j=2}^{n} \frac{1}{j} \sum_{k=2}^{j} \frac{1}{k} \;-\; 4 \left(\sum_{j=2}^{n} \frac{1}{j} \right)^2 \;+\; c_3 \;<\; c_1 \;<\; \infty, \end{aligned} \tag{36}$$

where c_1, c_2, and c_3 are constants that apply for all n. The proof is completed. □

Corollary A1. *The variables $\delta_n(U_n)$, $n \ge 2$, converge almost surely and in L^1 for $n \to \infty$ to a non-degenerate variable $\delta(U)$ with mean zero.*

Proof of Corollary A1. Equation (28) in Lemma A1 implies that $\delta_n(U_n) - K_n$, $n \ge 2$, is a uniformly integrable martingale. According to the martingale convergence theorem and Lévy's theorem (e.g., Hoffmann-Jørgensen 1994) this implies that $\delta_n(U_n) - K_n$, $n \ge 2$, converge almost surely and in L^1 to a variable, say $\delta(U)$, with expectation zero. But $K_n \to 0$, $n \ge 2$, almost surely and in L^1 (according to the law of large numbers, e.g. Hoffmann-Jørgensen 1994) and, thus, $\delta_n(U_n)$, $n \ge 2$, converge to $\delta(U)$. If $\delta(U)$ is constant, it is zero and in turn $\delta_n(U_n) = K_n$ for all n (because $\delta_n(U_n) - K_n = E(\delta(U)|\mathcal{F}_n)$ according to Lévy's theorem). This contradicts the definition of $\delta_n(U_n)$, and the corollary is proved. □

Proof of Theorem 3.2. First note that $u = \int_0^t v(z)dz$ implies $t = u - \frac{a}{2}u^2 + O(u^3)$ (after some manipulations). Hence,

$$\begin{aligned}\sqrt{n}\,(\hat{\phi}_n - \psi_0) &= \frac{\sqrt{n}}{n-1}\sum_{j=2}^{n}\left[(j-1)X_{jn} - \psi_0\right] \\ &= \frac{\psi_0\sqrt{n}}{n-1}\sum_{j=2}^{n}\left[\frac{j(j-1)}{2}W_j - 1\right] \\ &\quad - \frac{\psi_0\sqrt{n}}{n-1}\sum_{j=2}^{n}(j-1)\left[\frac{a}{2}U_{jn}^2 - O(U_{jn}^3)\right].\end{aligned}$$

According to Appendix 7.4, the last term converges to 0 in probability, and the first term converges to a standard normal distribution with variance ψ_0^2; hence

$$\sqrt{n}\,(\hat{\phi}_n - \psi_0) \to N(0, \psi_0^2)$$

in distribution, as required. □

Proof of Theorem 3.3. The theorem follows readily from properties of (composite) transformation models (see Barndorff-Nielsen et al. 1989). □

7.2. Appendix: Known scale

In this section proofs of the theorems given in Section 4 are derived.

The first derivative of the log-likelihood can be written, using the series expansion of the exponential, as

$$\begin{aligned}\frac{\partial l_n(T_n,\beta)}{\partial\beta} &= \sum_{j=2}^{n} T_{jn} + \frac{1}{\beta^2}\sum_{j=2}^{n}(j-1)\,(e^{\beta T_{jn}} - 1) - \frac{1}{\beta}\sum_{j=2}^{n}(j-1)T_{jn}\,e^{\beta T_{jn}} \\ &= \delta_n(T_n) - \frac{1}{\beta^2}\sum_{j=2}^{n}(j-1)\sum_{i=3}^{\infty}\frac{(\beta T_{jn})^i}{i!}(i-1) \\ &= \delta_n(T_n) - R_n(\beta, T_n), \qquad (37)\end{aligned}$$

where $R_n(\beta, T_n)$ denotes the power series. Note that the derivative is strictly decreasing in β.

