Percolation model for the existence of a mitochondrial Eve

Armando G. M. Neves
Carlos H. C. Moreira
UFMG - Departamento de Matemática
Av. Antônio Carlos, 6627 - Caixa Postal 702
30123-970 - B. Horizonte - MG

March 15, 2005

Abstract

We look at the process of inheritance of mitochondrial DNA as a percolation model on trees equivalent to the Galton-Watson process. The model is exactly solvable for its percolation threshold p_c and percolation probability critical exponent. In the approximation of small percolation probability, and assuming limited progeny number, we are also able to find the maximum and minimum percolation probabilities over all probability distributions for the progeny number constrained to a given p_c. As a consequence, we can relate existence of a mitochondrial Eve to quantitative knowledge about demographic evolution of early mankind. In particular, we show that a mitochondrial Eve may exist even in an exponentially growing population, provided that the average number of children per individual is constrained to a small range depending on the probability p that a newborn child is a female.

1 Introduction

Percolation originally meant the passage of water (or some other fluid) through the soil (or some other porous medium). In this paper we use the same word in the generalized sense probabilists gave to it [3], i.e. a class of models to describe fluid percolation through a porous medium, in which the medium is described as an infinite set of points (vertices) linked by edges. Edges may be either open or closed to the passage of the fluid according to some statistical law. Typically, but not in this paper, vertices are considered to be the points in the d-dimensional square lattice Z^d and edges are open with probability p and closed with probability $1-p$,

independently. Besides fluid percolation, percolation models have also found other applications, such as the spread of diseases in a population or disordered electrical networks [9].

One is usually interested in calculating the percolation probability, defined as follows. Choose some vertex O to be the origin and some configuration of open and closed edges. Consider the cluster of all vertices connected to the origin by some path of open edges. We say that this configuration percolates if the cluster of vertices connected to O is infinite. The percolation probability is defined as the probability $\theta(p)$ of the percolation event over all configurations.

Percolation models attract so much attention because $\theta(p)$ usually exhibits a "second-order phase transition": there exists a percolation threshold p_c such that $\theta(p)$ is strictly zero for $p < p_c$ and positive for $p > p_c$. They look thus like toy models for Statistical Mechanics and in fact many techniques developed in percolation may be adapted to Statistical Mechanics and vice-versa. But percolation models may be not at all trivial [9].

As the reader is going to see in a while, the percolation model we shall develop for the existence of a mitochondrial Eve is one in which the medium is a genealogical tree, the edges are parental relationships and the "fluid" is mitochondrial DNA (mtDNA). Before we say more on that, let us briefly review the very basic facts about mtDNA and the mitochondrial Eve.

Although most of the genetic information in higher animals is located at cells' nuclei, some DNA may be found in the subcellular organelles called mitochondria. These structures are present in large numbers in nearly every cell and play a key role in metabolism. mtDNA is very short; in humans it consists in only 16,569 base pairs carrying the information for 37 genes. But mtDNA of humans and many other species has been the object of recent intensive research [1], both for its availability (because mitochondria are so numerous) and peculiar inheritance mechanism. Unlike nuclear DNA, which is inherited in equal parts from mother and father, mtDNA is inherited only from the mother. So, in the absence of mutations, the mtDNA of an individual would be identical to the mtDNA of a single ancestor out of his or her 2^n ancestors n generations before, namely the mother of the mother ... of his or her mother.

This simple feature allows one to use mtDNA comparisons among living individuals to look far back in time and draw conclusions about the separation of subpopulations in one species or the speciation process, in which several extant species may descend from one extinct species [1].

An experiment performed in the late 80's brought press popularity to mtDNA. By examining mtDNA of 147 living humans and taking into account mutations, Cann, Stoneking and Wilson asserted in [5] that mtDNA of all living humans could be described as mutations in the mtDNA of a single woman. As we would all be her descendants, this woman was called the *mitochondrial Eve*. By using known

mutation rates and exploiting geographical correlations, it could be inferred that the mitochondrial Eve has lived in Africa more or less 200,000 years ago.

