Estimating Neandertal contribution to the Upper Paleolithic *H. sapiens* gene pool using genetic drift effect in branching process model

KRZYSZTOF A. CYRAN[1,2], MAREK KIMMEL[2]
[1]Institute of Informatics
Silesian University of Technology
Akademicka 16, 44-100 Gliwice
POLAND
krzysztof.cyran@polsl.pl

[2]Department of Statistics
William Marsh Rice University
6100 Main Street, 77005 Houston, TX
USA
kimmel@rice.edu

Abstract: - In the paper, there is considered the effect of the genetic drift, which could eliminate the hypothetical Neandertal mtDNA admixture from the Upper Paleolithic gene pool of anatomically modern humans. To model the demography, the slightly supercritical Markov's branching process (BP) based on the O'Connell model has been proposed. Relying on relatively fast convergence to the O'Connell's limiting properties it is possible to estimate the time of extinction of the Neandertals relatively to the time of the root of the mtDNA polymorphism of modern humans. The results of the study indicate that the maximum hypothetical contribution of Neandertal mtDNA which could be eliminated by the genetic drift at 0.05 significance level is about 12% for studying mtDNA record only. However, the expected value of the admixture, estimated to be about 3.9 % based on mtDNA record only, is reduced to around 2.3 % if additional nuclear genome data is utilized for Bayesian inference. Relevance of the paper lies in treating mtDNA-based studies as complementary approaches to those based on nuclear DNA sequenced by the Neandertal genome project.

Key-Words: scientific computations, Bayesian inferring, mitochondrial DNA, branching process, genetic drift, Neandertals-H. sapiens interactions

1 Introduction

The period of coexistence of Neandertals with the Upper Paleolithic *H. sapiens* in Europe and Western Asia is a basis for the intriguing problem about the interbreeding between the two (sub)species. Charles Darwin described the speciation as that “mystery of mysteries”, however nowadays, to the great extent, this process is in general explained [1]. The advances in understanding of how new species form are based on current views regarding the geographic context and genetic variation for speciation.

Recently more and more is known about how the barriers to gene flow could have evolved. Although the general mechanisms of generating new species are accepted, much remains to be learned in relation to particular speciation processes – such as for example, the problem of (non)existence of the gene barrier between anatomically modern humans and Neandertals. The synthesis in which the ecological, physiological, developmental, and genetic bases for population divergence can be fully integrated is still a matter of future, however, as presented in the paper, at least genetic results based on nuclear and mitochondrial DNA can be aligned, giving more accurate predictions relying one on the other.

Speciation is an important issue for understanding the birth of genus *Homo*, however it seems from the recent genetic evidence that this process did not proceed in relation to anatomically modern humans and Neandertals to the extent that would cause the irreversible gene flow barriers. Contrary to first inferences performed purely on mitochondrial DNA (mtDNA) [2, 3], the latest studies as well as the current paper, indicate that there was some interbreeding between the two populations and genetic record allows for quite accurate estimate of the size of admixture.

Whatever is the taxonomic relation between *H. sapiens* and *H. neanderthalensis*, the issue of interbreeding is at least as intriguing for understanding our ancestry, as the hypothetical, reconstructed physiognomy of Neandertals. There is visible a clear shift in reconstruction of Neandertal face from the earliest views, according to which Neandertals were resembling apes to the current opinions about Neandertal physiognomy resembling the modern human (Fig. 1).

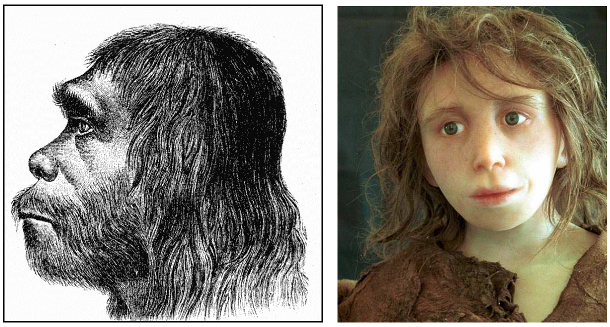

Fig. 1. The first (left) and the most recent (right) reconstruction of Neandertals. [Pictures in public domain]

This latter view is convincingly corroborating with the genetic evidence of interbreeding between the two populations some 30 000 – 50 000 years ago. After this period, that is after several thousand years of coexistence with *H. sapiens* in Europe, the Neandertals became extinct. (see Fig. 2 for sites where Neandertal fossils indicate the geographical range of their activity before extinction).