Proof of Theorem 4.1. Existence and uniqueness as well as relations (12) and (13) follow easily from (37). Consider $\hat{\beta}_n > 0$. It follows that

$$\{\hat{\beta}_n > 0\} \supseteq \bigcap_{j=2}^{n}\left\{T_{jn} - \frac{1}{2}(j-1)T_{jn}^2 > 0\right\} = \bigcap_{j=2}^{n}\left\{\frac{2}{j-1} > T_{jn}\right\} = A_n. \qquad (38)$$

But $P(A_n) > 0$ and hence $P(\hat{\beta}_n > 0) > 0$. Similarly,

$$\{\hat{\beta}_n = 0\} \supseteq \bigcap_{j=2}^{n}\left\{\frac{2}{j-1} \le T_{jn}\right\} = B_n, \qquad (39)$$

and $P(\hat{\beta}_n = 0) \geq P(B_n) > 0$. Finally, $T_{jn} = \log(1 + \beta_0 U_{jn})/\beta_0 \to 0$ for $\beta_0 \to \infty$, hence $P(A_n) \to 1$. This completes the proof. □

The following lemma is needed in the proof of Theorem 4.2.

Lemma A2. *Assume $\beta_0 > 0$. The variables, $\delta_n(T_n)$, $n \geq 2$, converge almost surely and in L^1 for $n \to \infty$ to a variable $\delta(T)$ with positive expectation.*

Proof of Lemma A2. Rewrite $\delta_n(T_n)$ as

$$\delta_n(T_n) = \sum_{j=2}^{n} T_{jn} - \frac{1}{2}\sum_{j=2}^{n}(j-1)T_{jn}^2 = \frac{1}{\beta_0}\sum_{j=2}^{n}\log(1+\beta_0 U_{jn}) - \frac{1}{2\beta_0^2}\sum_{j=2}^{n}(j-1)\{\log(1+\beta_0 U_{jn})\}^2. \tag{40}$$

Note that $f_1(x) = x - \log(1+x)$ and $f_2(x) = x^2 - \{\log(1+x)\}^2$ are increasing functions in x for $x \geq 0$ and further that $x^2/2 \geq f_1(x) \geq 0$ and $x^3 \geq f_2(x) \geq 0$. It follows that the variables

$$D_{1n} = \frac{1}{\beta_0}\sum_{j=2}^{n} f_1(\beta_0 U_{jn}) \quad \text{and} \quad D_{2n} = \frac{1}{2\beta_0^2}\sum_{j=2}^{n} f_2(\beta_0 U_{jn}) \tag{41}$$

are positive and increasing in n and bounded in L^1; in fact

$$0 \leq E(D_{1n}) \leq \frac{\beta_0}{2}\lim_{n\to\infty}\sum_{j=2}^{n} E(U_{jn}^2) = 4\beta_0, \tag{42}$$

and

$$0 \leq E(D_{2n}) \leq \frac{\beta_0}{2}\lim_{n\to\infty}\sum_{j=2}^{n}(j-1)E(U_{jn}^3) = 12\beta_0 \tag{43}$$

(see Appendix 7.4). Combining the above with Corollary A1 it is found that

$$\delta_n(T_n) = -D_{1n} + D_{2n} + \delta_n(U_n) \tag{44}$$

converges almost surely and in L^1 for $n \to \infty$ to a variable, say $\delta(T)$. This completes the first part of the lemma, and it will now be shown that $\delta(T)$ has positive mean.

Note that for $\beta_0/2 < \beta < 2\beta_0$, $\partial f_n(\underline{t}, \beta)/\partial\beta = f_n(\underline{t}, \beta)\, \partial l_n(\underline{t}, \beta)/\partial\beta$ is bounded in β;

$$\left|\frac{\partial f_n(\underline{t}, \beta)}{\partial\beta}\right| \leq \frac{n!\,(n-1)!}{2^{n-1}}\exp\left\{2\beta_0\sum_{j=2}^{n} t_j - \frac{2}{\beta_0}\sum_{j=2}^{n}(j-1)e^{\beta_0 t_j/2} + \frac{n(n-1)}{\beta_0}\right\}.$$

$$\times \left\{ \sum_{j=2}^{n} t_j \; + \; \frac{1}{2} \sum_{j=2}^{n} (j-1) t_j^2 \; + \; \frac{2}{\beta_0} \sum_{j=2}^{n} (j-1) t_j e^{2\beta_0 t_j} \right\} \; = \; k_n(\underline{t}, \beta_0), \tag{45}$$