The mitochondrial Eve should not be confused with the biblical Eve. Unlike the latter, the mitochondrial Eve is not supposed to be the only woman living at her time. Many other men and women contemporary to her have probably left traces of their nuclear DNA in modern humans, see e.g. [7] or [6] for proofs of this in the case of a population with a fixed size. But they left no trace of their mtDNA.

The time and the place in which the mitochondrial Eve lived are considered a strong evidence for the *Out of Africa* model for the origin of our own species [16]. This model proposes that modern humans, *Homo sapiens*, evolved in Africa and subsequently colonized the rest of the world replacing archaic forms such as the *Homo neanderthalensis* and the *Homo erectus*, without mixing with them. The competing *Multi-regional Evolution* model [15] suggests instead that modern humans evolved from archaic forms concurrently in different regions of the world, with occasional genetic flow among regions, necessary to preserve uniqueness of our species.

In order to explain why all mtDNA lineages stemming from women coeval to the mitochondrial Eve were extinct, Brown proposed in [4] that a severe bottleneck, in which human population dropped to only a few individuals, must have existed after the mitochondrial Eve. This is a very strong hypothesis, considering that human population, at least in historical times, has been steadily growing, and that achievement of important technological developments in prehistory made possible for humans to spread all over the world. In [2] Avise, Neigel and Arnold argued that a prehistorical population bottleneck is not strictly necessary by showing that stochastic mtDNA lineage extinction can be rapid enough even in *stable-sized* populations.

Our main purpose in this paper is to show that the existence of a mitochondrial Eve can be explained even if the population *grows exponentially*, provided that the growth rate is not too large. Although our model and formalism are similar to the ones in [2], we show that one solution discarded there is biologically plausible. We are also able to generalize their methods to arbitrary progeny distributions and to display the model explicitly in the language of an exactly-solvable percolation problem.

2 Percolation model for mtDNA inheritance

The assumptions in our model for mtDNA inheritance are the following:

(A1) Generations are nonoverlapping.

(A2) The numbers of children for each individual are statistically independent and identically distributed random variables assuming value $r \in \{0, 1, 2, \ldots\}$ with probability Q_r. The values for the Q_r's are time- and population size-independent.

(A3) A newborn child is a female with probability p and a male with probability $1-p$. The value for p is also time- and population size-independent.

(A4) There always exist males enough to mate with all females.

Assumption (A1) grants some formal simplicity to the model. (A2) disregards interactions that might come from a number of different sources such as fertility correlations among members of a family, competition (for food supplies, mating partners, etc), cooperation and geographical aspects. One natural attempt to improve the model would be to make the Q_r's dependent on the total population, thus accounting for saturation effects. Although feasible in computer simulations, that would ruin the linearity on which our theoretical analysis relies. The time independence of the Q_r's is also questionable because it disregards the effects of climatic changes. However this assumption may be adequate over an initial period of time long enough to produce most of the lineage extinctions. (A3) adds some generality to the model with respect to the one in [2], in which $p = 1/2$. Usually p varies from species to species and even for modern humans $p \approx 0.488$, strictly less that $1/2$ [12]. This fact will be of quantitative relevance in our results. (A4) is assumed, since we do not keep track of the male population. It is a reasonable assumption because males from all concurrent mitochondrial lineages may participate in any single one, without interfering in the mitochondrial inheritance. Also, even for p only slightly less than $1/2$, the results in this paper will show that male extinction is much less probable than female extinction for humans.

Consider now the genealogic tree of an ancestral woman constructed according to the above assumptions. By genealogic tree we mean the graph obtained by considering as vertices the ancestral herself, located at the origin O and all her descendants after an *infinite* number of generations, drawing edges joining each father or mother to their children of either sex. Define as *open* any edge linking a mother to her children and *closed* all other edges. It is clear that the mtDNA lineage of the woman will survive if and only if the configuration of open edges percolates. Therefore mtDNA inheritance may be posed as a problem of edge percolation in a tree graph, in which edges are open with probability p and closed with probability $1-p$. Unfortunately, unlike most percolation models, edges may not be statistically independent because any child with both parents in the graph must appear twice, separately linked to each of the parents. An example is shown in figure 1(a).

