Mitochondrial Eve Dating based on Computer Simulations of Coalescence Distributions for Stochastic vs. Deterministic Population Models

KRZYSZTOF A. CYRAN
Institute of Informatics
Silesian University of Technology
16 Akademicka Street, 44-100 Gliwice
POLAND

Abstract: - One of the crucial issues in contemporary evolutionary genetics is dating of the common ancestors of different species. Applicability of several existing approaches based on coalescence theory is limited to deterministic population trajectories, known to be unrealistic. In the paper the computer simulation based approach is presented, which is capable to deal with different population history scenarios, including populations evolving stochastically and with changing environment. This approach arises from comparison of O'Connell's and Fisher-Wright models. It is applied to estimate the age of our most recent female common ancestor, called Mitochondrial Eve, based on the genetic material from mitochondrial DNA belonging to contemporary humans and Neanderthal fossils. Obtained results indicate that after changing the outgroup from chimpanzee to Neanderthals, the stochastic genetic models with different assumptions tend to give similar predictions, and therefore these predictions are much more reliable than they were before.

Key-Words: Stochastic trajectories, coalescent distributions, Mitochondrial DNA, Neanderthal fossils, Mitochondrial Eve dating, Branching processes, Genetic information processing, Stochastic computer simulations.

1 Introduction

The 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 a lot of research work has been focused on inferring human population history from genetic diversity data [1, 2]. The majority of methods were based on the Fisher-Wright (FW) model of genetic drift which assumes multinomial sampling between generations and thus asymptotically Poisson distribution of the number of progeny for any individual. Since this model is not always accurate, there exists a problem of the influence of the departure from FW model on the distribution of the coalescence time and further analysis of genetic variation.

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 address this problem, we performed an extensive computer simulations, estimating the coalescence distribution for populations evolving according to various stochastic scenarios. 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 FW models and in the O'Connell (OC) model ([3] corrected in [4]). For this purpose we used the genetic data from hyper variable region I (HVRI) and hyper variable region II (HVRII) of mitochondrial DNA (mtDNA) of modern humans and Neanderthal fossils.

To draw conclusions, we simulated over 10^5 human population trajectories over time period of 10^4 generations (it is equivalent to about 200,000 years, comparable to time elapsed from mtEve until present, if we assume the human generation length to be approximately 20 years) The estimates we obtain based on genetic data from HVRI and HVRII of mtDNA of modern humans and Neanderthal fossil [5] are very similar to those obtained with the use of phylogenetic trees. Similar estimates, based on conceptually different methods but applied to the very same underlying biological system which processes the information during genetic evolution, make these methods more reliable. The experimental confirmation of this fact is one of the relevant results of presented here computer simulation based study.

2 Estimation of the expected time to coalescence

This section presents briefly models for calculating the distributions of time to coalescence of a pair of alleles.

In FW models we use the Bobrowski coalescence distribution [6], whereas the analytical asymptotic coalescence distribution for population following a slightly-supercritical branching process is based on OC model [3]. Then we present results of simulations for different population scenarios and perform Kolmogorov-Smirnow test for equality of distributions. We also give 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 3..

2.1 Fisher-Wright Model

Let us consider the population of haploid individuals, say mtDNA sequences, which at time $t \geq 0$ has the size Z_t. Since FW 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 [6]:

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

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

2.2 O'Connell's 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 $\mathrm{Var}(\xi_0) = \sigma^2 + O(1/T)$ it holds that [7]:

$$\lim_{T\to\infty} E(Z_T \mid N_0 = x) = \frac{\sigma^2 T_a}{2\lambda\alpha}\left(e^{\alpha} - 1\right), \qquad (2)$$

where $T_a = \lambda T$ is the equivalent of T expressed in years (λ years per generation) and N_t denotes the number of individuals at n t who persist alive also in generation T.

2.3 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 for the sake of simplicity and without loss of generality we assume that considered ancestor's time of the death is also the moment of offspring birth, then the equation $T_c = T - D_T$ relates these three time events.

2.3.1 Deterministic Cases

In the case of deterministic trajectory of the population we deal with FW models and consider special cases of the Bobrowski distribution (1). This distribution is presented for piecewise constant and for exponential growth population scenarios.

