

# Simulating Branching Processes in the Problem of Mitochondrial Eve Dating Based on Coalescent Distributions

Krzysztof A. Cyran

**Abstract**—The paper addresses the problem of dating the most recent common ancestor of modern humans based on mitochondrial DNA record. The applicability of several existing methods which are based on coalescence theory is limited to deterministic trajectories of population, despite the fact that it is known to be unrealistic. In the paper there are described computer simulations which are capable of dealing with different population history scenarios, including populations evolving stochastically and with changing in time environment. Such novel approach arises from comparison of O’Connell’s and Fisher-Wright models. Mitochondrial Eve dating considered in the paper is based on the genetic material from mitochondrial DNA belonging to contemporary humans and Neanderthal fossils. Results indicate that the change of the outgroup species from chimpanzee to Neanderthals is an important factor in terms of reliability and robustness of inferences.

**Keywords**—Branching processes, coalescent distributions, Mitochondrial Eve dating, stochastic computer simulations.

## I. INTRODUCTION

THIS is a well known fact that results of analysis of genetic variation, including such problems as heterozygosity, allele distribution, or linkage disequilibrium, are affected by population history. Therefore the estimation of the probable long-term demographic history of a population has become one of the main problems in statistical genetics, and in the last decade, with the advances of new numerical methods used for estimation of experimental distributions, a lot of research work was focused on inferring human population history from genetic diversity data [1, 2]. In this broad trend there are included studies performed by the author reported in [3], this text being the corrected and extended version of the paper. The majority of methods were based on the Wright-Fisher (WF) model of genetic drift which assumes multinomial sampling between generations and thus asymptotically Poisson distribution of the number of progeny for any individual. Since the assumptions of this model are not always fulfilled in reality, there exists a problem of the influence of

the departure from WF model on the distribution of the coalescence time and further analysis of genetic variation. The author tries to solve the problem using time-forward, numerical simulations of branching processes and numerically approximated distribution of coalescence time for a pair of alleles.

It turns out that the coalescent events, i.e. moments of finding in the genealogy the common ancestors of two individuals, are dependent on many demographical events having the stochastic nature. Therefore, to solve this problem, there were performed extensive computer simulations, numerically estimating the coalescence distribution of a pair of alleles. In these simulations there were considered populations evolving accordingly to various stochastic trajectories. The paper presents how to estimate the time to the most recent female common ancestor (MRFCA) of modern humans, called Mitochondrial Eve (mtEve), by comparison of coalescence time distributions in WF models and in the O’Connell (OC) model ([4] corrected in [5]). The genetic material from hyper variable region I (HVRI) and hyper variable region II (HVRII) of mitochondrial DNA (mtDNA) of *H. sapiens* and *H. neanderthalensis* fossils was applied to these models.

To address the problem, there was performed simulation of over  $10^5$  human population histories evolving for  $10^4$  generations. Assuming the human generation length to be approximately 20 years, each simulation history corresponds to 200,000 years, comparable to time elapsed from mtEve epoch. Simulations of so many trajectories modeling such long periods in an unbiased way excluded the use of built-in pseudo-random number generator. The reason for that is either too short range of generator aperiodicity or failing some statistical tests based on overlapping pairs sparse occupancy (OPSO) [6]. Therefore there was implemented an advanced random number generator being the composition of two other generators. The first was Fibonacci random number generator with period  $2^{120}$  and the second was a generator with period  $2^{24}-1$ , as described in [7]. The resulting advanced generator had the desirable aperiodicity length  $2^{144}$ , moreover, it satisfied known statistical tests. The estimates obtained in the study based on mitochondrial genetic data reported in [8] are very similar to those obtained lately by other researchers with the use of phylogenetic trees, which increases reliability

Manuscript received December 10, 2006; Revised version received June 1, 2007. This work was supported in part under the habilitation grant number BW/RGH-5/Rau-0/2007, under SUT statutory activities, and under MNiSW grant number 3T11F 010 29.