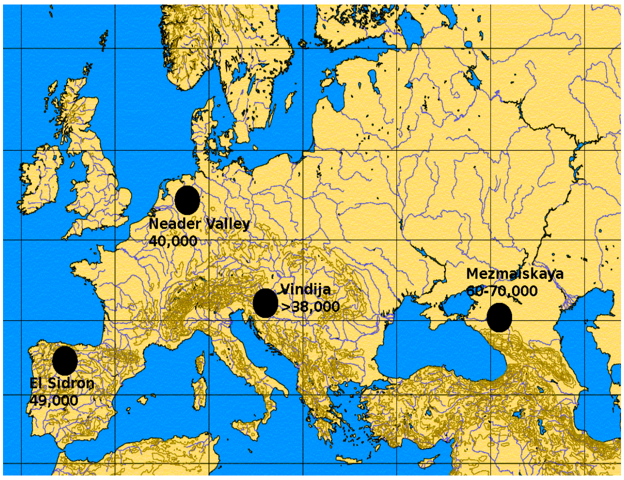

Fig. 2. Sites where Neandertal fossils have been found [based on Briggs et al., Science 325, 318 - 321 (2009)]

However, Neandertal nuclear genes seem to survive in the genome of modern humans, whereas, the corresponding putative admixture of mitochondrial DNA which they contributed to the Upper Paleolithic *H. sapiens* gene pool, has finally disappeared as a result of the genetic drift.

This latter fact is confirmed by many mtDNA-based studies, from the earliest [2, 3] to the most recent [4], which indicate that *H. neanderthalensis* is an outgroup in the mtDNA polymorphism of present-day humans (see Fig. 3). It is interesting that the first studies based on sequencing of the mtDNA retrieved from Neandertal fossils, which resulted in phylogenetic trees similar to this presented in Fig. 3, erroneously claimed an evidence of no interbreeding (for example [3]). While no interbreeding could be certainly the reason for the observed pattern, yet basing solely on such pattern, it cannot be excluded that the mtDNA polymorphism around 30 000 years ago contained also Neandertal sequences, and the currently observed pattern is the result of the genetic drift.

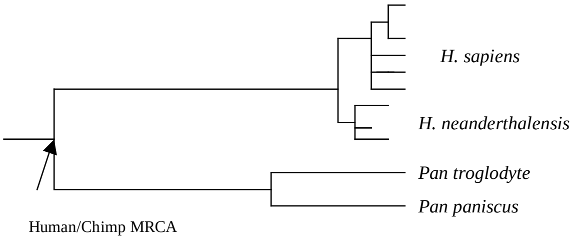

Fig. 3 mtDNA phylogeny of Neandertals and their close relatives. [based on Briggs et al., Science 325, 318 -321 (2009)]

This latter hypothesis corroborates with the recent studies performed within the Neandertal Genome Project (NGP), which are based on nuclear DNA. The loci having Neandertal ancestry in modern humans are estimated by Green et al. (see [5]) to constitute between 1 and 4% of the nuclear genome. This is significantly less than the amount reported by Wall et al. (see [6]), who found the signatures of the archaic ancestry in present day *H. sapiens* to be as abundant as 12% in the nuclear genome. However, the latter estimate can be explained by more ancient gene flows, probably from *H. erectus* (see Fig. 4).

2 Problem Formulation

Let us consider a family of slightly supercritical time-homogeneous Markov branching processes with the expected numbers of offspring per individual equal to $E(X_0) = 1 + \alpha/T + o(1/T)$ and the variance equal to $Var(X_0) = \sigma^2 + O(1/T)$, as $T \to \infty$. In such branching process, the expected number of progeny of any individual is slightly larger than the critical value (i.e slightly larger than 1). Let us assume that this branching process represents the abundance of Neandertal mtDNA evolving within the post-Neandertal modern human population (Fig. 4) after putative admixture indicated in Fig. 4 by the bold horizontal arrow.