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:

$$\begin{cases} P(T - D_T = t) = P(T_C = t) = \dfrac{(Z-1)^{t-1}}{Z^t} = \dfrac{1}{Z}\left(\dfrac{Z-1}{Z}\right)^{t-1}, \quad t = 1, 2, \ldots, T-1 \\ P(T - D_T = T) = P(T_c = T) = 1 - \sum_{t=1}^{T-1} P(T_c = t). \end{cases} \qquad (3)$$

Hence, the expected time to coalescence is

$$E(T_c) = \sum_{t=1}^{T} tP(T_c = t) = \sum_{t=1}^{T-1} tP(T_c = t) + TP(T_c = T) = \\ = \frac{1}{Z}\sum_{t=1}^{T-1} t\left(\frac{Z-1}{Z}\right)^{t-1} + T\left[1 - \frac{1}{Z}\sum_{t=1}^{T-1}\left(\frac{Z-1}{Z}\right)^{t-1}\right] \qquad (4)$$

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

$$\ln\left(\frac{Z-1}{Z}\right)^{t-1} = (t-1)\ln\left(1 - \frac{1}{Z}\right) \approx -\frac{t-1}{Z} \qquad (5)$$

and

$$\left(\frac{Z-1}{Z}\right)^{t-1} \approx e^{-\frac{t-1}{Z}} \qquad (6)$$

and therefore, this time can be approximated by

$$E(T_c) \approx \frac{1}{Z}\sum_{t=1}^{T-1} te^{-\frac{t-1}{Z}} + T\left(1 - \frac{1}{Z}\sum_{t=1}^{T-1} e^{-\frac{t-1}{Z}}\right). \qquad (7)$$

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

$$E(T_c) \approx \frac{1}{Z}\sum_{t=1}^{T-1} t + T\left(1 - \frac{1}{Z}\sum_{t=1}^{T-1} 1\right) = \frac{(T-1)T}{2Z} + T\left(1 - \frac{T-1}{Z}\right) = T - \frac{T(T-1)}{2Z} \qquad (8)$$

or

$$E\left(\frac{T_c}{T}\right) \approx 1 - \frac{T-1}{2Z}. \qquad (9)$$

Exponential growth

In this scenario, even though in calculations we use 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. Assuming that $Z_{t+1} = R\,Z_t$ yields the following distribution of coalescence time

$$P(T_c = t) = \left[\prod_{k=0}^{t-2}\left(R^{-k}Z_T - 1\right)\right]\left[R^{\frac{-t(t-1)}{2}}Z_T^t\right]^{-1},$$
$$t = 1, 2, ..., T-1, \quad P(T_c = T) = 1 - \sum_{t=1}^{T-1} P(T_c = t), \tag{10}$$

and therefore, the expected coalescence time is given by

$$E(T_c) = \sum_{t=1}^{T-1} t\left[\prod_{k=0}^{t-2}\left(R^{-k}Z_T - 1\right)\right]\left[R^{\frac{-t(t-1)}{2}}Z_T^t\right]^{-1} +$$
$$+ T\left(1 - \sum_{t=1}^{T-1}\left[\prod_{k=0}^{t-2}\left(R^{-k}Z_T - 1\right)\right]\left[R^{\frac{-t(t-1)}{2}}Z_T^t\right]^{-1}\right), \tag{11}$$

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

2.3.2 Coalescence in Stochastic Models

O'Connell's distribution for the branching process

Let us consider slightly supercritical time-homogenous Markov branching process with the expected number of offspring $E(\xi_0) = 1 + \alpha/T + o(1/T)$ and variance $\mathrm{Var}(\xi_0) = \sigma^2 + O(1/T)$ where the time interval $[0,T]$ of a variable t is expressed as a unit interval $[0,1]$ of variable $r = t/T$. Additionally let us assume this branching process approximates the long-term history of human population. Then [3, 4] for long times T we are able to compute the tail of the distribution of D_T, i.e the time of death of last common ancestor, given that we start the population history from x individuals having descendants at T.

Fisher-Wright based distributions for time-homogeneous branching processes

Let us calculate 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) we consider are, Poisson (P) distribution

$$f(s) = \sum_{k=0}^{\infty} s^k e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda+s\lambda}, \tag{12}$$

binary fission (BF) distribution

$$f(s) = p^2 + 2p(1-p)s + (1-p^2)s^2 = [p + (1-p)s]^2 \tag{13}$$

and linear fractional (LF) distribution

$$f(s) = \frac{1-b-p}{1-p} + \sum_{k=1}^{\infty} s^k b p^{k-1} = 1 - \frac{b}{1-p} + \frac{bs}{1-ps}. \tag{14}$$

Fisher-Wright based distributions for time inhomogeneous process

By inhomogeneity in time we understand process evolving with variable in time parameters. Thus we generalize the time-homogeneous scenario in which parameters of process are constant. We introduce time-inhomogeneity 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. The most influential on our problem moment, then mean μ of the offspring number distribution is then given by $\mu(t) = h(t)$.