and that k_n is L^1-integrable (wrt Lebesgue measure). This can be seen applying the transform $v_j = 2(e^{\beta_0 t_j/2} - 1)/\beta_0$ to the integral of k_n. As a consequence

$$0 \; = \; \frac{\partial}{\partial \beta} \int f_n(\underline{t}, \beta) d\underline{t} \; = \; \int \frac{\partial f_n(\underline{t}, \beta)}{\partial \beta} d\underline{t} \; = \; E\left(\frac{\partial l_n(T_n, \beta)}{\partial \beta} \right), \tag{46}$$

evaluated in $\beta = \beta_0$. From (46) and (37),

$$0 < E(\delta_n(T_n)) \; = \; E(R_n(\beta_0, T_n)) \; \leq \; E(R_{n+1}(\beta_0, T_{n+1})) \; = \; E(\delta_{n+1}(T_{n+1})),$$

and thus,

$$E(\delta(T)) \; = \; \lim_{n \to \infty} E(\delta_n(T_n)) \; \geq \; E(\delta_n(T_n)) \; > \; 0,$$

because $\delta_n(T_n) \to \delta(T)$ in L^1. The proof is completed. □

One can prove that $\delta_n(T_n)$ converges almost surely under the assumption of Theorem 3.1 and not just under the assumption of an exponentially growing population. However, the proof of Lemma A2 needs convergence in L^1 which cannot be guaranteed under the assumptions of Theorem 3.1.

Proof of Theorem 4.2. Consider the second term, $R_n(\beta, T_n)$, in equation (37). For arbitrary $\beta \geq 0$,

$$\begin{aligned} R_n(\beta, T_n) \; &= \frac{1}{\beta^2} \sum_{j=2}^{n} (j-1) \sum_{i=3}^{\infty} \frac{(\beta T_{jn})^i}{i!} (i-1) \\ &= \beta \sum_{i=0}^{\infty} \frac{\beta^i}{i!(i+1)(i+3)} A_{in} \; \to \; \beta \sum_{i=0}^{\infty} \frac{\beta^i}{i!(i+1)(i+3)} A_{i\infty} \end{aligned} \tag{47}$$

for $n \to \infty$, where $A_{in} = \sum_{j=2}^{n} (j-1) T_{jn}^{i+3}$. $R_\infty(\beta, T_\infty) = \lim_{n\to\infty} R_n(\beta, T_n)$ is bounded by

$$\begin{aligned} \beta \sum_{i=0}^{\infty} \frac{\beta^i}{i!} A_{i\infty} \; &= \beta \sum_{i=0}^{\infty} \sum_{j=2}^{\infty} (j-1) T_{j\infty}^3 \frac{(\beta T_{j\infty})^i}{i!} \\ &= \beta \sum_{j=2}^{\infty} (j-1) T_{j\infty}^3 \exp(\beta T_{j\infty}) \; \leq \; \beta \exp(\beta T_{2\infty}) \sum_{j=2}^{\infty} (j-1) T_{j\infty}^3, \end{aligned}$$

which is finite almost surely according to Appendix 7.4 and $T_{j\infty} \leq U_{j\infty}$. From Lemma A2 and above

$$\lim_{n \to \infty} \frac{\partial l_n(T_n, \beta)}{\partial \beta} \; = \; \delta(T) \; - \; R_\infty(\beta, T_\infty) \tag{48}$$

exists and is finite. Further, for $\delta(T) > 0$ there exists a unique solution, $\hat{\beta}_\infty$, to $\delta(T) - R_\infty(\beta, T_\infty) = 0$; for $\delta(T) \le 0$ let $\hat{\beta}_\infty = 0$. Because of (48) and that $\partial l_n(T_n, \beta)/\partial\beta$ is decreasing in β, it follows that $\hat{\beta}_n \to \hat{\beta}_\infty$ almost surely as $n \to \infty$. This implies that $\hat{\beta}_n$, $n \ge 2$, converge in distribution to $\hat{\beta}_\infty$. This proves the first part of the theorem.

