J. Math. Biol. 45, 279–293 (2002)
Digital Object Identifier (DOI):
10.1007/s002850200148

Mathematical Biology

Peter Olofsson · Chad A. Shaw

Exact sampling formulas for multi-type Galton-Watson processes

Received: 10 June 2001 / Revised version: 21 November 2001 /
Published online: 23 August 2002 –

Abstract. Exact formulas for the mean and variance of the proportion of different types in a fixed generation of a multi-type Galton-Watson process are derived. The formulas are given in terms of iterates of the probability generating function of the offspring distribution. It is also shown that the sequence of types backwards from a randomly sampled particle in a fixed generation is a non-homogeneous Markov chain where the transition probabilities can be given explicitly, again in terms of probability generating functions. Two biological applications are considered: mutations in mitochondrial DNA and the polymerase chain reaction.

1. Introduction

Multitype Galton-Watson branching process models have many applications to the study of biological populations. These models are natural tools to study biological systems because they explicitly consider systems of proliferating individuals or particles. The multi-type setting allows individuals to have different probabilistic behavior depending on their type. In cell and molecular biology, type may for example be genotype, a dichotomous variable indicating mutant vs. wild-type or the number of accumulated mutations.

The literature on branching processes is vast and mostly focused on asymptotic theory. In comparison, relatively little has been done to address problems of sampling in finite time from a branching process. This is a problem which is relevant in many biological applications. In PCR, the polymerase chain reaction, genetic material is amplified and sampled after a fixed number of cycles. In cell cultures, cells are grown and harvested after a fixed number of population doublings. Also, many branching processes arising in these applications are intrinsically reducible in the sense that some types can only have certain other types in their ancestries. It is well known that in such processes, limiting distributions on the type space are typically degenerate and of no practical use.

P. Olofsson: Rice University, Department of Statistics MS-138, P.O.Box 1892, Houston, Texas 77251, USA. e-mail: olofsson@stat.rice.edu

C.A. Shaw: Baylor College of Medicine, Department of Molecular and Human Genetics, One Baylor Plaza, Smith Medical Research Building, Houston, Texas, 77030, USA.

Mathematics Subject Classification (2000): Primary 60J80, Secondary 92D10, 92D25

Key words or phrases: Multi-type Galton-Watson process – sampling formula – PCR – mitochondrial DNA

In a sequence of papers, Waugh, [17], and Joffe and Waugh, [7], [8], [9], address the so called kin number problem in Galton-Watson populations. They establish exact formulas for the probability distributions of family trees of a randomly sampled individual in a fixed generation. The most extensive treatment is of the single-type case, [7], the multi-type case is addressed in [8] and [9]. Our results and methods are similar but our focus differs slightly since we are motivated by explicit biological applications and aim at developing theoretical results which are relevant and useful in such applications.

The first main result of this paper is to provide exact formulas for the mean and variance of the proportion of a fixed type in a fixed generation. These are given in terms of the probability generating function of the offspring distribution.

The second main result is that the sequence of types in the ancestry of a randomly sampled particle is a non-homogeneous Markov chain where the transition probabilities can be given explicitly, again as a formula invoking the probability generating function of the offspring distribution.

Two biological applications are considered. The first concerns a certain mutation in mitochondrial DNA. Mitochondria are organelles in cells carrying their own DNA. Just like nuclear DNA, mitochondrial DNA (mtDNA for short) is subject to mutations which may take the form of base substitutions, duplications or deletions. This application focuses on one particular mutation, the mtDNA^{4977} deletion. This is a mutation which causes a deletion of about one third of the mitochondrial genome, thus producing a DNA molecule which is significantly smaller than normal. It has been observed that high levels of deletions are associated with certain degenerative diseases, for example Kearns-Sayre syndrome, [4]. These levels may be as high as 40%-50%. Low levels have been observed in different regions of the brain of healthy humans. There is a wide variety of issues involved such as different levels in different types of tissue but we will not attempt to address any of these. Instead, we focus on how the process of replication of mitochondrial DNA can be described as a multi-type Galton-Watson process and how the sampling formulas of the previous sections can be applied to explore how deletions accumulate over time. The idea of using branching processes in this application was first described in the unpublished manuscript [13]. We model the population of mitochondrial DNA as a two-type branching process where the types are normal and mutant molecules.

The second application is to PCR, the polymerase chain reaction. PCR provides a method for geometric amplification of genetic material. Invented less than two decades ago, [12], the technique is currently used in almost every facet of molecular biology, from DNA sequencing to gene expression studies. The PCR operates by performing repeated rounds of DNA synthesis in a test tube. As the reaction proceeds, newly created molecules serve as templates for the fabrication of further molecules in later reaction cycles, and for this reason the reaction is well represented by a branching process model. Mutations are known to occur during the PCR procedure. These become important when analysis focuses on the characteristics of individual molecules drawn from the reaction. In many applications mutations during amplification hinder analysis of the initial sample, as is the case in forensic PCR, [18]. PCR mutations also obscure results in experiments to assess genetic variation in a single cell or a small collection of cells, [5]. In other settings

PCR mutations are desirable, as is the case in site directed mutagenesis or artificial evolution experiments, [3]. We demonstrate how our results can be used to run simulations of the accumulation of mutations in a fixed number of PCR cycles.

