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 interferences.

***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 hy-per 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 Fibbonacci 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

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]

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

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 $\mathrm{Var}(\xi_0) = \sigma^2 + O(1/T)$ the probability $P^x(Z_t > 0)$ (P^x denotes probabilities starting the process with x individuals) is given by (see also [4])

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

From this it follows that [10]

$$E^x(Z_T \mid Z_T > 0) = \frac{\sigma^2 T_a}{2\lambda\alpha}\left(e^{\alpha} - 1\right), \quad as \quad T \to \infty \tag{3}$$

where $T_a = \lambda T$ is the equivalent of T expressed in years (λ 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:

$$\begin{cases} P(T - D_T = t) = P(T_C = t) = \dfrac{(Z-1)^{t-1}}{Z^t} = \dfrac{1}{Z}\left(\dfrac{Z-1}{Z}\right)^{t-1}, & t = 1, 2, \ldots, T-1 \\ P(T - D_T = T) = P(T_c = T) = 1 - \sum_{t=1}^{T-1} P(T_c = t). \end{cases} \tag{4}$$

Hence, the expected time to coalescence is

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

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

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

and

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

and therefore, this time can be approximated by

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

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

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

or

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

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

$$\begin{aligned} P(T_c = t) &= \left[\prod_{k=0}^{t-2}\left(R^{-k}Z_T - 1\right)\right]\left[R^{\frac{-t(t-1)}{2}} Z_T^t\right], \quad t = 1, 2, \ldots, T-1, \\ P(T_c = T) &= 1 - \sum_{t=1}^{T-1} P(T_c = t), \end{aligned} \tag{11}$$

and therefore, the expected coalescence time is given by

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

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

O'Connell distribution for the branching process. It is assumed that slightly supercritical branching process defined in O'Connell model approximates the long-term history of human population. Given that the population history starts from $N_T = x$ individuals having descendants at T and expressing the time interval $[0,T]$ of a variable t as a unit interval $[0,1]$ of variable $r = t/T$, the distribution of D_T for long times T has the form [4, 5, 11]

$$P(D_T > rT \mid N_T = x) = \\ = \frac{2q_r^x}{(x-1)!}\left[-(1-q_r)^{-x}(x-1)! - F(x-1, 1-q_r)\right], \quad (13)$$

where

$$q_r = \left(e^{-r\alpha} - e^{-\alpha}\right)\left(1 - e^{-\alpha}\right)^{-1} \quad (14)$$

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

$$F(n,y) = \frac{\partial^n}{\partial y^n}\left[y^{-2}\ln(1-y)\right]. \quad (15)$$

It is worth to notice that O'Connell distribution is continuous, however, in order to compare it with discrete empirical distributions described below, it is converted to the discretized version, counted only at points r corresponding to integer values of $t = rT$. For the sake of terminological simplicity, we will refer to this discretized version of distribution as to OC distribution in further text.

Distributions for time-homogeneous branching processes. Let us consider Bobrowski distributions (1) assuming that the long-term demographic history is approximated by a time-homogenous branching process with different offspring distributions. The offspring distributions and their corresponding probability generating functions (pgfs) considered are Poisson (P) distribution

$$f(s) = \sum_{k=0}^{\infty} s^k e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda+s\lambda}, \quad (16)$$

binary fission (BF) distribution

$$f(s) = p^2 + 2p(1-p)s + (1-p^2)s^2 = [p + (1-p)s]^2 \quad (17)$$

and linear fractional (LF) distribution

$$f(s) = \frac{1-b-p}{1-p} + \sum_{k=1}^{\infty} s^k b p^{k-1} = 1 - \frac{b}{1-p} + \frac{bs}{1-ps}. \quad (18)$$

Distributions for time-inhomogeneous branching processes. By inhomogeneity in time there is understood the process evolving with variable in time parameters. This is generalization of the time-homogeneous scenario in which parameters of a process are constant. Time-inhomogeneity is introduced to be able to model the history with variable environmental influence on the reproduction abilities of the population. In particular some extra-genetic inferences about the population growth can be incorporated into this approach by applying a deterministic function $h(t)$ to change moments of the offspring number distribution in time.