According to Lemma A2, $E(\delta(T)) > 0$ which implies $P(\hat{\beta}_\infty > 0) > 0$. Define δ_3 such that $\delta(T) = \delta_3 + T_{2\infty} - T_{2\infty}^2/2$. Then

$$P(\hat{\beta}_\infty = 0) = \int P(\delta(T) \le 0 \mid T_{3\infty} = t, \delta_3 = d)\, dP_{T_{3\infty},\delta_3}(t, d)$$
$$= \int P(2(1+t)V_{2\infty} - V_{2\infty}^2 \le a \mid T_{3\infty} = t, \delta_3 = d)\, dP_{T_{3\infty},\delta_3}(t, d) > 0,$$

where $a = -2(d+t) + t^2$ and $V_{2\infty} = T_{2\infty} - T_{3\infty}$. This completes the proof. □

Proof of Theorem 4.3. Put $\rho = \beta/\beta_0$. Consider $\partial l_n(T_n, \rho)/\partial\rho = \beta_0 \partial l_n(T_n, \beta)/\partial\beta$ with T_{jn} replaced by $\log(1 + \beta_0 U_{jn})/\beta_0$;

$$\frac{\partial l_n(T_n, \rho)}{\partial\rho} = \sum_{j=2}^{n} \log(1 + \beta_0 U_{jn}) - \frac{n(n-1)}{2\rho^2\beta_0}$$
$$+ \frac{1}{\rho^2\beta_0} \sum_{j=2}^{n} (j-1)\{1 - \rho \log(1 + \beta_0 U_{jn})\}(1 + \beta_0 U_{jn})^\rho. \quad (49)$$

Further, apply the transform $\rho = 1 + r/\log(\beta_0)$ to (49). Then,

$$\frac{\partial l_n(T_n, r)}{\partial r} = \frac{1}{\log(\beta_0)} \frac{\partial l_n(T_n, \rho)}{\partial\rho} \to (n-1) - \exp(r) \sum_{j=2}^{n} (j-1) U_{jn} \quad (50)$$

almost surely as $\beta_0 \to \infty$. The derivative $\partial l_n(T_n, r)/\partial r$ is decreasing in r for all β_0 because $\partial l_n(T_n, \beta)/\partial\beta$ is decreasing, and it is concluded that the mle, $\hat{r}_n(\beta_0)$, of r fulfills

$$\hat{r}_n(\beta_0) \to -\log\left\{\frac{\sum_{j=2}^{n}(j-1)U_{jn}}{n-1}\right\}$$

almost surely as $\beta_0 \to \infty$ (relying on an argument used in the proof of Theorem 4.2). But

$$\hat{r}_n(\beta_0) = \log(\beta_0)\,(\hat{\rho}_n - 1) \quad \text{and} \quad Y_{n-1} := \frac{\sum_{j=2}^{n}(j-1)U_{jn}}{n-1} \sim \Gamma(n-1, n-1),$$

hence $\hat{r}_n(\beta_0) \to -\log(Y_{n-1})$ in distribution and $\hat{\beta}_n/\beta_0 \to 1$ in probability. By Fatou's Lemma, $\liminf_{n\to\infty} E(\hat{\rho}_n) \ge 1$. The proof is completed. □

7.3. Appendix: Unknown scale

In this section proofs of the theorems given in Section 5 are derived.

By differentiation of the log-likelihood,

$$\frac{\partial l_n(X_n,\xi,\psi)}{\partial \xi} = \sum_{j=2}^{n} X_{jn} + \frac{1}{\psi\xi^2}\sum_{j=2}^{n}(j-1)\,(e^{\xi X_{jn}} - 1) - \frac{1}{\psi\xi}\sum_{j=2}^{n}(j-1)X_{jn}e^{\xi X_{jn}}, \tag{51}$$

and

$$\frac{\partial l_n(X_n,\xi,\psi)}{\partial \psi} = -\frac{n-1}{\psi} + \frac{1}{\psi^2\xi}\sum_{j=2}^{n}(j-1)\,(e^{\xi X_{jn}} - 1). \tag{52}$$