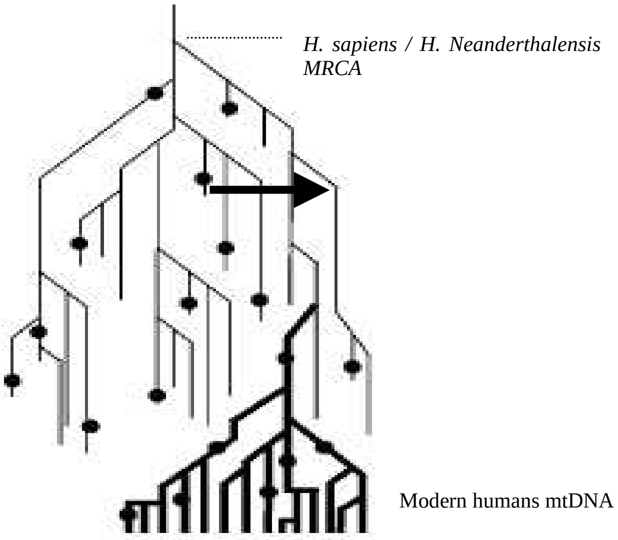

Fig. 4. Contribution of Neandertal mtDNA to Upper Paleolithic anatomically modern humans. The inter-population gene flow is indicated by the horizontal arrow. [Adapted from Cyran (2011)].

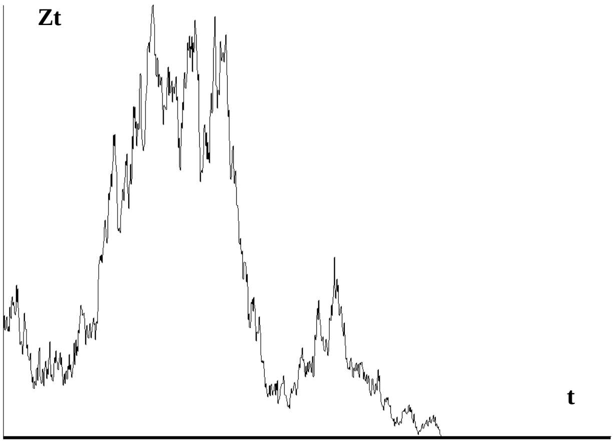

Fig. 5. Extinction of the branching process Z_t responsible for no observing Neandertal mtDNA sequences within contemporary modern humans mtDNA pool.

Observe that the upper Paleolithic *H. sapiens* mtDNA pool contained Neandertal sequences, subject to extinction by the genetic drift. Therefore, in the present day population no Neandertal mtDNA is found. The illustration of the possible time course of the number of Neandertal mtDNA within the modern human population is given in Fig.5. Observe that the final fate of these sequences is the extinction because of the genetic drift.

Hence, the following question arises: how abundant Neandertal mtDNA should be in Upper Paleolithic mtDNA pool of anatomically modern humans so that this amount is the most likely in the light of the recent results obtained within NGP for Neandertal nuclear genome.

3 Problem Solution

The beginning of this section follows the line of argument presented in Cyran and Kimmel (2005) [7], and Cyran (2011) [8], and serves as the introduction to the final Bayesian inference using prior probabilities based on NGP project results. The contribution of this paper is this final inference of the amount of Neandertal mtDNA admixture in Upper Paleolithic gene pool of *H. sapiens* using synthesis of Neandertal mtDNA and nuclear DNA data.

Denote the number of individuals in the process at time t by Z_t. As t, consider the time 30 000 years ago, when the Neandertals disappeared and their putative admixture in a gene pool of anatomically modern humans started to be a subject to the genetic drift with no further Neandertal contribution.