Our goal however was to observe the influence of environmental stochastic variability on the shape of the coalescence time distribution. Therefore instead of deterministic function $h(t)$ we change μ in time according to 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 we estimate Bobrowski coalescence distributions (1) assuming that population trajectories follow random environment branching processes. It should be noted that in distributions we use for offspring number calculation, the change of the mean also changes their variance.

Comparison of distributions

The influence of different population history scenarios on the shape of distributions of time to coalescence is presented in Fig. 1. We also conducted tests for equality of coalescence time distributions P_H and P_{INH} resulting from time homogenous and inhomogenous branching processes simulations respectively. Since we compare two empirical distributions based on numbers of non-extinct simulations n_1 and n_2 respectively, the testing statistics of Kolmogorov-Smirnov test has the form:

$$d = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\sup\left|F_H - F_{INH}\right|. \tag{15}$$

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 5%, but not at 2.5%, since $d = 0.372$. For larger value of standard deviation σ_{2e} the same null hypothesis can be rejected even at significance level 0.1% 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.

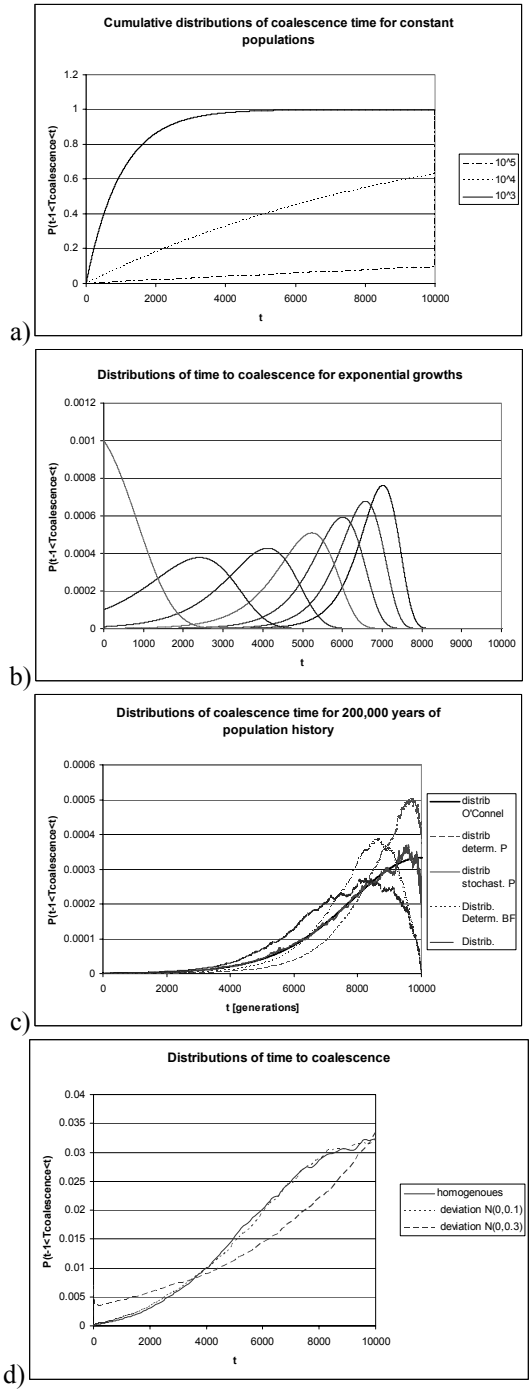

Fig. 1. .Distributions of time to coalescence for different population scenarios: a) cumulative distribution for constant effective population size 10^5, 10^4 and10^3 b) distributions for exponential growth from 1 to (from right to left) 10^9, 10^8, 10^7, 10^6, 10^5, 10^4 and 10^3 c) distributions for stochastic time homogeneous growths d) distributions for stochastic time-inhomogeneous growths.

These results contribute into 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 first and third row of Fig. 2). It is because environmental stochasticity, on contrary to demographic one, is not eliminated by the enlarging size of population.