In order to overcome the dependence problem, we define the female genealogic tree (FGT) as the tree obtained after stripping the genealogic tree of all descendants of male individuals, see figure 1(b). In the FGT all edges are open and statistically independent. Percolation in the complete genealogic tree is equivalent to the corresponding FGT being infinite generations long.

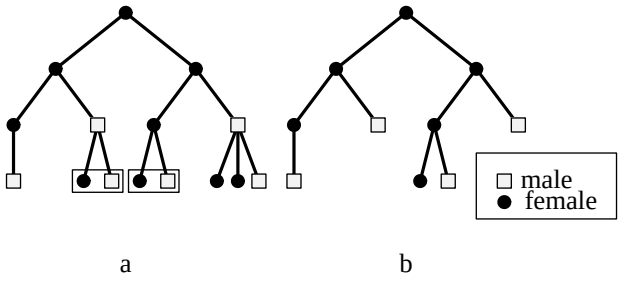

Figure 1: (a) Example of a genealogic tree of the type considered in this paper. The individuals enclosed in the rectangle have both their father and mother in the tree, exemplifying the possibility of statistically dependent edges.
(b) The female genealogic tree corresponding to the complete tree in (a).

If we denote as q_r, $r = 0, 1, 2, \ldots$, the probability that an individual has r children of female sex, then

$$q_r \;=\; \sum_{k=r}^{\infty} Q_k \left(\begin{array}{c} k \\ r \end{array} \right) p^r \, (1-p)^{k-r} \; . \tag{1}$$

The process of growing the FGT under the above assumptions, each female vertex originating r branches with probability q_r, is known as the Galton-Watson processes [10]. It was originally introduced to study the problem of extinction of family names.

Let $\theta_n(q_0, q_1, \ldots)$ be probability that the FGT is at least n generations long. We denote by $\overline{\theta}_n \equiv 1 - \theta_n$ the probability that the FGT is long $n-1$ generations or less. The end of the FGT in $n-1$ or less generations may happen either if the woman in its origin has no daughters (with probability q_0), or if she has b daughters, $b = 1, 2, \ldots$, each of them having a female genealogy at most $n-2$ generations long (with probability $q_b \, \overline{\theta}_{n-1}^{\,b}$). Thus,

$$\overline{\theta}_n \;=\; S(\overline{\theta}_{n-1}) \; , \tag{2}$$

where

$$S(x) \;=\; q_0 \,+\, q_1 \, x \,+\, q_2 \, x^2 \,+\, \ldots \tag{3}$$

is the generating function for the probability distribution of the number of daughters. The initial condition to be used in conjunction with (2) is $\overline{\theta}_1 = q_0$.

The normalization of the q_r's implies that 1 is a fixed point of S. Since S is non-decreasing with all derivatives non-decreasing in $[0, 1]$, it will have a second fixed point $\overline{\theta} \in [0, 1)$ if and only if $S'(1) > 1$. There are no other fixed points in $[0, 1]$. Regarding attractiveness [8], we may have three different regimes:

(i) If $S'(1) < 1$, then 1 is the only fixed point and it is attractive.

(ii) If $S'(1) > 1$, then 1 is a repulsive fixed point, whereas $\overline{\theta}$ is attractive.

(iii) If $S'(1) = 1$, then 1 is again the only fixed point in $[0,1]$ and it is *weakly* attractive.

If $\Theta \equiv \lim_{n\to\infty} \theta_n$ is the percolation probability, then

$$\Theta \;=\; \begin{cases} 0, & \text{if } \; S'(1) \leq 1 \\ \theta, & \text{if } \; S'(1) > 1 \end{cases} \;, \tag{4}$$

where $\theta \equiv 1 - \overline{\theta} \in (0,1]$. By writing $S'(1) = \sum_{r=1}^{\infty} r\, q_r$ and using (1) we get that condition $S'(1) > 1$ may be written as $p > p_c$. The percolation threshold p_c is

$$p_c \;=\; \frac{1}{\overline{N}} \;, \tag{5}$$

with

$$\overline{N} \;=\; \sum_{r=1}^{\infty} r\, Q_r \tag{6}$$

meaning the average number of children (of either sex) per woman. Regime (i) is then the subcritical percolation regime (i.e. $p < p_c$), (ii) is supercritical ($p > p_c$) and (iii) is critical ($p = p_c$).