Assume further the duration T of the whole branching process to be 200 000 years. Cyran and Kimmel (2010) [9] have shown that such process is faithfully modeling the evolution of *H. sapiens* mtDNA from the MRCA (mtEve) dated to live around 175 000 years ago.

Under these assumptions, it follows that the time to coalescence of a pair of mtDNA sequences randomly picked from a sample of contemporary modern humans, denoted as T_2, is roughly equal to 150 000 years. These values are based on results of presented by Cyran and Kimmel in [9], provided that the time to the most recent ancestor of modern humans and Neandertals is around 500 000 years ago (compare with Briggs et al. 2009 [4]). The mtDNA data used for the inference was taken from the paper by Green et al. (2008) [10].

Assuming the duration of the average population to be 20 years, the times *t* and *T* expressed in the units of number of generations are equal to: $t = 1\,500$, and $T = 10\,000$, respectively.

Based on [11] (see also Cyran and Kimmel, 2010 [9]), the Wright – Fisher model is equivalent to the branching process model with the number of offspring having the Poisson distribution. Under this assumption, the distribution of the time to coalescence of a pair of sequences in Wright –Fisher model is identical to the O'Connell (1995) (see [12]) distribution derived for slightly supercritical branching processes. Based on analyses performed by O'Connell (1995) and Cyran and Kimmel (2010), it follows that the feasible value of α is 10 and hence, $E(X_0) = \sigma^2 = 1.001$ for Poisson offspring number distribution with $\alpha = 10$ and $T = 10\,000$.

In such process, the probability of non-extinction of a lineage descending from a single Neandertal mtDNA, $P(Z_t > 0 \mid Z_0 = 1)$ is given by (O'Connell 1995) (see [11], compare with [8])

$$P(Z_t > 0 \mid Z_0 = 1) \sim \frac{2\alpha}{\sigma^2 T}\left(1-\exp\left(-\alpha\frac{t}{T}\right)\right)^{-1}, \quad as \quad T \to \infty \qquad (1)$$

Therefore, the probability of extinction of such lineage is equal to

$$P(Z_t = 0 \mid Z_0 = 1) = 1 - P(Z_t > 0 \mid Z_0 = 1) \qquad (2)$$

Treating the extinction of particular lineage as an event independent from the extinction of other lineages, the probability of extinction of lineages started by *n* hypothetical mtDNA sequences present in the Upper Paleolithic *H. sapiens* gene pool is given by

$$P(Z_t = 0 \mid Z_0 = x) = 1 - P(Z_t > 0 \mid Z_0 = 1)^x \qquad (3)$$

The graph of the resulting likelihood function $P(Z_t = 0 \mid Z_0 = x)$ as a function of *n* is given in the Fig. 6. Solving (3) for *n*, yields

$$n = \frac{\ln(P(Z_t = 0 \mid Z_0 = n))}{\ln(1 - P(Z_t > 0 \mid Z_0 = 1))} \qquad (4)$$

Using the values specified in the assumptions, it follows that $P(Z_t > 0 \mid Z_0 = 1) \approx 2.57 \times 10^{-3}$.

In order to compute the maximum admixture not contradicting the mtDNA record (see Fig. 3) at 0.05 significance level, lat us assume the probability $P(Z_t = 0 \mid Z_0 = n)$ to be 0.05. Hence, from (4) it follows that $n = 1\,166$ individuals (because of the equivalence of the Wright Fisher reproduction model and the branching process reproduction model used by O'Connell, this number constitutes directly the short-term inbreeding effective population size).

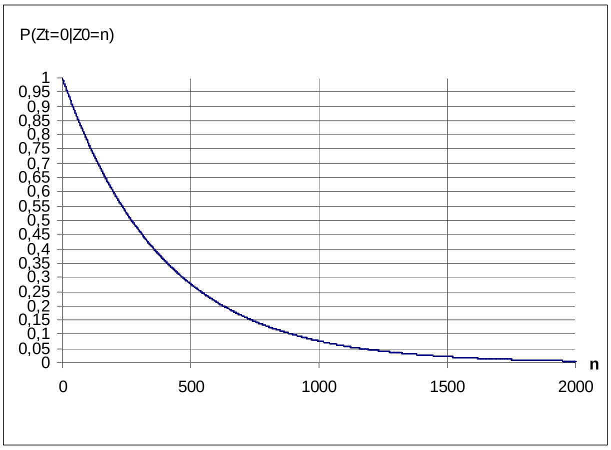

Fig. 6. The likelihood of the $P(Z_t = 0 \mid Z_0 = n)$ as a function of *n* [graph based on results obtained by Cyran (2011)].