For the considered problem the most relevant moment of the distribution is the mean μ of the offspring number distribution. In some cases it can be given by $\mu(t) = h(t)$, however, the goal of the study was to observe the influence of environmental stochastic variability on the shape of the coalescence time distribution. Therefore, instead of deterministic function $h(t)$ the mean μ was changed in time according to the formula: $\mu(t) = \mu_0 + \varepsilon(t)$, where μ_0 is constant, $\varepsilon(t) \sim N(0, \sigma_e)$ and σ_e indicates the scale of environmental variability. In other words Bobrowski coalescence distributions (1) are estimated assuming that population trajectories follow random environment branching processes. It should be noted that in distributions used for offspring number calculation, the change of the mean also changes their variance.

B. Comparison of Distributions

The influence of different population history scenarios on the shape of distributions of time to coalescence is presented in Fig. 1.

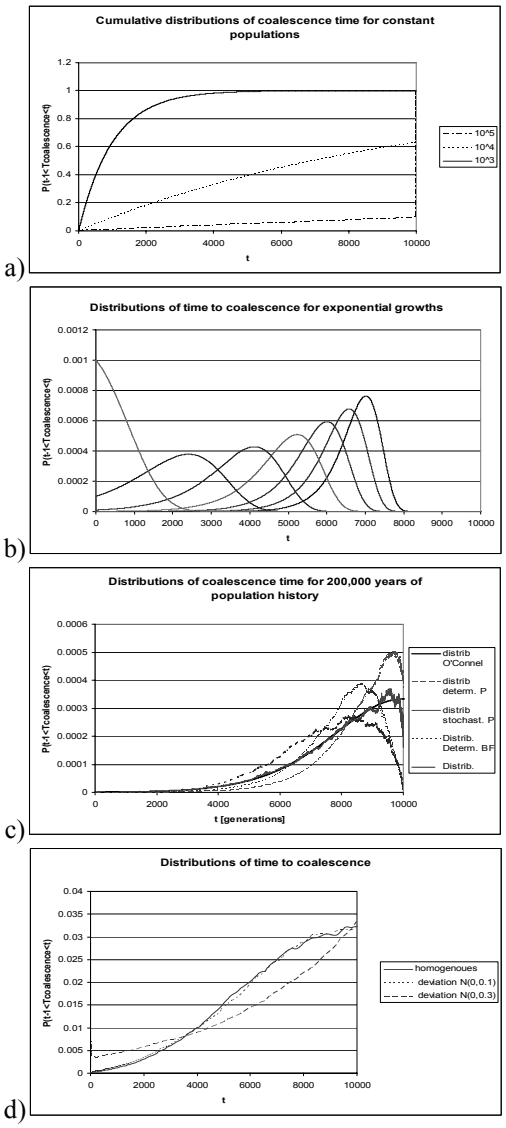

Fig. 1. Distributions of time to coalescence for different population scenarios: a) cumulative distribution for constant effective population size 10^5, 10^4 and 10^3; b) distributions for exponential growth from 1 to (from right to left) 10^9, 10^8, 10^7, 10^6, 10^5, 10^4 and 10^3; c) distributions for stochastic time homogeneous growths; d) distributions for stochastic time-inhomogeneous growths.

The comparison presented as a difference between chosen pairs of distribution is shown in Fig. 2. The formal statistical

comparisons of distributions of the time to coalescence were done with the use of Kolmogorov-Smirnov distributions.