The value of θ in terms of the q_r's may be obtained by solving $S(\overline{\theta}) = \overline{\theta}$. For $p > p_c$ and p close enough to p_c, we introduce what we shall call *small percolation probability approximation*. In this case, an approximate value θ_a for θ is obtained by replacing S by its Taylor polynomial of degree 2 around 1. We find

$$\theta \;\approx\; \theta_a \;=\; \frac{p - p_c}{p_c\, p^2 \sum_{r=2}^{\infty} \left(\begin{array}{c} r \\ 2 \end{array} \right) Q_r} \;. \tag{7}$$

If the critical exponent β for the percolation probability is defined by $\theta \sim (p - p_c)^{\beta}$, then (7) gives $\beta = 1$. From the non-negativity of S''', it follows that θ_a is actually a lower bound for θ.

The speed of convergence of θ_n to Θ, a useful parameter for simulation purposes, follows from the mean value theorem of Calculus. In cases (i) and (ii) $\theta_n - \Theta \;\sim\; e^{-n/\xi}$ for large n, where, as usual, the *correlation time*

$$\xi = -1/\ln[S'(1 - \Theta)] \tag{8}$$

diverges when $p \to p_c$. In the critical case (iii) exponential convergence is replaced by a much slower power law: $\theta_n \;\sim \frac{2}{S''(1)\, n}$ for large n.

3 Demographic considerations

Let W be the number of women coeval to the mitochondrial Eve. If different lineages may be considered as independent, then the number r of lineages remaining after n generations is a binomially distributed random variable with

$$\text{Prob}\{r = m\} = \left(\begin{array}{c} W \\ m \end{array} \right) \theta_n^m (1-\theta_n)^{W-m} \, . \tag{9}$$

This approach was used in [2], where the authors concentrated on the probability Π_n for the survival of two or more lineages after n generations. For some specific progeny distributions they found that in the supercritical regime Π_n tends to a positive value as $n \to \infty$; they considered this to be incompatible with a mitochondrial Eve. The subcritical regime was also discarded because it leads to quick extinction. Instead, the critical regime was selected because with $W = 1,000 \sim 10,000$ it leads to Π_n approaching zero in $n \approx 10^4$ generations. Such values for W and n are well within the range expected by geneticists and paleontologists. For $p = 1/2$, as used in [2], the critical regime yields $\overline{N} = 2$ and can only account for a stable-sized population.

We shall now argue that for a range of values of $\overline{N}$, the supercritical regime also provides a biologically plausible solution for the existence of a mitochondrial Eve in a growing population. We shall use $W = 5,000$ and $n = 10^4$, as suggested for example in [14]. We also use $p = 0.488$, obtained from the standard figure of 105 male births per 100 female births, the modern human *sex ratio at birth* [12], assuming that this ratio can be extrapolated to the times of early mankind.

From (9) we get that the expected number of surviving lineages after n generations is $W\theta_n$. To be consistent with the existence of a mitochondrial Eve as an event with a not too small probability, this number must not be much smaller or much larger than 1. For illustration purpose, we take it between $1/2$ and 2, implying θ_n between $1/(2W)$ and $2/W$. In order to estimate the range of values for $\overline{N}$ consistent with that, we first replace θ_n for θ_a given in (7). This approximation is valid as long as two conditions are fulfilled. Firstly, the number of generations n must be so large that θ_n is close to θ. Finally, the small percolation probability approximation must hold in order that θ is close to θ_a. The latter is true, because θ is of order W^{-1}, a small number if $W = 5,000$. The former, $\theta_n \approx \theta$, will be justified soon.

Also, for practical purposes we may truncate the series in the right-hand side of (7) at some order M (limited progeny assumption). The problem of estimating the maximum and minimum of θ_a for given $\overline{N}$ then becomes a standard linear programming problem, in which we must respectively minimize or maximize the linear function $\sum_{r=2}^{M} \left(\begin{array}{c} r \\ 2 \end{array} \right) Q_r$ under the constraints $\sum_{r=1}^{M} r\, Q_r = \overline{N}$, $\sum_{r=1}^{M} Q_r = 1$ and $0 \leq Q_r \leq 1$ for $r = 0, 1, \ldots, M$.