Using Bayesian rule, it follows that the posterior probability $P(Z_0 = n \mid Z_t = 0)$ is given by

$$P(Z_0 = n \mid Z_t = 0) = \frac{P(Z_0 = n)P(Z_t = 0 \mid Z_0 = n)}{P(Z_t = 0)}, \qquad (5)$$

and its shape is modeled by the prior distribution.

Relying on the estimate of the census population size of modern humans around 30 000 years ago to be at least 500 000, the census population size of females active in reproduction at that time was at least 100 000. This is based on the assumption that the population is composed of the same number of males and females and provided that, on average, one out of 2.5 females in a population is reproductively active.

Note, that if the actual variance of the number of offspring σ^2 is around 10 (what corresponds to standard deviation in the number of progeny about 3 – which seems feasible), then, the short-time inbreeding effective population size of anatomically modern human females living 30 000 years ago, N_e is about 10 000. Assuming particular shape of the prior probabilities $P(Z_0 = n)$ over the range [0, 10 000], and appropriate scaling factor $P(Z_t = 0)$

which is independent of n, it is possible to compute from (5) the distribution of $P(Z_0 = n \mid Z_t = 0)$. Having this distribution, the expected value $E(Z_0 \mid Z_t = 0)$ can be obtained as

$$E(Z_0 \mid Z_t = 0) = \sum_x nP(Z_0 = n \mid Z_t = 0) \qquad (6)$$

In what follows, four different prior distributions are considered. At first, consider the uniform distribution of the prior probabilities $P(Z_0 = n)$ over the range n ∈ [0, 10 000], what corresponds to the situation when no additional data about the putative admixture is available. The posterior probability distribution $P(Z_0 = n \mid Z_t = 0)$ in this case is (to the scaling factor) identical to the likelihood function $P(Z_t = 0 \mid Z_0 = n)$ (compare Fig. 6 and Fig. 7).

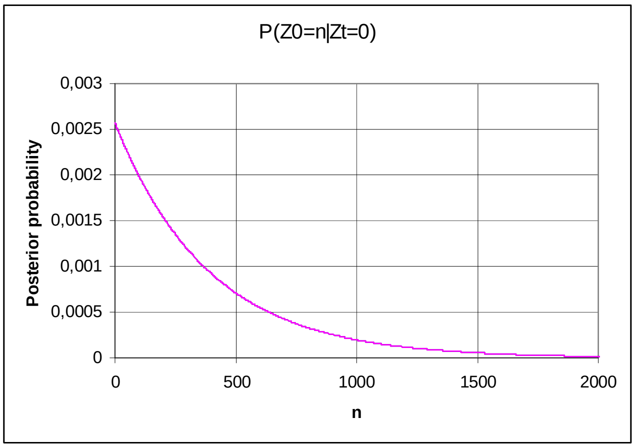

Fig. 7. Posterior probability $P(Z_t = 0 \mid Z_0 = n)$ for uniform prior distribution U_1 (0, 10 000).

In this situation, it follows that $E(Z_0 \mid Z_t = 0) = 388$ individuals (effective population size of Neandertal mtDNA sequences in the Upper Paleolithic *H. sapiens* mtDNA gene pool). To compute the expected value of the Neandertal mtDNA admixture in a gene pool, let us divide $E(Z_0 \mid Z_t = 0) = 388$ by $N_e = 10\,000$. This results in the expected admixture of about 3.9 %. Similarly, to compute the maximum admixture non contradicting the mtDNA testimony at significance level 0.05, let us divide $n = 1\,166$ by $N_e = 10\,000$. This results in the maximum hypothetical admixture of about 11.7 %.