2. Multi-type Galton-Watson processes

This section contains a brief description of multi-type Galton-Watson processes. For further reference, see [2], [6] or [11].

Suppose that a newborn individual inherits a type from a finite type space $\{0, 1, \ldots, r\}$. Given its type, the individual reproduces according to a probability distribution determined by this type. In both applications mentioned in the previous section, there are two different types and the type space can thus be taken as $\{0, 1\}$.

Denote by $X^{(j)}$ the number of offspring of type j and use subscripts to denote types of parents. The offspring distribution is

$$p_i(k_0, \ldots, k_r) = P_i(X^{(0)} = k_0, \ldots, X^{(r)} = k_r),$$

the probability that an i-type parent gets k_0 children of type $0, \ldots, k_r$ children of type r. The joint probability generating function of $(X^{(0)}, \ldots, X^{(r)})$ is defined as

$$\varphi_i(s_0, \ldots, s_r) = E_i[s_0^{X^{(0)}} \ldots s_r^{X^{(r)}}] = \sum_{k_0,\ldots,k_r} s_0^{k_0} \ldots s_r^{k_r} p_i(k_0, \ldots, k_r).$$

Let

$$\varphi_i^{(n)}(s_0, \ldots, s_r) = E_i[s_0^{Z_n^{(0)}} \ldots s_r^{Z_n^{(r)}}],$$

the probability generating function of the nth generation, $(Z_n^{(0)}, \ldots, Z_n^{(r)})$ starting from an ancestor of type i. Now let $\mathbf{s} = (s_0, \ldots, s_r)$ and let the functions $\varphi : R^{r+1} \to R^{r+1}$ and $\varphi^{(n)} : R^{r+1} \to R^{r+1}$ be defined by

$$\varphi(\mathbf{s}) = (\varphi_0(\mathbf{s}), \ldots, \varphi_r(\mathbf{s}))$$

and

$$\varphi^{(n)}(\mathbf{s}) = (\varphi_0^{(n)}(\mathbf{s}), \ldots, \varphi_r^{(n)}(\mathbf{s})).$$

Then there is the fundamental recursive relation

$$\varphi^{(n)}(\mathbf{s}) = \varphi(\varphi^{(n-1)}(\mathbf{s})).$$

or, coordinate-wise,

$$\varphi_i^{(n)}(\mathbf{s}) = \varphi_i(\varphi_0^{(n-1)}(\mathbf{s}), \ldots, \varphi_r^{(n-1)}(\mathbf{s})). \tag{2.1}$$

In the sequel, we will use the notation ψ_n for the probability generating function of $(Z_n^{(0)}, \ldots, Z_n^{(r)})$ when there is an arbitrary number of ancestors $(Z_0^{(0)}, \ldots, Z_0^{(r)})$, reserving the notation $\varphi^{(n)}$ for the case of one single ancestor.

3. Main results

3.1. Formulas for mean and variance

The following result gives the mean and variance of the proportion of type i individuals in the nth generation, conditioned on this generation being non-empty. Use the notation $|\mathbf{Z}_n|$ for the total number of individuals in the nth generation i.e. $|\mathbf{Z}_n| = \sum_{k=0}^{r} Z_n^{(k)}$.

Theorem 3.1. *Let* $\mathbf{u}$ *be a vector with all* u *entries except for a* v *in the* i^{th} *position:* $\mathbf{u} = (u, \ldots v, \ldots, u)$, $\mathbf{0} = (0, 0, \ldots, 0)$ *and denote by* ψ_n *the joint probability generating function of* $(Z_n^{(1)}, \ldots, Z_n^{(r)})$. *Then*

$$E\left[\left.\frac{Z_n^{(i)}}{|\mathbf{Z}_n|}\right| |\mathbf{Z}_n| > 0\right] = \frac{1}{1-\psi_n(\mathbf{0})}\int_0^1 \left.\frac{\partial}{\partial v}\psi_n(\mathbf{u})\right|_{u=v=s} ds$$

and

$$\begin{aligned} Var&\left[\left.\frac{Z_n^{(i)}}{|\mathbf{Z}_n|}\right| |\mathbf{Z}_n| > 0\right] \\ &= \frac{1}{1-\psi_n(\mathbf{0})}\int_0^1 -\log s \left(s\left.\frac{\partial^2}{\partial v^2}\psi_n(\mathbf{u})\right|_{u=v=s} + \left.\frac{\partial}{\partial v}\psi_n(\mathbf{u})\right|_{u=v=s}\right) ds \\ &\quad - \left(\frac{1}{1-\psi_n(\mathbf{0})}\int_0^1 \left.\frac{\partial}{\partial v}\psi_n(\mathbf{u})\right|_{u=v=s} ds\right)^2. \end{aligned}$$