K. A. Cyran is with the Institute of Informatics, Silesian University of Technology, Gliwice, 44-100 Poland, phone: +48-32-237-2500; fax: +48-32-237-2733; e-mail: krzysztof.cyran@polsl.pl.

of both estimates obtained by conceptually different methods.

II. PROBLEM FORMULATION AND METHODOLOGY

A. Estimation of the Expected Coalescence Time

This section presents briefly models for calculating the distributions of time to coalescence of a pair of alleles. In WF models there is used the Bobrowski coalescence distribution [9], whereas the analytical asymptotic coalescence distribution for population following a slightly-supercritical branching process is based on OC model [4]. Next there are presented results of simulations for different population scenarios and Kolmogorov-Smirnow test performed for equality of distributions. There are also given estimates of mtEve time, parameterized by genetic diversity data. Applying genetic data from HVRI and HVRII of mtDNA sequences belonging to *H. sapiens* and *H. neanderthalensis* is postponed until section 4.

**Wright-Fisher Model.** Let us consider the population of haploid individuals, say mtDNA sequences, which at time  $t \geq 0$  has the size  $Z_t$ . Since WF model of genetic drift assumes the multinomial distribution of the number of offspring, two individuals at generation  $t + 1$  are descendants of the single member of generation  $t$  with probability  $p_t = 1/Z_t$  and with probability  $q_t = 1 - p_t$  they are descendants of two different members. Thus the distribution of the time to coalescence of two randomly drawn alleles has the form [9]

$$P(T_c = t) = \prod_{k=T-t}^{T-1} q_k - \prod_{k=T-t-1}^{T-1} q_k = p_{T-t-1} \prod_{k=T-t}^{T-1} q_k, \quad (1)$$

where  $T$  is the number of generations considered and for the sake of mathematical consistency  $q_{-1} = 0$  and  $p_{-1} = 1$ .

**O'Connell Model.** For slightly supercritical time-homogenous Markov branching process with the expected number of offspring  $E(\xi_0) = 1 + \alpha/T + o(1/T)$  and variance  $\text{Var}(\xi_0) = \sigma^2 + O(1/T)$  the probability  $P^x(Z_t > 0)$  ( $P^x$  denotes probabilities starting the process with  $x$  individuals) is given by (see also [4])

$$P^x(Z_t > 0) \sim \frac{2\alpha x}{\sigma^2 T} \left[ 1 - \exp\left(-\alpha \frac{t}{T}\right) \right]^{-1}, \quad \text{as } T \rightarrow \infty. \quad (2)$$

From this it follows that [10]

$$E^x(Z_t | Z_t > 0) = \frac{\sigma^2 T_a}{2\lambda\alpha} (e^\alpha - 1), \quad \text{as } T \rightarrow \infty \quad (3)$$

where  $T_a = \lambda T$  is the equivalent of  $T$  expressed in years ( $\lambda$  years per generation) and  $E^x$  denotes the expected value for process starting with  $x$  individuals. Observe the surprising fact of independence of  $E^x$  with respect to  $x$ , explained in [10].

**Distributions of Coalescence Time.** Let us denote by  $D_T$  the time of the death of the most recent common ancestor (MRCA) of two alleles under consideration, and by  $T_c$  the time to coalescence of these two alleles, counted from the present moment  $T$  backwards into the past. If we assume that ancestor's death time is also the moment of offspring birth, then  $T_c = T - D_T$ . In the case of deterministic trajectory of the population we deal with WF models and consider special cases of the Bobrowski distribution (1). This distribution is presented for piecewise constant and for exponential growth

population scenarios. In the case of stochastic trajectory the O'Connell model and Wright-Fisher model are considered. Finally, the comparison of the distributions is presented.