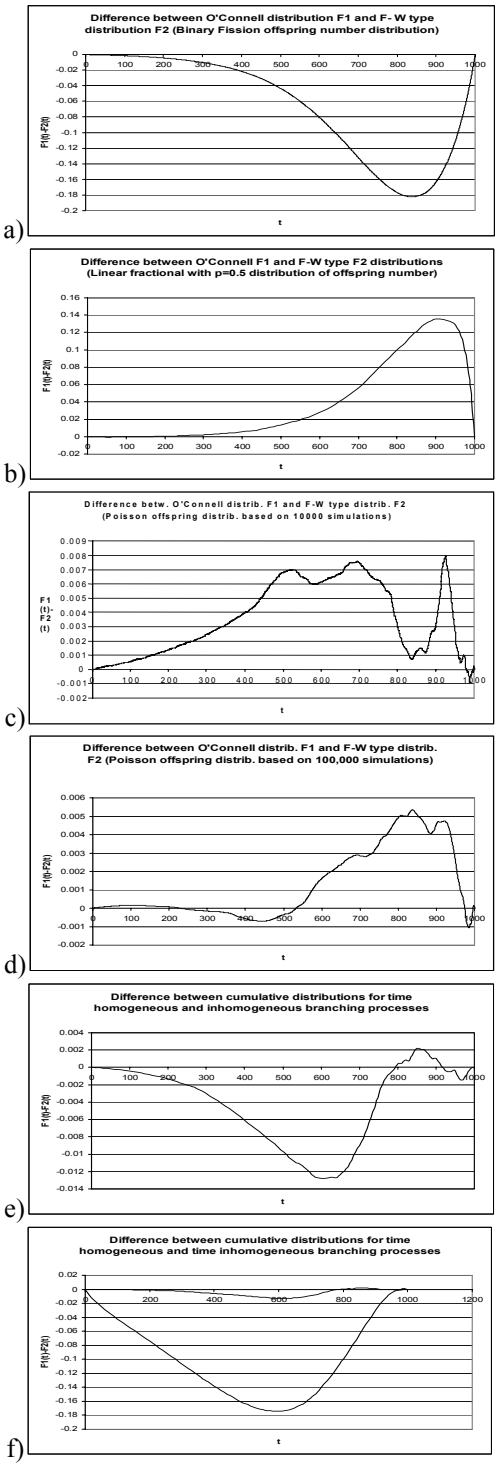

Fig. 2. Pairwise comparison of coalescence time cumulative distributions: a) O'Connell vs. WF type with BF; b) O'Connell vs. WF type with LF; c) O'Connell vs. WF type with P based on 10^4 simulations; d) O'Connell vs. WF type with P based on 10^5 simulations; e) WF type with P time-homogeneous vs. WF type with P time-inhomogeneous $\sigma_e = 0.09\times\mu$; f) WF type with P time-homogeneous vs. F-W type with P time-inhomogeneous $\sigma_e = 0.27\times\mu$ (lower curve) and vs. F-W type with P time-inhomogeneous $\sigma_e = 0.09\times\mu$ (upper curve).

The null hypothesis H_0 stated that Bobrowski distribution $P_{B\text{-}P}$, obtained from n non-extinct simulations of time homogenous branching process with Poisson offspring distribution was equal to the theoretical O'Connell distribution denoted below as P_C. The test versus alternative hypothesis H_1: $P_{B\text{-}P} \neq P_C$ for the statistics

$$d = \sqrt{n}\sup\left|F_{B-P} - F_C\right| \tag{19}$$

was then performed. The obtained value $d = 0.235$ compared to critical value 0.35 of one-sample Kolmogorov-Smirnov distribution at significance level 0.05 indicates that there is no reason for rejecting H_0 at 0.05 significance level. This result is obtained for $n = 1929$ non-extinct branching processes (out of total 10^5). Similar tests for Bobrowski distributions with BF or LF distributions of offspring indicated that they are significantly different from OC distribution. Note that the expected time to coalescence in the case of binary fission offspring distribution is shorter than analogous time for OC distribution. The opposite is true for linear fractional offspring distribution (with $p = 0.5$). This is the effect of different variances of BF, P and LF distributions.