3.2. The Markov property

Next, we investigate the dependence structure in the sequence of types in the lineage of a particle in the nth generation. We may think of this particle as sampled at random and denote its type by T_n. Since

$$P(T_n = i) = E\left[\left.\frac{Z_n^{(i)}}{|\mathbf{Z}_n|}\right| |\mathbf{Z}_n| > 0\right]$$

the probability $P(T_n = i)$ can be obtained from Theorem 3.1. Denote the type of this particle's parent by T_{n-1}, its grandparent's type by T_{n-2} and so on; we thus obtain a sequence of types $T_n, T_{n-1}, \ldots, T_0$, the type of the ancestor. It turns out that, conditional on non-extinction, this sequence is a non-homogeneous Markov chain with transition probabilities given by a formula invoking the probability generating function of the offspring distribution. This can be utilized for simulations to assess the type variation in lineages of sampled particles.

Theorem 3.2. *Let* $\mathbf{u} = (u, u, \ldots, u)$, $\mathbf{v} = (v, v, \ldots, v)$ *and let* $(\mathbf{u}_j)$ *have* u *in the* j*th position and* v *otherwise. Then*

$$P(T_k = i | T_{k+1} = j) = \frac{P(T_k = i, T_{k+1} = j)}{P(T_{k+1} = j)}$$

where

$$P(T_k = i, T_{k+1} = j) = \frac{1}{P(Z_n > 0)} \times \int_0^1 \frac{d}{du} \psi_k(\varphi_0^{(n-k)}(\mathbf{v}), \ldots, \varphi_i^{(n-k)}(\mathbf{u}_j), \ldots, \varphi_r^{(n-k)}(\mathbf{v}))\,|_{u=v=s}\, ds$$

and

$$P(T_{k+1} = j) = \frac{1}{P(Z_n > 0)} \times \int_0^1 \psi_k(\varphi_0^{(n-k-1)}(\mathbf{v}), \ldots, \varphi_j^{(n-k-1)}(\mathbf{u}), \ldots, \varphi_r^{(n-k-1)}(\mathbf{v}))\,|_{u=v=s}\, ds$$

4. Applications

4.1. Mutations in mitochondrial DNA

The population of mitochondrial DNA is modeled as a two-type process where the types are 0 (normal) and 1 (mutant). A normal can give birth to either two normals or, if there is a mutation, one normal and one mutant. The latter happens with probability λ and we refer to λ as the mutation rate. Mutants can only give birth to mutants. A DNA molecule may also die without reproducing (so called mitochondrial turnover, see [1]) and we let the survival probabilities be p and q for normals and mutants respectively. This gives the following offspring distributions:

$$p_0(0, 0) = 1 - p, \quad p_0(2, 0) = p(1 - \lambda), \quad p_0(1, 1) = p\lambda$$

for normals and

$$p_1(0, 0) = 1 - q, \quad p_1(0, 2) = q$$

for mutants. This gives the joint probability generating functions

$$\varphi_0(u, v) = 1 - p + p\lambda uv + p(1 - \lambda)u^2 \tag{4.1}$$

and

$$\varphi_1(u, v) = 1 - q + qv^2. \tag{4.2}$$

The proportion of mutants in the nth generation is

$$\frac{Z_n^{(0)}}{Z_n^{(0)} + Z_n^{(1)}}$$

and we can use Theorem 3.1 to compute its mean and variance. For now, we assume that the population is started from one normal ancestor, that is, $(Z_0^{(0)}, Z_0^{(1)}) = (1, 0)$.

To apply Theorem 3.1, note that in this case $\psi_n = \varphi_0^{(n)}$ and by the relation (2.1) of the previous section we get, omitting the argument (u, v),

$$\psi_n = \varphi_0^{(n)} = \varphi_0\left(\varphi_0^{(n-1)}, \varphi_1^{(n-1)}\right) = 1-p+p\lambda\varphi_0^{(n-1)}\varphi_1^{(n-1)}+p(1-\lambda)\left(\varphi_0^{(n-1)}\right)^2$$

and

$$\varphi_1^{(n)} = \varphi_1\left(\varphi_0^{(n-1)}, \varphi_1^{(n-1)}\right) = 1 - q + q\left(\varphi_1^{(n-1)}\right)^2.$$