Proof of Theorem 5.1. Let

$$\hat{\psi}_\xi = \frac{1}{n-1}\sum_{j=2}^{n}(j-1)\,(e^{\xi X_{jn}} - 1)\frac{1}{\xi} \tag{53}$$

be the solution to $\frac{\partial l_n(X_n,\xi,\psi)}{\partial \psi} = 0$ for fixed $\xi \geq 0$. If $\xi = 0$, $\hat{\psi}_\xi = \frac{1}{n-1}\sum_{j=2}^{n}(j-1)X_{jn}$. By insertion of (53) into (51) one obtains,

$$g_n(\xi) = \sum_{j=2}^{n} X_{jn} + \frac{1}{\hat{\psi}_\xi\xi^2}\sum_{j=2}^{n}(j-1)\left\{1 - \xi X_{jn}\right\}e^{\xi X_{jn}} - \frac{n(n-1)}{2\hat{\psi}_\xi\xi^2}.$$

In particular,

$$g_n(0) = \sum_{j=2}^{n} X_{jn} - \frac{(n-1)\sum_{j=2}^{n}(j-1)X_{jn}^2}{2\sum_{j=2}^{n}(j-1)X_{jn}} = \gamma_n(X_n),$$

and

$$g_n(\infty) = \sum_{j=2}^{n} X_{jn} - (n-1)X_{2n} \leq 0,$$

with equality if and only if $X_{jn} = X_{2n}$ for all j. Define $G_n(\xi)$ by

$$G_n(\xi) = \xi\sum_{j=2}^{n} X_{jn} - (n-1)\log(\hat{\psi}_\xi).$$

Then $\frac{d}{d\xi}G_n(\xi) = g_n(\xi)$ and it will be shown that $-G_n(\xi)$ is convex (in which case $g_n(\xi)$ is decreasing). It is sufficient to prove that $\log(\hat{\psi}_\xi)$ is convex. Rewrite $\hat{\psi}_\xi$ as

$$\hat{\psi}_\xi = \frac{n}{2\xi}\left[\frac{2}{n(n-1)}\sum_{j=2}^{n}(j-1)e^{\xi X_{jn}} - 1\right] = \frac{n}{2\xi}\left[L(\xi) - 1\right] = K(\xi),$$

where $L(\xi)$ is the Laplace transform of

$$P(X = x_j) = \frac{2(j-1)}{n(n-1)}, \tag{54}$$

$2 \le j \le n$ (in that a fixed outcome of $X_{jn} = x_j$ is considered). If $c\,K(\xi)$ is a Laplace transform for some $c > 0$ then $\log(K(\xi))$ is convex (Widder 1946). According to Widder (1946) this is so with $c = 2/[nE(X)]$. It follows that $\log(\hat{\psi}_\xi)$ is convex.

Equations (20) and (21) follow from the fact that $-G_n(\xi)$ is convex. If the tree is perfectly star-shaped ($X_{jn} = X_{2n}$ for all j) then $\hat{\xi}_n = \infty$ (because $g_n(\infty) = 0$) and hence $\hat{\psi}_n = \infty$. The functional relations for $\hat{\psi}_n$ and $\hat{\xi}_n$ follow from Theorem 3.3. □

Proof of Corollary 5.1. Follows from Theorem 3.3. □

Proof of Theorem 5.2. Consider $g_n(\xi)$ from Theorem 5.1. Rewrite as,

$$g_n(\xi) = \sum_{j=2}^{n} X_{jn} - \frac{1}{2\hat{\psi}_\xi}\sum_{j=2}^{n}(j-1)X_{jn}^2 - \frac{1}{\hat{\psi}_\xi}R_n(\xi, X_n) = \sum_{j=2}^{n} X_{jn}$$
$$- \frac{1}{2\psi_0}\sum_{j=2}^{n}(j-1)X_{jn}^2 + \left(\frac{\hat{\psi}_\xi - \psi_0}{2\psi_0\hat{\psi}_\xi}\right)\sum_{j=2}^{n}(j-1)X_{jn}^2 - \frac{1}{\hat{\psi}_\xi}R_n(\xi, X_n),$$

similar to Equation (37). From the proof of Theorem 4.2, $R_n(\xi, X_n)$ is almost surely convergent for $n \to \infty$. Further,

$$\hat{\phi}_n < \hat{\psi}_\xi < \hat{\phi}_n + \frac{\xi}{n-1}\exp(\xi X_{2n})\sum_{j=2}^{n}(j-1)X_{jn}^2 = \hat{\phi}_n + R_n^*(\xi, X_n),$$