In order to incorporate the additional information available based on nuclear DNA data obtained within the NGP, the prior probability distribution will use the information that the putative Neandertal component in nuclear genome of contemporary humans is between 1 and 4 % [5]. Since nuclear genes, due to recombination, are evolving to great extent independently, therefore, on average, they are not prone to the effect of genetic drift. Hence, it is likely, that similar component of nuclear DNA was present in Upper Paleolithic anatomically modern human nuclear genome. However, under Wright-Fisher model, this implies, that at that time the mtDNA gene pool contained the same percentage of Neandertal sequences.

Therefore, for the model considered in the paper, it is feasible to account for this additional information by assuming appropriate priors. Three such prior distributions are considered: the uniform distribution U_2 (100, 400) as well as two normal distributions: N_1 (250, 150), and N_2 (250, 75). The resulting posterior probability distributions are given in Fig. 8, Fig. 9, and Fig. 10, respectively.

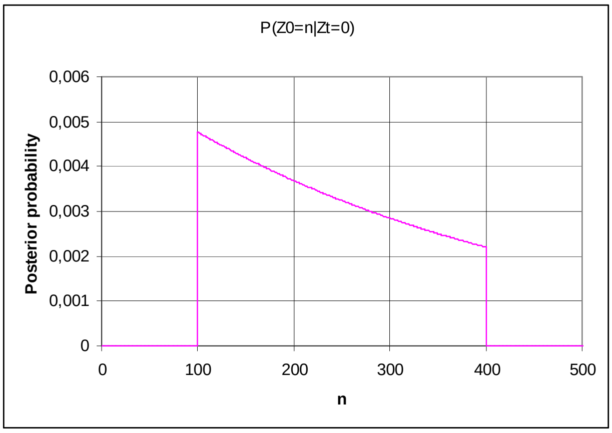

Fig. 8. Posterior probability $P(Z_t = 0 \mid Z_0 = n)$ for uniform prior distribution U_2 (100, 400).

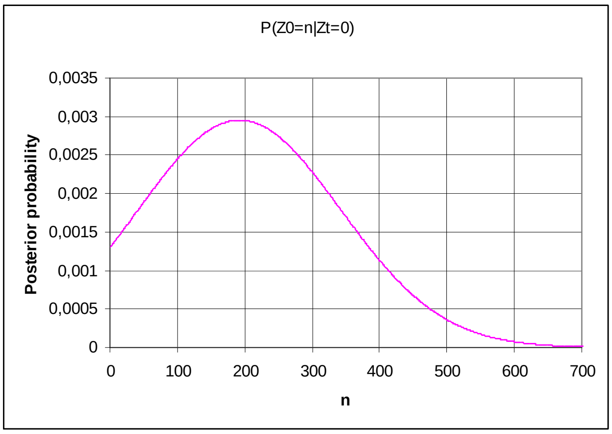

Fig. 9. Posterior probability $P(Z_t = 0 \mid Z_0 = n)$ for normal prior distribution N_1 (250, 150).

For short-term inbreeding effective population size equal to 10 000, these distributions expressed in the units of percentage of the admixture, correspond to distributions: U_2 (1, 4), N_1 (2.5, 1.5), and N_2 (2.5, 0.75). The normal distributions are chosen in such a way, that the proposed by nuclear DNA studies limits (1% and 4%) are in the distance of one and two standard deviations for N_1, and N_2 respectively.

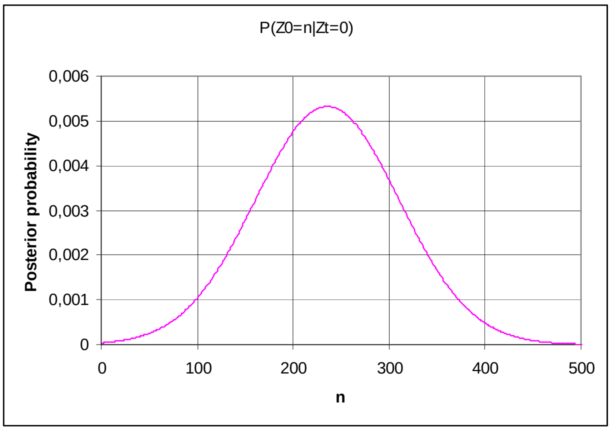

Fig. 10. Posterior probability $P(Z_t = 0 \mid Z_0 = n)$ for normal prior distribution N_1 (250, 75).