Differentiating with respect to v gives

$$\frac{d}{dv}\varphi_0^n(n) = p\lambda\left(\frac{d}{dv}\varphi_0^{(n-1)}\varphi_1^{(n-1)} + \varphi_0^{(n-1)}\frac{d}{dv}\varphi_1^{(n-1)}\right)+2p(1-\lambda)\varphi_0^{(n-1)}\frac{d}{dv}\varphi_0^{(n-1)}$$

and

$$\frac{d}{dv}\varphi_1^{(n)} = 2q\varphi_1^{(n-1)}\frac{d}{dv}\varphi_1^{(n-1)}.$$

Recall that we are interested in

$$\frac{d}{dv}\varphi_0^{(n)}(u, v)\bigg|_{u=v=s}$$

and with the notation

$$F_n(s) = \varphi_0^{(n)}(s, s), \quad f_n(s) = \frac{d}{dv}\varphi_0^{(n)}(u, v)\bigg|_{u=v=s}$$

$$G_n(s) = \varphi_0^{(n)}(s, s), \quad g_n(s) = \frac{d}{dv}\varphi_1^{(n)}(u, v)\bigg|_{u=v=s}$$

we get the following recursive scheme:

$$\begin{aligned}
F_n(s) &= 1 - p + p\lambda F_{n-1}(s)G_{n-1}(s) + p(1 - \lambda)F_{n-1}^2(s)\\
f_n(s) &= p\lambda(f_{n-1}(s)G_{n-1}(s) + F_{n-1}(s)g_{n-1}(s)) + 2p(1 - \lambda)F_{n-1}(s)f_{n-1}(s)\\
G_n(s) &= 1 - q + qG_{n-1}^2(s)\\
g_n(s) &= 2qG_{n-1}(s)g_{n-1}(s)
\end{aligned}$$

with the initial conditions

$$\begin{aligned}
F_1(s) &= 1 - p - ps^2\\
f_1(s) &= p\lambda s\\
G_1(s) &= 1 - q + qs^2.\\
g_1(s) &= 2qs
\end{aligned}$$

Note that $f_1(s)$ is not the derivative of $F_1(s)$ with respect to s, but rather the derivative of $\varphi_0(u, v)$ with respect to v evaluated at the point $(u, v) = (s, s)$. To compute $\varphi_0^{(n)}(\mathbf{0})$, use the recursion formulas for F_n and G_n for $s = 0$.

For the variance formula, we get a similar recursion scheme (which is omitted here) where also second derivatives are included. For the initial conditions, note that

$$\frac{d^2}{dv^2}\varphi_0(u, v) = 0$$

and

$$\frac{d^2}{dv^2}\varphi_1(u, v) = 2q.$$

For the computations, we used the mutation rate $\lambda = 6 \cdot 10^{-8}$, which is the estimate obtained in [15]. To model non-dividing tissue such as the brain, the death rates $p = q = 1/2$ were used. This means that the population of mtDNA is kept constant on the average but that mitochondrial DNA keeps reproducing also in non-dividing cells, [1]. The generation length was taken as one month, [1]. We assume that the process starts from a normal molecule and compute the mean and variance of the proportion of deleted molecules at a given age. In Figure 1 the mean is plotted vs. age and in Figure 2 the mean plus one, two and three standard deviations respectively.

If there are $Z_0^{(0)}$ ancestors of type 0 and $Z_0^{(1)}$ of type 1, denote the joint probability generating function of $(Z_n^{(0)}, Z_n^{(1)})$ by ψ_n to obtain

$$\begin{aligned}\psi_n(u, v) &= E\left[u^{Z_n^{(0)}} v^{Z_n^{(1)}} \,\middle|\, Z_0^{(0)}, Z_0^{(1)}\right] \\ &= \left(\varphi_0^{(n)}(u, v)\right)^{Z_0^{(0)}} \left(\varphi_1^{(n)}(1, v)\right)^{Z_1^{(0)}}\end{aligned}$$

Hence, by differentiating this with respect to v we get

$$\begin{aligned}\frac{d}{dv}\psi_n(u, v) &= Z_0^{(0)} \left(\varphi_0^{(n)}(u, v)\right)^{Z_0^{(0)}-1} \frac{d}{dv}\varphi_0^{(n)}(u, v) \left(\varphi_1^{(n)}(1, v)\right)^{Z_1^{(0)}} \\ &\quad + \left(\varphi_0^{(n)}(u, v)\right)^{Z_0^{(0)}} Z_0^{(1)} \left(\varphi_1^{(n)}(1, v)\right)^{Z_1^{(0)}-1} \frac{d}{dv}\varphi_1^{(n)}(1, v)\end{aligned}$$