where $R_n^*(\xi, X_n)$ denotes the sum (this is obtained similarly to the bound on $R_n(\beta, T_{jn})$ in the proof of Theorem 4.2). The term, $\sqrt{n}R_n^*(\xi, X_n)$, converges almost surely to zero (Appendix 7.4); thus $\hat{\psi}_\xi \to \psi_0$ almost surely and in distribution, and $\sqrt{n}(\hat{\psi}_\xi - \psi_0) \to N(0, \psi_0^2)$ in distribution (as in Theorem 3.2). It follows that

$$g_n(\xi) \to \psi_0 \lim_{n\to\infty}\delta_n(X_n/\psi_0) - \frac{1}{\psi_0}R_\infty(\xi, X_\infty) \tag{55}$$

almost surely for $n \to \infty$. Reasoning similar to reasoning in the proof of Theorem 4.2 gives $\hat{\xi}_n \to \hat{\xi}_\infty$ almost surely (because $g_n(\xi)$ is strictly decreasing for all n). This in turn gives,

$$\hat{\phi}_n < \hat{\psi}_n = \hat{\psi}_{\hat{\xi}_n} < \hat{\phi}_n + R_n^*(\hat{\xi}_n, X_n) < \hat{\phi}_n + R_n^*(\hat{\xi}_\infty + \epsilon, X_n)$$

for $\epsilon > 0$ and sufficiently large n. Further,

$$\sqrt{n}\,R_n^*(\hat{\xi}_\infty + \epsilon, X_n) \to 0$$

in probability for $n \to \infty$ (similar to $\sqrt{n}R_n^*(\xi, X_n) \to 0$), hence

$$\sqrt{n}\,(\hat{\psi}_n - \psi_0) \to N(0, \psi_0^2)$$

in distribution, as required. Finally, it follows that $\hat{\beta}_\infty = \hat{\xi}_\infty \psi_0$ because $\hat{\psi}_n \to \psi_0$ almost surely, and from (55) that $\hat{\beta}_\infty$ is as defined in Theorem 4.2. □

Proof of Theorem 5.3. The proof of this theorem uses the same techniques as in Theorems 4.2 and 5.2 and will not be given here. □

7.4. Appendix: Moments

The following formulas hold for moments of U_{jn}:

$$E(U_{jn}) = \frac{2}{j-1} - \frac{2}{n}, \tag{56}$$

$$E(U_{jn}^2) = 8\sum_{i=j-1}^{n} \frac{1}{i^2} - 8\left(\frac{1}{j-1} - \frac{1}{n} + \frac{1}{(j-1)n}\right), \tag{57}$$

$$E(U_{jn}^3) = 48\left(\frac{1}{j-1} - \frac{1}{n} - 2\right)\sum_{i=j-1}^{n} \frac{1}{i^2} + 96\left(\frac{1}{j-1} - \frac{1}{n} + \frac{1}{(j-1)n}\right), \tag{58}$$

$$16\left(\frac{1}{j-1} - \frac{1}{n}\right)^4 - \frac{k_1}{(j-1)^5} < E(U_{jn}^4) < 16\left(\frac{1}{j-1} - \frac{1}{n}\right)^4 + \frac{k_1}{(j-1)^5}, \tag{59}$$

$$\sum_{j=2}^{n} E(U_{jn}) = 2\sum_{j=2}^{n} \frac{1}{j}, \tag{60}$$

$$\sum_{j=2}^{n} E(U_{jn}^2) = 8\left(1 - \frac{1}{n}\sum_{j=1}^{n} \frac{1}{j}\right), \tag{61}$$

$$\sum_{j=2}^{n} E\{(j-1)U_{jn}\} = n - 1, \tag{62}$$

$$\sum_{j=2}^{n} E\{(j-1)U_{jn}^2\} = 4\sum_{j=2}^{n} \frac{1}{j}, \tag{63}$$

$$\sum_{j=2}^{n} E\{(j-1)U_{jn}^3\} = 24\left(1 - \frac{1}{n}\sum_{j=1}^{n} \frac{1}{j}\right), \tag{64}$$