Similar tests were conducted for equality of coalescence time distributions P_H and P_{INH} resulting from time homogenous and inhomogenous branching processes respectively. However, since comparison of two empirical distributions was performed based on numbers of non-extinct simulations n_1 and n_2 respectively, this time the testing statistics had to be changed into

$$d = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\sup\left|F_H - F_{INH}\right|. \tag{20}$$

The inhomogeneity was introduced by random walk of the expected number of offspring with $\sigma_{1e} = 0.09\times\mu$ and $\sigma_{2e} = 3\sigma_{1e} = 0.27\times\mu$. For the first, smaller standard deviation σ_1 the null hypothesis H_0: $P_H = P_{INH}$ can be rejected at significance level 0.05, but not at 0.025, since $d = 0.372$. For larger value of standard deviation σ_{2e} the same null hypothesis can be rejected even at significance level 0.001 since $d = 6.731$ and appropriate 0.1% point of the Kolmogorov-Smirnov distribution is 0.949. So with the increase of stochastic environmental variation, the difference between resulting coalescence time distribution and analogous distribution for constant in time environmental influence is also growing. These results contribute to conclusion that completely random environmental changes have influence on the coalescence time distribution similar to that caused by decreased (with respect to Poisson) variance of offspring distribution, however spanned over longer time (compare plot (a) with plots (e) and (f) in Fig. 2). It is because environmental stochasticity, contrary to demographic one, is not eliminated by enlarging the size of population.

III. Parameterized Estimates of Mitochondrial Eve

The average genetic distance in a sample of n sequences

denoted by d_{avg} and the genetic divergence rate denoted by δ, for $n << Z_T$ are related by the formula

$$E\left(d_{avg} \mid N_0 = x\right) = \delta\lambda E\left(T - D_T \mid N_0 = x\right) = T\lambda\delta E\left(\frac{T_C}{T} \mid N_0 = x\right) \quad (21)$$

yielding moment based estimate for T

$$\hat{T} = d_{avg}\left[\lambda\delta E\left(\frac{T_C}{T} \mid N_0 = 1\right)\right]^{-1}. \quad (22)$$

In WF models $E(T_c/T \mid N_0=1)$ can be numerically obtained from simulations and in OC model required parameters α and T can be simultaneously estimated from (3) and (22). Table 1 presents estimates of relative time of coalescence with respect to total population history length T. In the next section genetic data is incorporated to these parameterized estimates.

Table 1. Estimation of relative time to coalescence of a pair of alleles for different population trajectories. Apart from constant population size, the history starts with 1 individual and ends with number indicated in 5th column. The 1st column defines the population scenario.

Population trajectory	$\gamma = E(T_c/T \mid N_0=1)$	σ	Final population size	Equal to O'Connell distribution?
O'Connell	0.801	0.159	10^7	–
WF, P offspring distr.	0.802	0.159	10^7	Yes
WF, BF offspring distr	0.735	0.17	0.5×10^7	No
WF, LF offspring distr	0.844	0.156	2×10^7	No
WF, P, time inh. σ_{e1}	0.794	0.17	10^7	Not sure
WF, P, time inh. σ_{e2}	0.699	0.269	2×10^7	No
WF, const. pop. size	1	0	10^9	No
WF, const. pop. size	0.995	0.057	10^6	No
WF, const. pop. size	0.95	0.174	10^5	No
WF, const. pop. size	0.632	0.359	10^4	No
WF, const. pop. size	0.1	0.1	10^3	No
WF, exp. growth	0.674	0.062	10^9	No
WF, exp. growth	0.627	0.07	10^8	No
WF, exp. growth	0.565	0.079	10^7	No
WF, exp. growth	0.482	0.092	10^6	No
WF, exp. growth	0.366	0.1	10^5	No
WF, exp. growth	0.216	0.097	10^4	No
WF, exp. growth	0.066	0.05	10^3	No

IV. Applying Genetic Data to Models: Results

Until recently, the estimates of the divergence rate could only rely on time of separation of human and chimpanzee lineages. However, due to relatively long time to that event, all estimates were inaccurate, yielding results from 4 to 9 million years. Therefore estimates of time to mtEve could not be accurate, ranging from 200,000 to 300,000 years ago for methods based on phylogenetic trees. For other methods, including OC method, they were even larger, reaching up to 1 million years. So, these estimates were not only dependent on inaccurate inference about human-chimpanzee divergence time. They depended also on the method applied for inferring.