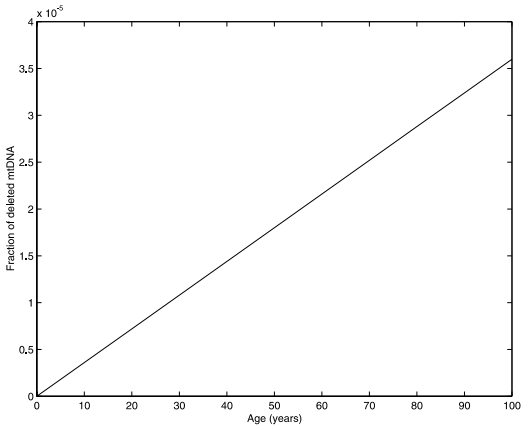

Fig. 1. Mean number of mutants as a function of age

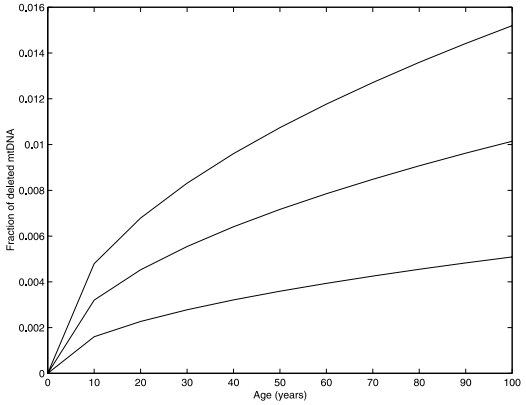

Fig. 2. Mean number of deletions plus one, two and three standard deviations as functions of age

where the functions $\varphi_0^{(n)}$ and $\varphi_1^{(n)}$ as well as their derivatives are computed by the recursive formulas from above. For the variance formula, we need to differentiate once more.

4.2. The polymerase chain reaction

We demonstrate how Theorems 3.1 and 3.2 can be used to run simulations of the accumulation of mutations in a fixed number of PCR cycles. In order to simplify the analysis of replication during the PCR, interest has focused on a single stranded model of the PCR procedure, [18], [10]. We follow this convention, thus regarding single DNA strands as the basic units in the branching process, referring to them as "particles" . A more biologically accurate semi-conservative model is considered in [14].

Note that a given particle at any given time either existed before the last PCR cycle or is newly created. The process is modeled as a two-type process where the type space is {0, 1}, "0" for "old" and "1" for "new". The distinction is crucial to mutation studies since new mutations only arise on newly created particles. The offspring distribution is

$$p_0(1,0) = 1-p, \quad p_0(1,1) = p$$
$$p_1(1,0) = 1-p, \quad p_1(1,1) = p$$

where p is the cycle efficiency, i.e. the probability that a given molecule replicates successfully in a given PCR cycle. This leads to joint probability generating functions

$$\varphi_0(u,v) = \varphi_1(u,v) = (1-p)u + puv.$$

For the simulations, Theorem 3.1 was used to compute the distribution of a randomly sampled particle in generation n and Theorem 3.2 to compute the transition probabilities. Simulations were then performed in which a particle was sampled

at random from generation n and the sequence of types in its lineage back to the ancestor generated. Each time a particle of type 1 appeared, it was independently assigned a new mutation with probability λ. The values $n = 30$, $p = 0.7$ and $\lambda = 0.05$ were used, [18].

Figure 3 shows a histogram of the number of mutations in the lineage of a randomly sampled particle in generation 30. This is based on 100,000 simulation runs of the Markov chain. As mentioned in Section 3, the transition probabilities $P(T_k = i|T_{k+1} = j)$ converge to a limiting distribution as $n \to \infty$ and in this particular application, the convergence is rapid. The limiting transition probabilities can be computed as

$$P(T_k = i|T_{k+1} = j) = \frac{\nu(i)M(i, j)}{\rho\nu(j)}$$

where $M(i, j) = E_i[X^{(j)}]$, the (i, j)th entry in the mean reproduction matrix

$$M = \begin{pmatrix} M(0,0) & M(0,1) \\ M(1,0) & M(1,1) \end{pmatrix} = \begin{pmatrix} 1 & p \\ 1 & p \end{pmatrix},$$

ρ is the largest eigenvalue of M and ν is the left eigenvector of M corresponding to ρ. In this case,

$$\rho = 1 + p, \quad \nu(0) = \frac{p}{1+p}, \quad \nu(1) = \frac{1}{1+p}$$

which gives, in the limit

$$P(T_k = 1|T_{k+1} = j) = \frac{p}{1+p} \approx 0.41$$

for both $j = 0$ and $j = 1$. The computations reveal that this limit is effectively reached after less than ten generations. For comparison, we ran simulations where