$$\sum_{j=2}^{n} E\{(j-1)^3 U_{jn}^4\} < 16\sum_{j=2}^{n} \frac{1}{j} + k_2, \tag{65}$$

$$\sum_{j=2}^{n} E\{\,(j-1)(j-2)U_{jn}^{3}\,\} \;<\; 8\sum_{j=2}^{n}\frac{1}{j} \;+\; k_3, \tag{66}$$

and

$$\sum_{j=2}^{n} E\left\{\,(j-1)U_{jn}^{2}\cdot\left(\sum_{k=2}^{j-1}\frac{1}{k}\right)\right\} \;<\; -6\sum_{j=2}^{n}\frac{1}{j} \;+\; 4\sum_{j=2}^{n}\frac{1}{j}\sum_{k=2}^{j}\frac{1}{k} \;+\; k_4. \tag{67}$$

The number k_1 is a constant, independent of j and n, and k_2, k_3, and k_4 are constants, independent of n.

Acknowledgements. O. Pybus, P. Donnelly, J. Hein, and the Mathematical Genetics Group at the department are thanked for helpful comments. Anonymous reviewers are thanked for providing many useful corrections and suggestions. The author was supported by a grant from the Medical Research Council, UK, by a grant from the Biotechnology and Biological Sciences Research Council, UK (BBSRC 43/MMI09788), and by the Carlsberg Foundation, Denmark.

References

Barndorff-Nielsen, O.E., Blæsild, P., Eriksen, P.S.: Decomposition and Invariance of Measures, and Statistical Transformation Models. Lecture Notes in Statistics **58**, Springer-Verlag, New York 1989

Chakraborty, R.: Distribution of nucleotide differences between two randomly chosen cistrons in a population of variable size. Theor. Pop. Biol. **11**, 11–22 (1977)

Chang, J.T.: Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. Math. Biosc. **137**, 51–73 (1996)

Donnelly, P., Tavaré, S.: Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. **29**, 401–421 (1995)

Felsenstein, J.: Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. **59**, 139–147 (1992)

Fu, Y.-X., Li, W.-H.: Maximum likelihood estimation of population parameters. Genetics, **134**, 1262–1270 (1993)

Griffiths, R.C., Tavaré, S.: Sampling theory for neutral alleles in a varying environment. Phil. Trans. R. Soc. Lond. B, **344**, 403–410 (1994)

Hoffmann-Jørgensen, J.: Probability with a View Towards Statistics. Chapman & Hall, New York 1994

Joyce, P.: Likelihood ratios for the infinite alleles model. J. Appl. Prob. **31**, 595–605 (1994)

Jukes, T.H., Cantor, C.R.: Evolution of protein molecules. In Mammalian Protein Metabolism, **II**, H. N. Munro, ed, pp 21–132. Academic Press, New York 1969

Kingman, J.F.K.: The Coalescent. Stoch. Proc. Appl. **13**, 235–248 (1982a)

Kingman, J.F.K.: Exchangeability and the evolution of large populations. In Exchangeability in Probability and Statistics, G. Koch and F. Spizzichino, eds, pp 97-112. North Holland Publishing Company, Amsterdam 1982b

Klein, E.T., Austerlitz, F., Larédo, C.: Some statistical improvements for estimating population size and mutation rate from segregating sites in DNA sequences. Theor. Pop. Biol. **55**, 235–247 (1999)

Polanski, A., Kimmel, M., Chakraborty, R.: Application of a time-dependent coalescence process for inferring the history of population size changes from DNA sequence data. Proc. Nat. Aca. Sc. USA, **95**, 5456–5461 (1998)

Pybus, O.G., Rambaut, A., Harvey, P.H.: An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics **155**, 1429–1437 (2000)
Schierup, M., Hein, J.: Consequences of recombination on traditional phylogenetic analysis. Genetics **155**, 879–891 (2000)
Slatkin, M.W., Hudson, R.R.: Pairwise comparisons of mitochondrial DNA sequences in stable and exponential growing populations. Genetics **129**, 555–562 (1991)
Stephens, M., Donnelly, P.: Inference in molecular population genetics (with discussion). J. R. Stat. Soc. Series B **62**, 605–655 (2000)
Widder, D.V.: The Laplace Transform. Princeton University Press, Princeton 1946