Fortunately, in 1997 when for the first time the mtDNA from *H. neanderthalensis* dated about 40,000 years ago [12] was sequenced [13], the situation changed. Despite the fact that only fewer than 400 base pairs were sequenced and hence any estimates based on this data were not very reliable, the next successful sequencings of Neanderthal mtDNA in 1999 [8] and 2000 [14, 15] confirmed the accuracy of the first experiment and qualitatively changed the situation in problems of estimating the last female common ancestor of modern humans. At present, divergence rate no longer has to be guessed basing only on problematic dating of human-chimpanzee split. Since it is evident from genetic data [8] that *H. neanderthalensis* did not contribute any detectable mtDNA to modern humans, the time of mtEve can be reasonably placed after *H. sapiens – H. neanderthalensis* separation. For the sample of almost 700 modern humans the average pairwise number of segregating sites in DNA taken from HVRI and HVRII was equal to 35.3 ± 2.3 [8]. Since the analyzed sequences have the total length equal to 600 nucleotides, the average genetic distance $d_{avg\text{M-N}}$, being the parameter in the model studied, is equal to 5.9 %.

The average number of segregating sites in analyzed regions within contemporary human population was 10.9 ± 5.1 [8], and therefore the average genetic distance among contemporary humans d_{avg} can be estimated to a value 1.8 %. The ratio of estimates of $d_{avg\text{M-N}}$ and d_{avg} indicates that the average genetic difference between Neanderthals and modern humans is about 3 times greater than that counted within contemporary humans. Since it is still small enough, it is possible to ignore reverse mutations occurring on both lineages from the time of their divergence T_d some 500,000 years ago [8]. In the infinite allele model (where no reverse mutations are allowed), the parameter called rate of divergence δ can be estimated as equal to $d_{avg\text{M-N}}/T_d$, and therefore its approximate value is $0.06/500{,}000 = 1.2 \times 10^{-7}$. This estimate is within the confidence interval $[5.9 \times 10^{-8}, 1.4 \times 10^{-7}]$ reported in [16]. Using this value of rate of divergence, the time to the most recent female ancestor of contemporary humans expressed in years is $T_a = \lambda T$. The estimates of this time, assuming $\delta = 1.2 \times 10^{-7}$ and $d_{avg} = 0.018$ for different population histories are presented in the Table 2 and in the Table 3 for stochastic and deterministic population scenarios, respectively.

Table 2. Estimates of the time to mtEve $E(T_a)$. In models assuming stochastic scenarios homogeneous in time, letters P, BF and LF state for Poisson, Binary Fission, and Linear Fractional offspring distributions, respectively. In stochastic time inhomogeneous growth models the Poisson offspring distribution was used with the mean (and thus variance) equal to σ_{e1} and $\sigma_{e1} = 3\times\sigma_{e2}$ respectively. The numbers in the bottom row of a table are expressed in thousands of years units.

Stochastic growth					
OC model	WF time-homogeneous			WF time-inhomogeneous	
	P	BF	LF	σ_e 1	σ_e 2
187	187	204	178	189	215

By comparison of the Table 2 with 95 % confidence interval $[111 \times 10^3, 260 \times 10^3]$ [8] it can be concluded that all

predictions under stochastic models fall into it, even though particular coalescence time distributions (see Table 1) are not equal to OC distribution according to Kolmogorov-Smirnov test. Therefore, the predictions of the WF models are not sensitive to actual departures from assumption about multinomial sampling, despite their statistically significant influence on the coalescence time distributions.

Table 3. Estimates of the time to mtEve $E(T_a)$. In deterministic growth scenarios the label $PS10^9$ denotes the final population size equal to 10^9 individuals, and identical notation is applied to labels $PS10^8$ $PS10^7$ and $PS10^6$.