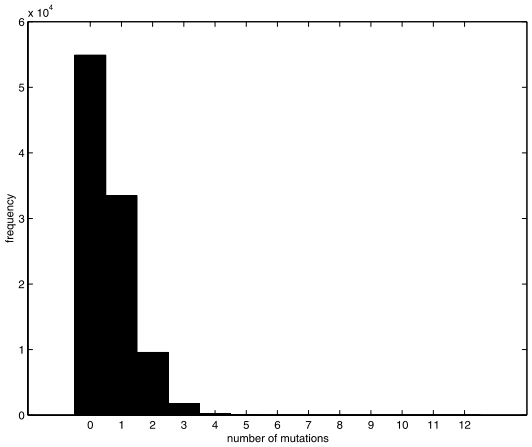

Fig. 3. Histogram of the number of mutations in 30 PCR generations. Mutation rate: 0.05. Mean: 0.59, standard deviation: 0.76

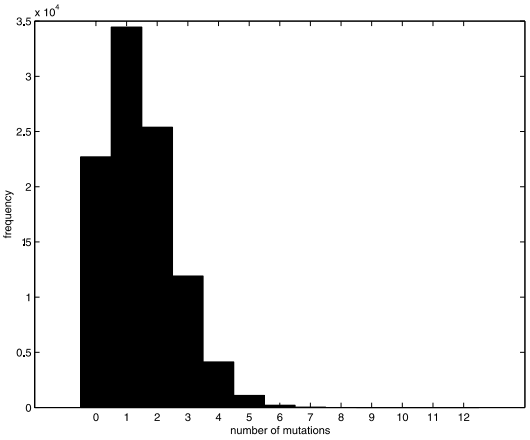

Fig. 4. Histogram of the number of mutations in 30 PCR generations using limiting transition probabilities. Mutation rate: 0.05. Mean: 1.45, standard deviation: 1.17

the limiting transition probabilities were used. Figure 4 shows a histogram for this case. Figures 5 and 6 repeat the two simulation schemes with mutation rate 0.1.

5. Discussion

We stated results about the mean and variance of the proportion of types in a fixed generation in a multi-type branching process. Also, the transition probabilities for the sequence of types from a randomly sampled particle in a fixed generation was obtained. It is important to notice that our formulas are exact and in finite time (fixed generation). This is of importance in both the applications we presented.

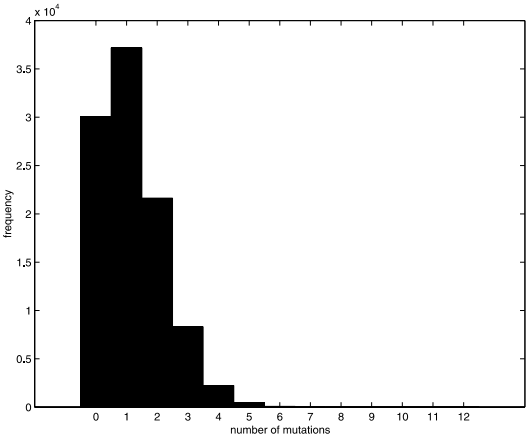

Fig. 5. Histogram of the number of mutations in 30 PCR generations. Mutation rate: 0.1. Mean: 1.17, standard deviation: 1.06

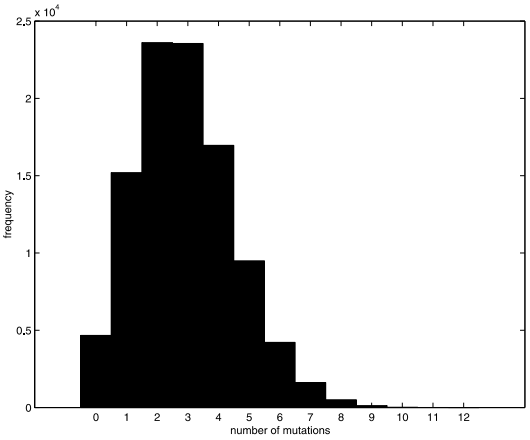

Fig. 6. Histogram of the number of mutations in 30 PCR generations using limiting transition probabilities. Mutation rate: 0.1. Mean: 2.91, standard deviation 1.62

The first application was to the accumulation of a particular deletion in human mitochondrial DNA. The branching process has types 0 for normal and 1 for mutant. We notice that in this case, there is no back-mutation. Thus, as time goes on, the proportion of mutants increases steadily to 1 so the only stable limiting distribution has probability 1 for mutant and 0 for normal. Therefore, asymptotic approximations are useless and the exact, finite-time formulas we have developed are necessary. In [16], it was reported that the level of deletions in the brain in an 80-year-old normal individual was 0.5%. In Figure 2, we notice that at age 80, this level is well within two standard deviations of the mean and is thus well explained by our model as a natural phenomenon.