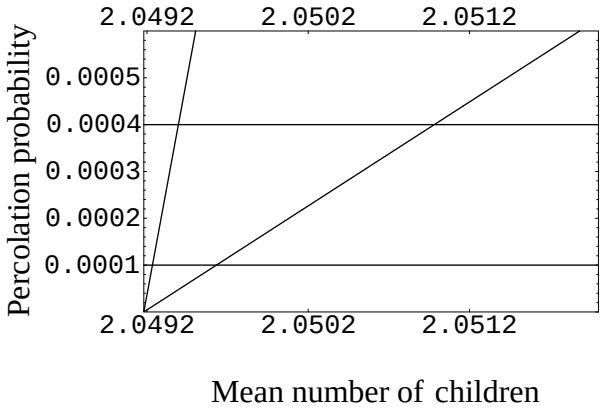

Figure 2: Maximum and minimum values for θ_a as functions of $\overline{N}$ with $M = 10$. Horizontal lines correspond to the values $1/(2W)$ and $2/W$ for θ_a.

In figure 2 we show a plot of the maximum and minimum values for θ_a as functions of $\overline{N}$, taking $M = 10$. The values for $\overline{N}$ consistent with existence of a mitochondrial Eve as a not very rare event are thus such that the maximum θ_a lies over $1/(2W)$, which yields $\overline{N} > 2.0492$, and the minimum lies below $2/W$, which yields $\overline{N} < 2.0510$. For $\overline{N}$ in this range the probability that only one lineage survives is approximately constant around 30% and the population will grow exponentially at a non-negligible rate.

The value $M = 10$ for the maximum number of children is arbitrary. However it turns out that the maximum values for θ_a become independent of M for $M \geq 3$ and are obtained with $Q_r \neq 0$ only for $r = 2$ or 3. On the other hand, the minimum values for θ_a do depend on M and tend to 0 as $M \to \infty$. Nonetheless, this minimum value is attained when $Q_r = 0$, $r = 0, 1, 2, \ldots, M-1$ and $Q_M = \overline{N}/M$, which is quite an artificial probability distribution for the number of children. We believe that imposing realistic conditions on the Q_r's would constrain even more the average number of children consistent with existence of a mitochondrial Eve.

Results plotted in figure 2 assume, as already mentioned, that θ_n can be approximated by its limit θ when $n \to \infty$. In other words, the results will be valid provided $n \gg \xi$, where the correlation time ξ is given by (8). Although we do not know how to give precise bounds to ξ independent of the progeny distribution, as we did for θ_a, we noticed that it is fairly constant when $\overline{N}$ is fixed and the progeny distribution varied. In particular, for the largest value $\overline{N} = 2.0510$ compatible with the existence of the mitochondrial Eve, ξ is of the order of 1,100 generations. As geneticists assume that the mitochondrial Eve lived more or less 10,000 generations ago, the approximation is justified for $\overline{N} = 2.0510$. On the other hand ξ diverges when $\overline{N}$ approaches $1/p$. This means that the range of values for $\overline{N}$ compatible with

the Eve should be extended down to $1/p$.

4 Conclusions and perspectives

We have proposed and solved a percolation model for mtDNA inheritance and lineage extinction compatible with the existence of a mitochondrial Eve in a growing population. According to this model, a mitochondrial Eve is likely only if the mean number of children per individual is constrained to a narrow range. The exact values for this range depend critically on the number of women W at the time of the mitochondrial Eve, but not on the number of generations since then. Of course, our model should be thought of as a simplified version of reality. We think it is a valuable starting point for further work, including simulations of more realistic models [13].

One tacit assumption in our analysis, supported by biologists, is that all W original mtDNA lineages are equally fit, i.e. there is no natural selection acting on lineage sorting. It also follows from our results that within the range of values for $\overline{N}$ in which a mitochondrial Eve is likely, there is a probability of approximately 70% that the number of surviving mtDNA lineages is different from 1. As this number can also be 0, we may explain extinction of other hominid species which had existed for some periods, if they had similar demography as our own species.