Deterministic growth				
OC model	WF exponential growth			
	$PS10^9$	$PS110^8$	$PS10^7$	$PS10^6$
187	223	239	266	311

V. Discussion and Conclusions

One of the goals of this paper was to implement time-forward numerical simulations of a population following branching processes and to compute experimental distributions of coalescence. The second purpose was to compare distributions of the time to coalescence of a pair of alleles under various population scenarios. For stochastic trajectories the distribution was approximated by more than 10^5 simulated trajectories over time period of 2×10^5 years. In simulations there was considered environmental influence on the number of offspring both constant and randomly changing in time. Resulting WF coalescence time distributions for different offspring distributions were compared with OC coalescence time distribution.

The Kolmogorov-Smirnov test indicated at significance level 0.05 that WF based distributions are equal to OC distribution only if the offspring number follows Poisson distribution. However, by application of advanced numerical methods for computing coalescence distributions it was determined that the expected time to coalescence for any reasonable departures from these requirements is not very sensitive to these departures. This is an important result, since it validates WF models also for population histories not satisfying all assumptions of the model. Having in mind this robustness, considered approach is more general than OC model, as it is applicable to calculate coalescence time distribution for populations evolving both stochastically and with variable in time environmental impacts.

Finally, presented approach was used to estimate the age of mtEve based on the genetic material from contemporary humans and Neanderthal fossil. For all stochastic trajectories the resulting time falls into 95% confidence interval of the estimate based on phylogenetic trees. However, presented results, with the average of 193×10^3 years, indicate a systematic shift of 30×10^3 years towards the past compared to phylogenetic tree based estimates. Since this is not much, the study also showed that after changing the outgroup from chimpanzee to Neanderthals, stochastic genetic models with different assumptions tend to give similar predictions, and therefore these predictions are much more reliable as compared to estimates obtained before sequencing of the hyper variable region II locus in mitochondrial DNA of Neanderthal fossils.

The computer program designed and written by the author is applicable for numerical computations of coalescence time distributions. Such experimentally obtained distributions were used in considered mtEve study, as well as in a paper dealing with a problem of estimating the upper limit of possible Neanderthal admixture in mtDNA of early *H. sapiens* [11]. The program is available at: the location:

http://www.stat.rice.edu/~kimmel/software/coalescence.

Acknowledgments

The author would like to thank Prof. M. Kimmel from Department of Statistics at Rice University in Houston TX, USA, for advice and long discussions concerning branching processes and coalescent theory in application for dating Mitochondrial Eve.

References

[1] S. Wooding and A. Rogers, "A Pleistocence population X-plosion?," *Human Biology*, vol. 72, 2000, pp. 693-695.

[2] S. Wooding and A. Rogers, "The matrix coalescence and an application to human single-nucleotide polymorphisms", *Genetics*, vol. 161, 2002, pp. 1641-1650.

[3] K. A. Cyran, "Mitochondrial Eve dating based on computer simulations of coalescence distributions for stochastic vs. deterministic population models," in *Proc. the 7th WSEAS International Conf. on Systems Theory and Scientific Computation*, Athens, Greece, August 2007, pp. 107-112.

[4] N. O'Connell, "The genealogy of branching processes and the age of our most recent common ancestor," *Adv. Appl. Prob.*, vol. 27, 1995, pp. 418-442.

[5] M. Kimmel and D. Axelrod, *Branching Processes in Biology*. New-York: Springer-Verlag, 2002, pp. 80-83.

[6] G. Marsaglia, "Monkey tests for random number generators," *Comput. Math. Appl*., vol. 9, 1993, pp. 1-10.

[7] G. Marsaglia, A. Zaman, and W. W. Tsang, "Toward a universal random number generator," *Stat. Prob. Lett*., vol. 8, 1990, pp. 35-39.

[8] M. Krings, H. Geisert, R. Schmitz, H. Krainitzki, and S. Pääbo, "DNA sequence of the mitochondrial hypervariable region II from the Neandertal type specimen," *Proc. Natl. Acad. Sci. USA*, vol. 96, 1999, pp. 5581-5585.

[9] A. Bobrowski and M. Kimmel, "Asymptotic behavior of joint distributions of characteristics of a pair of randomly chosen individuals in discrete-time Fisher-Wright models with mutations and drift," *Theoretical Population Biology*, vol. 66, 2003, pp. 355-367.