To be able to address questions of how the mutants can quickly accumulate to the levels observed for example in individuals with Kearns-Sayre syndrome, more sophisticated modeling should be carried out, taking into account the various stages in the development of a human being. During the early stages, until the brain is fully developed, the population of mtDNA is growing and the values of p and q should be chosen accordingly. There is also the question of what happens at the very early stages since it is known that the oocyte contains roughly 100,000 mtDNA, about a hundred times as many as other cells, [4]. The purpose in this paper is merely to point out the potential usefulness of the sampling formulas and we leave the biological problems to be pursued in detail elsewhere. We will however mention that to assess the importance of initial proportions of mutants, we need to start the process from an arbitrary distribution of normals and mutants. This can be carried out according to the formulas described at the end of Section 4.1.

The second application was to PCR. Here, the types in the branching process were 0 for an old particle and 1 for a newly synthesized particle. Each new particle was independently assigned a mutation with probability λ. It was noticed that in this case there is a rapid convergence towards limiting distributions. This means

that after about ten generations, the proportion of types has stabilized so that, regardless of the type of the sampled particle, its parent is of type 1, and thus prone to mutation, with a fixed probability. Thus, as long as our analysis only concerns later generations, approximations based on limiting distributions work well. However, when we run simulations from a sampled particle backwards to the ancestor, the approximation obtained by using the limiting probabilities gets worse the closer to the ancestor we get. The difference can be clearly seen by comparing Figures 3 and 5 where the exact formula has been used to Figures 4 and 6 where the approximation has been used. Using the approximation overestimates the number of accumulated mutations. The reason for this is that we start from a mutation-free particle and only after mutations have started to accumulate do the proportions settle in towards their limits.

6. Appendix: Proofs

The methods of proof are inspired by those of [7] and [17]. The proofs are based on the following lemma, which is of interest in its own right.

Lemma 6.1. *Let X and Y be two non-negative integer valued random variables with the joint probability generating function $\varphi_{X,Y}(u, v)$. Let*

$$f_1(s) = \frac{\partial}{\partial u}\varphi_{X,Y}(u, v)\bigg|_{u=v=s}$$

and

$$f_2(s) = \frac{\partial^2}{\partial u^2}\varphi_{X,Y}(u, v)\bigg|_{u=v=s}.$$

Then

$$E\left[\frac{X}{X+Y}\,\middle|\, X+Y>0\right] = \frac{1}{1-\varphi_{X,Y}(0,0)}\int_0^1 f_1(s)ds \tag{6.1}$$

and

$$\begin{aligned} &Var\left[\frac{X}{X+Y}\,\middle|\, X+Y>0\right] \\ &= \frac{1}{1-\varphi_{X,Y}(0,0)}\int_0^1 -\log s(sf_2(s)+f_1(s))ds \\ &\qquad -\left(\frac{1}{1-\varphi_{X,Y}(0,0)}\int_0^1 f_1(s)ds\right)^2 \end{aligned}$$

Note. If X and Y are independent and strictly positive with probability generating functions φ_X and φ_Y respectively, (6.1) reduces to

$$E\left[\frac{X}{X+Y}\right] = \int_0^1 \varphi_X'(s)\varphi_Y(s)ds.$$

Proof. First note that, if $j \geq 0$,

$$\int_0^1 s^j ds = \frac{1}{j+1} \tag{6.2}$$

and, with the standard notation $E[X; A] = E[XI_A]$,

$$\begin{aligned} E\left[\frac{X}{X+Y}\middle| X+Y>0\right] &= \frac{1}{P(X+Y>0)} E\left[\frac{X}{X+Y}; X+Y>0\right] \\ &= \frac{1}{1-\varphi_{X,Y}(0,0)} E\left[\frac{X}{X+Y}; X+Y>0\right]. \end{aligned}$$

Also, by elementary properties of probability generating functions,

$$\frac{\partial}{\partial u}\varphi_{X,Y}(u,v)\Big|_{u=v=s} = E\left[Xu^{X-1}v^Y\right]\Big|_{u=v=s} = E\left[Xs^{X+Y-1}\right]. \tag{6.3}$$

By (6.2), (6.3) and an application of Fubini's Theorem,

$$\begin{aligned} E\left[\frac{X}{X+Y}; X+Y>0\right] &= E\left[X\int_0^1 s^{X+Y-1}ds\right] = \int_0^1 E\left[Xs^{X+Y-1}\right]ds \\ &= \int_0^1 \frac{\partial}{\partial u}\varphi_{X,Y}(u,v)\Big|_{u=v=s} ds = \int_0^1 f_1(s)ds \end{aligned}$$

and the first part is shown. For the second part, first note that

$$\begin{aligned} Var\left[\frac{X}{X+Y}\middle| X+Y>0\right] &= E\left[\frac{X(X-1)}{(X+Y)^2}\middle| X+Y>0\right] \\ &\quad + E\left[\frac{X}{(X+Y)^2}\middle| X+Y>0\right] \\ &\quad - \left(E\left[\frac{X}{X+Y}\middle| X+Y>0\right]\right)^2 \end{aligned}$$