Returning to the two competing models on human origins, we should say that existence of an African mitochondrial Eve only proves that a sizeable part of early mankind did originate in that continent, but not the whole of it. mtDNA lineages originated somewhere else may have simply become extinct, which in our model appears as a highly probable event even in a growing population. We mean that existence of an African mitochondrial Eve does not rule out the multi-regional evolution model.

While preparing this paper, we discovered another application for our methods. Unlike most other animals, in which the sex of the offspring is genetically determined (X and Y chromosomes in mammals, for example), the sex of the offspring in some living reptile species depends on the temperature during egg incubation. Although a definitive proof still lacks, Miller, Summers and Silber [11] conjecture that the same sex determination by temperature might hold for dinosaurs. Should that be true, a predominance of males could have arisen as a consequence of worldwide climate change after the impact of a large meteor. By numerically solving a mathematical model based on differential equations, they show that in this case dinosaur populations would decrease with time. If the sex skew were not too severe or lasted only for a short time, populations would start growing again.

We notice here that our model can easily account for this phenomenon. In fact, an increase in male proportion is equivalent to lowering the value of p, while p_c is

held constant. On the other hand, the analysis carried out in [11] seems to disregard the fact that even in the worst cases, populations would increase again after some time. Here we see an important difference between their population modelling with differential equations and our "discrete" model. Differential equations are deterministic because they use the mean behavior for all individuals in the population. Extinction occurs in their model because of decrease in population to less than one individual. On the other hand, our model accounts for statistical fluctuations. If $2 < \overline{N} < 1/p$, population will increase in average, but extinction will still happen with positive probability due to statistical fluctuations. In other words, modelling populations with differential equations is analogous to mean-field theories in Statistical Mechanics, which we know can lead to wrong results.

We finally note that as Monte Carlo simulations are being increasingly used in Genetics, it is important to use the methods from the Physics of critical phenomena to better understand results derived from these simulations.

References

[1] J. C. Avise. Mitochondrial dna and the evolutionary genetics of higher animals. *Phil. Trans. R. Soc. Lond. B*, 312:325, 1986.

[2] J. C. Avise, J. E. Neigel, and J. Arnold. Demographic influences on mitochondrial dna lineage survivorship in animal populations. *J. Molec. Evol.*, 20:99, 1984.

[3] S. R. Broadbent and J. M. Hammersley. Percolation processes i. crystals and mazes. *P. Camb. Philos. Soc.*, 53:629, 1957.

[4] W. M. Brown. Polymorphism in mitochondrial dna of humans as revealed by restriction endonuclease analysis. *Proc. Natl. Acad. Sci USA*, 77:3605, 1980.

[5] R. L. Cann, M. Stoneking, and A. C. Wilson. Mitochondrial dna and human evolution. *Nature*, 325:31, 1987.

[6] J. T. Chang. Recent common ancestors of all present-day individuals. *Adv. Appl. Probab.*, 31:1002, 1999.

[7] B. Derrida, S. C. Manrubia, and D. H. Zanette. Statistical properties of genealogical trees. *Phys. Rev. Lett.*, 82:1987, 1999.

[8] R. L. Devaney. *A First Course in Chaotic Dynamical Systems*. Addison Wesley, Reading, 1992.

[9] G. Grimmett. *Percolation*. Springer-Verlag, Berlin, 2nd edition, 1999.

[10] T. E. Harris. *The Theory of Branching Processes*. Dover, New York, 1989.

[11] D. Miller, J. Summers, and S. Silber. *Fertility and Sterility*, 81(4):954, 2004.

[12] C. Newell. *Methods and models in Demography*. John Wiley and Sons, Chichester, 1988.

[13] D. L. T. Rohde, S. Olson, and J. T. Chang. Modelling the recent common ancestry of all living humans. *Nature*, 431:562, 2004.

[14] N. Takahata. Allelic genealogy and human evolution. *Mol. Biol. Evol.*, 10(1):2, 1993.

[15] A. G. Thorne and M. H. Wolpoff. The multiregional evolution of humans. *Sci. Am.*, 266(4):76, 1992.

[16] A. C. Wilson and R. L. Cann. The recent african genesis of humans. *Sci. Am.*, 266(4):68, 1992.