[10] K. A. Cyran, "Robustness of the dating of the most recent common female ancestor of modern humans," in *Proc. the Tenth National Conf. on Application of Mathematics in Biology and Medicine*, Swięty Krzyż, Poland, 2004, pp. 19-24.

[11] K. A. Cyran and M. Kimmel, "Interactions of Neanderthals and modern humans: what can be inferred from mitochondrial DNA?" *Math. Biosci. Eng.*, vol. 2, 2005, pp. 487-498.

[12] R. Schmitz, G. Bonani, and F. H. Smith, "New research at the Neandertal type site in the Neander Valley of Germany.," in *Annu. Meeting of the Paleoanthropology Society*, Denver, March 2002, pp. 19-20.

[13] M. Krings M, A. Stone, R. Schmitz, H. Krainitzki, M. Stoneking, and S. Pääbo, "Neandertal DNA sequences and the origin of modern humans," *Cell*, vol. 90, 1997, pp. 19-30.

[14] I. Ovchinnikov, A. Götherström, G. Romanova, V. Kharitonov, K. Lidén, and W. Goodwin W, "Molecular analysis of Neanderthal DNA from the Northern Caucasus, *Nature*, vol. 404, 2000, pp. 490-493.
[15] M. Krings, C. Capelli, F. Tschentscher, H. Geisert, S. Meyer, A. von Haeseler, K. Grossschmidt, G. Possnert, M. Paunovic, and S. Pääbo, " A view of Neandertal genetic diversity," *Nature Genetics*, vol. 26, 2000, pp. 144-146.
[16] J. Adachi and M. Hasegawa, Improved dating of the human-chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites," *J. Mol. Evol.*, vol. 40, 1995, pp. 622-628.

Krzysztof A. Cyran was born in Cracow, Poland, in 1968. He received MSc degree in computer science (1992) and PhD degree (with honours) in technical sciences with specialty in computer science (2000) from the Silesian University of Technology SUT, Gliwice, Poland. His PhD dissertation addresses the problem of image recognition with the use of computer generated holograms applied as ring-wedge detectors.

He has been an author and co-author of more than 60 technical papers in journals (several of them indexed by Thomson Scientific) and conference proceedings. These include scientific articles like: K. A. Cyran and A. Mrózek, "Rough sets in hybrid methods for pattern recognition," *Int. J. Intel. Syst.*, vol. 16, 2001, pp. 149-168, and K. A. Cyran and M. Kimmel, "Interactions of Neanderthals and modern humans: what can be inferred from mitochondrial DNA?" *Math. Biosci. Eng.*, vol. 2, 2005, pp. 487-498, as well as a monograph: U. Stańczyk, K. Cyran, and B. Pochopień, *Theory of Logic Circuits*, vol 1 and 2, Gliwice: Publishers of the Silesian University of Technology, 2007. Dr. Cyran (in 2003-2004) was a Visiting Scholar in Department of Statistics at Rice University in Houston, US. He is currently the Assistant Professor and the Vice-Head of the Institute of Informatics at Silesian University of Technology, Gliwice, Poland. His current research interests are in image recognition and processing, artificial intelligence, digital circuits, decision support systems, rough sets, computational population genetics and bioinformatics.

Dr. Cyran has been involved in numerous statutory projects led at the Institute of Informatics and some scientific grants awarded by the State Committee for Scientific Research. He also has received several awards of the Rector of the Silesian University of Technology for his scientific achievements. In 2004-2005 he was a member of International Society for Computational Biology. Currently he is a member of the Editorial Board of Journal of Biological Systems, member of the Scientific Program Committee of WSEAS international conferences in Malta (ECC'08), Rodos (AIC'08, ISCGAV'08, ISTASC'08) and multiconference in Crete (CSCC'08) as well as a reviewer for Studia Informatica and such journals indexed by Thompson Scientific as: Optoelectronic Review, Mathematical Biosciences and Engineering, and Journal of Biological Systems.