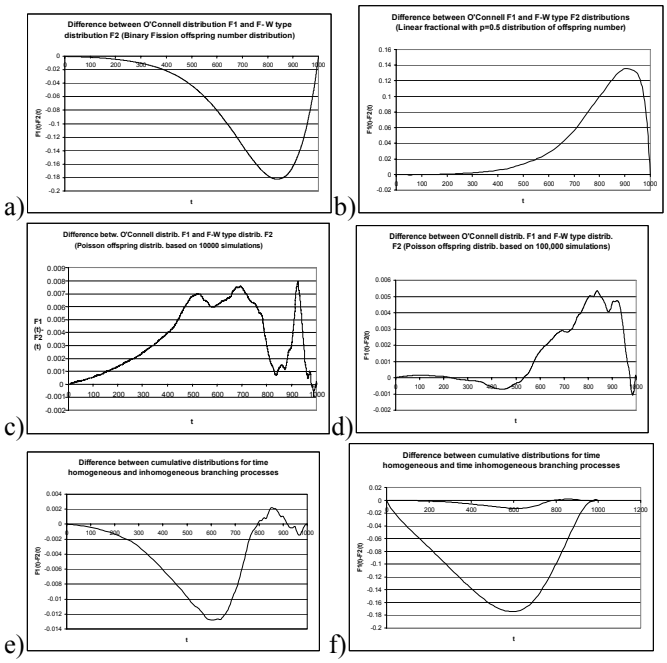

Fig. 2. Pairwise comparison of coalescence time cumulative distributions: a) O'Connell's vs. FW type with BF b) O'Connell's vs. FW type with LF c) O'Connell's vs. FW type with P based on 10^4 simulations d) O'Connell's vs. FW type with P based on 10^5 simulations e) FW type with P time-homgeneous vs. FW type with P time-inhomogeneous $\sigma_e = 0.09\times\mu$ f) FW type with P time-homgeneous 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).

In the Table 1 we summarize estimations of relative time of coalescence with respect to total population history length *T*. In next sections we substitute to these parameterized estimates genetic data.

3 Applying Genetic Data to Models

Until recently, the estimation of the divergence rate could rely only on human-chimpanzee divergence data. However due to relatively long time to this divergence, all estimates of this time were very inaccurate ranging from 4 to 9 million years. Consequently estimated divergence rate and time to mtEve could not be accurate, ranging from 200,000 to 300,000 years ago for methods

based on phylogenetic trees. These estimates not only were dependent on inaccurate inference about human-chimpanzee divergence time. They depended also on the method applied for inferring.

Situation has changed after 1997 [9] when for the first time the mtDNA from *H. neanderthalensis* dated about 40,000 years ago [10] was sequenced. However, only less than 400 base pairs were sequenced, hence any estimates based on this data were not very reliable. The next successful sequencings of Neanderthal mtDNA in 1999 [5] and 2000 [11, 12] confirmed the accuracy of the first experiment and qualitatively changed the situation in problems of estimating the last female common ancestor of modern humans. The present divergence rate no longer has to be guessed relying on the assumption of its constancy over a few million years, and problematic dating of human-chimpanzee divergence.

Since it is evident from genetic data [7] that *H. neanderthalensis* did not contribute any mtDNA to modern humans, the time of mtEve has to be clearly placed after *H. sapiens – H. neanderthalensis* divergence. 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 [5]. Since the analyzed sequences have the total length equal to 600 nucleotides, the average genetic distance $d_{avgM\text{-}N}$, being the parameter in our model, is equal to 5.9 %.

Table 1. Estimation of relative time to coalescence of a pair of alleles for different pop. histories. Apart from constant pop. size, the history starts with 1 individual and ends with number indicated in 3rd column. The 1st column defines scenario

Population trajectory	$\gamma = E(T_c /T \mid N_0=1)$	Final population size
O'Connell's	0.801	10^7
FW, P offspring distr.	0.802	10^7
FW, BF offspring distr	0.735	0.5×10^7
FW, LF offspring distr	0.844	2×10^7
FW, P, time inh. σ_{e1}	0.794	10^7
FW, P, time inh. σ_{e2}	0.699	2×10^7
FW, const. pop. size	1	10^9
FW, const. pop. size	0.995	10^6
FW, const. pop. size	0.95	10^5
FW, const. pop. size	0.632	10^4
FW, const. pop. size	0.1	10^3
FW, exp. growth	0.674	10^9
FW, exp. growth	0.627	10^8
FW, exp. growth	0.565	10^7
FW, exp. growth	0.482	10^6
FW, exp. growth	0.366	10^5
FW, exp. growth	0.216	10^4
FW, exp. growth	0.066	10^3