4 Conclusion

The estimate of the putative Neandertal mtDNA admixture obtained solely on the basis of mtDNA record analyzed in slightly supercritical branching process model is corroborating with the recent results obtained based on nuclear DNA sequenced from the Neandertal fossils in the Neandertal genome project (see Green et al. 2010, [5]). Therefore, it is likely that these data do not contradict each other – rather they can be used together in a single model which is able to make use of both of them. In the paper, such model is presented. The inferences about the putative admixture of Neandertal mtDNA are utilizing not only mtDNA data, but also the results from nuclear analysis, which serves in a model as a supply of the feasible prior distributions. Interestingly, no mater what type of feasible prior distribution is used, the expected value of putative admixture is very similar and ranges from 2.2 % for prior N_1, through 2.3 % for prior U_2, up to 2.35 % for prior N_2.

This shows, that the mean of prior distribution based on nuclear genome data (2.5 %) has stronger impact on the obtained results than the mean of posterior distribution 3.9 %, provided non-informational prior U_1. Making use of both, the mtDNA and nuclear DNA data, as presented in the proposed model makes the Bayesian inference about putative mtDNA admixture more reliable than relying solely on nuclear data (covered in prior distribution) or solely in mtDNA record (covered in posterior distribution with non-informational priors). In fact, the probability that the observed pattern in contemporary mtDNA contains no Neandertal sequences has increased 13 to 14 times by including the meaningful priors U_2, N_1, or N_2 as compared to non-informative prior $U_{1.}$

Acknowledments

The research leading to these results has received funding from the PEOPLE Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement no. 298995, which supported research work of Krzysztof A. Cyran.

Marek Kimmel was supported by grant NCN no: NN519579938.

References:

[1] Feder J. L. (2009) The Mystery of Speciation, lecture VII presented at III STOQ International Conference on Biological Evolution: Facts and Theories. A Critical Appraisal 150 Years After "The Origin of Species", Rome, March 3-7 2009

[2] Krings et al. (1997) Neanderthal DNA sequences and the origin of modern humans. Cell 90, 19-30.

[3] Krings et al. (1999) DNA sequence of the mitochondrial hypervariable region II from the Neanderthal type specimen, Proc. Natl. Acad. Sci. USA 96, 5581-5585.

[4] Briggs A.W. et al. (2009) Targeted retrieval and analysis of five Neandertal mtDNA genomes, Science 325, 318 –321.

[5] Green R.E. et al. (2010) A draft sequence of the Neandertal genome, Science 328, 710-722.

[6] Wall J.D., Lohmueller K.E., Plagnol V. (2009), Detecting ancient admixture and estimating demographic parameters in multiple human populations, Mol. Biol. Evol. 26, 1823.

[7] 7. K. Cyran, M. Kimmel (2005) Interactions of Neanderthals and modern humans: what can be inferred from mitochondrial DNA, Mathematical Biosciences and Engineering, 2(3), 487-498.

[8] Cyran K.A. (2011) Artificial Intelligence, Branching Processes and Coalescent Methods in Evolution of Humans and Early Life, Publishers of the Silesian University of Technology, Gliwice.

[9] Cyran K.A., Kimmel M. (2010) Alternatives to the Wright-Fisher model: The robustness of mitochondrial Eve dating, Theoretical Population Biology, 78(3), 165-172.

[10] Green R.E. et al. (2008) A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing, Cell 134, 416-426.

[11] Haccou P., Jagers P., Vatutin V.A. (2005) Branching processes. Variation, Growth, and Extinction of Populations, Cambridge Univ. Press.

[12] O'Connell N. (1995) The genealogy of branching processes and the age of our most recent common ancestor, Adv. Appl. Prob. 27, 418-442.