*Constant and piecewise constant population size.* The assumption about constant population size is unrealistic for a long term population trajectory, however, a piecewise constant trajectory can approximate an arbitrary complex one. This approach was utilized in [2] for inference of the population scenario in ML-based, matrix coalescence method, and it may help to grasp the range of variation of the expected coalescent time  $E(T_c)$  for hypothetical population sizes  $Z$ . We have the following distribution of the time to coalescence of a pair of alleles:

$$\begin{cases} P(T - D_T = t) = P(T_c = t) = \frac{(Z-1)^{t-1}}{Z^t} = \frac{1}{Z} \left( \frac{Z-1}{Z} \right)^{t-1}, & t = 1, 2, \dots, T-1 \\ P(T - D_T = T) = P(T_c = T) = 1 - \sum_{t=1}^{T-1} P(T_c = t). \end{cases} \quad (4)$$

Hence, the expected time to coalescence is

$$\begin{aligned} E(T_c) &= \sum_{t=1}^T t P(T_c = t) = \sum_{t=1}^{T-1} t P(T_c = t) + T P(T_c = T) \\ &= \frac{1}{Z} \sum_{t=1}^{T-1} t \left( \frac{Z-1}{Z} \right)^{t-1} + T \left[ 1 - \frac{1}{Z} \sum_{t=1}^{T-1} \left( \frac{Z-1}{Z} \right)^{t-1} \right] \end{aligned} \quad (5)$$

As  $Z \rightarrow \infty$ , i.e. practically for  $Z > 10^3$  and for  $T < Z$ , we have

$$\ln\left(\frac{Z-1}{Z}\right)^{t-1} = (t-1) \ln\left(1 - \frac{1}{Z}\right) \approx -\frac{t-1}{Z} \quad (6)$$

and

$$\left(\frac{Z-1}{Z}\right)^{t-1} \approx e^{-\frac{t-1}{Z}} \quad (7)$$

and therefore, this time can be approximated by

$$E(T_c) \approx \frac{1}{Z} \sum_{t=1}^{T-1} t e^{-\frac{t-1}{Z}} + T \left[ 1 - \frac{1}{Z} \sum_{t=1}^{T-1} e^{-\frac{t-1}{Z}} \right]. \quad (8)$$

Furthermore, for  $T/Z \rightarrow 0$ , (i.e. practically for  $T/Z < 10^{-3}$ ) we can write

$$E(T_c) \approx \frac{1}{Z} \sum_{t=1}^{T-1} t + T \left( 1 - \frac{1}{Z} \sum_{t=1}^{T-1} 1 \right) = \frac{(T-1)T}{2Z} + T \left( 1 - \frac{T-1}{Z} \right) = T - \frac{T(T-1)}{2Z} \quad (9)$$

or

$$E\left(\frac{T_c}{T}\right) \approx 1 - \frac{T-1}{2Z}. \quad (10)$$

*Exponential growth.* In this scenario, even though in calculations there is used a purely exponential trajectory, we remember that it should be properly rounded to the nearest integer value. The model is unrealistic, mainly due to its homogeneity in time. Assumption that  $Z_{t+1} = R Z_t$  yields the following distribution of coalescence time

$$P(T_c = t) = \left[ \prod_{k=0}^{t-2} (R^{-k} Z_T - 1) \right] \left[ R^{-\frac{t-1}{2}} Z_T^t \right], \quad t = 1, 2, \dots, T-1, \quad (11)$$

$$P(T_c = T) = 1 - \sum_{t=1}^{T-1} P(T_c = t),$$

and therefore, the expected coalescence time is given by

$$\begin{aligned} E(T_c) &= \sum_{t=1}^{T-1} t \left[ \prod_{k=0}^{t-2} (R^{-k} Z_T - 1) \right] \left[ R^{-\frac{t-1}{2}} Z_T^t \right] \\ &+ T \left( 1 - \sum_{t=1}^{T-1} \left[ \prod_{k=0}^{t-2} (R^{-k} Z_T - 1) \right] \left[ R^{-\frac{t-1}{2}} Z_T^t \right] \right), \end{aligned} \quad (12)$$

where  $R = (Z_T / Z_0)^{1/T}$ .