The estimates of the time to mitochondrial Eve, assuming the values of parameters: $\delta = 1.2 \times 10^{-7}$ 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

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

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	FW exponential growth			
	$PS10^9$	$PS10^8$	$PS10^7$	$PS10^6$
187	223	239	266	311

By comparison of the Table 2 with 95 % confidence interval $[111 \times 10^3, 260 \times 10^3]$ of the mitochondrial Eve epoch [5] we conclude that all predictions under stochastic models fall into it, despite that particular coalescence time distributions are not equal to OC distribution according to Kolmogorov-Smirnov test. Therefore, the predictions of the FW models are not sensitive on actual departures from assumption about multinomial sampling, despite their statistically significant influence on the coalescence time distributions.

4 Conclusion

One of the goals of this paper 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 so many simulations we considered environmental influence on the number of offspring both constant and randomly changing in time. Resulting FW 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 FW based distributions are equal to OC distribution only if the offspring number follows Poisson distribution. However, we also determined that the expected time to coalescence for any reasonable departures from these requirements is not very sensitive to these departures. This is important and original result. It validates FW models used in many population genetic studies also for population histories not satisfying all assumptions of the model. Moreover, having in mind this robustness of FW model, we consider our approach 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 what cannot be studied in O'Connell model.

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 fall into 95% confidence interval of the estimate based on phylogenetic trees. However our 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, we also showed in this paper, 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.

Written by the author computer program used for computations of coalescence time distributions described in this paper, as well as in a problem of estimating the upper limit of possible Neanderthal admixture in mtDNA of early *H.* sapiens [8] is available on the web page:
http://www.stat.rice.edu/~kimmel/software/coalescence

Acknowledgments. The author would like to acknowledge the financial support of this work under the habilitation grant number BW/RGH-5/Rau-0/2007 and SUT statutory activities BK/2007. Also, the author would like to thank to Prof. Marek Kimmel from Department of Statistics at Rice University in Houston TX, USA, for long discussions and advices concerning branching processes theory and its applications to biological problems.

References:

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

[2] S. Wooding, A. Rogers, The Matrix Coalescence and an Application to Human Single-Nucleotide Polymorphisms, *Genetics*, Vol. 161, 2002, p. 1641-1650.

[3] 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.

[4] M. Kimmel, D. Axelrod, *Branching Processes in Biology*. Springer-Verlag, New-York, 2002.

[5] M. Krings, H. Geisert, R. Schmitz, H. Krainitzki, 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.

[6] A. Bobrowski, 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.

[7] K. A. Cyran, M. Kimmel, Robustness of the dating of the most recent common female ancestor of modern humans, *Proc. Tenth national Conference on Application of Mathematics in Biology and Medicine*, Święty Krzyż, Poland, 2004, pp. 19-24.

[8] K.A. Cyran, M. Kimmel, Interactions of Neanderthals and Modern Humans: What Can Be Inferred from Mitochondrial DNA?, *Math. Biosci. Eng.*, Vol. 2, 2005, pp. 487-498.

[9] M. Krings, A. Stone, R. Schmitz, H. Krainitzki, M. Stoneking, S. Pääbo, Neandertal DNA Sequences and the Origin of Modern Humans, *Cell*, Vol. 90, 1997, pp. 19-30.

[10] R. Schmitz, G. Bonani, F.H. Smith, New Research at the Neandertal Type Site in the Neander Valley of Germany, *Paper presented at: Annual Meeting of the Paleoanthropology Society*, Denver, March 19-20, 2002.

[11] I. Ovchinnikov, A. Götherström, G. Romanova, V. Kharitonov, K. Lidén, W. Goodwin, Molecular Analysis of Neanderthal DNA from the Northern Caucasus, *Nature*, Vol. 404, 2000, pp. 490-493.

[12] M. Krings, C. Capelli, F. Tschentscher, H. Geisert, S. Meyer, A. von Haeseler, K. Grossschmidt, G. Possnert, M. Paunovic, S. Pääbo, A View of Neandertal Genetic Diversity, *Nature Genetics*, Vol. 26, 2000, pp. 144-146.