where the expression for the third term is given by the first part of the lemma. For the two other terms, note that

$$\frac{\partial^2}{\partial u^2}\varphi_{X,Y}(u,v)\Big|_{u=v=s} = E\left[X(X-1)s^{X+Y-1}\right] \tag{6.4}$$

and make repeated use of (6.2), (6.3) and Fubini's Theorem to obtain

$$\begin{aligned} &E\left[\frac{X}{(X+Y)^2}; X+Y>0\right] \\ &= E\left[X\int_0^1 \frac{t^{X+Y-1}}{X+Y}dt\right] \\ &= E\left[X\int_0^1\int_0^t \frac{1}{t}s^{X+Y-1}dsdt\right] \end{aligned}$$

$$
\begin{aligned}
&= E\left[X\int_0^1\int_s^1 \frac{1}{t}s^{X+Y-1}dtds\right]\\
&= E\left[X\int_0^1 -\log s\cdot s^{X+Y-1}ds\right]\\
&= -\int_0^1 \log s\, E\left[Xs^{X+Y-1}\right]ds = -\int_0^1 \log s f_1(s)ds
\end{aligned}
$$

and for the first term, in a similar fashion,

$$
\begin{aligned}
&E\left[\frac{X(X-1)}{(X+Y)^2};\, X+Y>0\right]\\
&\quad= -\int_0^1 \log s\, E\left[X(X-1)s^{X+Y-1}\right]ds\\
&\quad= -\int_0^1 s\log s\, E\left[X(X-1)s^{X+Y-2}\right]ds\\
&\quad= -\int_0^1 s\log s f_2(s)ds.
\end{aligned}
$$

Proof of Theorem 3.1. Since

$$
|\mathbf{Z}_n| = Z_n^{(i)} + \sum_{j\neq i} Z_n^{(j)},
$$

let $X = Z_n^{(i)}$ and $Y = \sum_{j\neq i} Z_n^{(j)}$ to obtain

$$
\varphi_{X,Y}(u,v) = E[u^X v^Y] = E\left[u^{Z_n^{(i)}} v^{\sum_{j\neq i} Z_n^{(j)}}\right] = \psi_n(\mathbf{u})
$$

and the result follows immediately from Lemma 6.1. □

Proof of Theorem 3.2. The probability $P(T_{k+1}=j, T_k=i)$ equals the expected proportion of individuals in the nth generation who had ancestors of types i and j in the kth and $(k+1)$th generation respectively. Thus, let $Z_{n-j}(m)$ denote the number of individuals in generation n stemming from the mth individual in generation j and apply Lemma 6.1 to

$$
\sum_{l=1}^{Z_k^{(i)}}\sum_{m=1}^{X_l^{(j)}} Z_{n-k-1}(m)
$$

and

$$
\sum_{l=1}^{Z_k^{(i)}}\sum_{d\neq j}\sum_{m=1}^{X_l^{(d)}} Z_{n-k-1}(m) + \sum_{d\neq i}\sum_{l=1}^{Z_k^{(d)}} Z_{n-k}(l).
$$

For the probability $P(T_{k+1} = j)$ apply the Lemma to

$$\sum_{l=1}^{Z_{k+1}^{(j)}} Z_{n-k-1}(l)$$

and

$$\sum_{m \neq j} \sum_{l=1}^{Z_{k+1}^{(m)}} Z_{n-k-1}(l).$$

□

References

[1] Arking, R.: *Biology of Aging, 2nd ed.* Sinauer, Sunderland, MA, 1998

[2] Athreya, K., Ney, P.: *Branching Processes*. Springer Verlag, New York, 1972

[3] Caldwell, R., Joyce, G.: Randomization of genes by PCR mutagenesis. *PCR Methods and Applications*, **2**, 28–33 (1992)

[4] Chinnery, P.F., Turnbull, D.M.: Mitochondrial DNA and disease. *Lancet*, **354**, 17–21 (1999)

[5] Hsu, L.Y.H., Kaffe, S., Jenkins, E.C., Alonso, L., Benn, P.A., David, K., Hirschhorn, K., Lieber, E., Shanske, A., Shapiro, L.R., Schutta, E., Warburton, D.: Proposed guidelines for diagnosis of chromosome mosaicism in amniocytes based on data derived from chromosome mosaicism and pseudomosaicism studies. *Prenatal diagnosis*, **12**, 555–573 (1992)

[6] Jagers, P.: *Branching Processes with Biological Applications*. New York: Wiley, 1975

[7] Joffe, A., Waugh, W.: Exact distributions of kin numbers in a Galton-Watson process. *J. Appl. Probab.*, **19**, 767–775 (1982)

[8] Joffe, A., Waugh, W.A.O'N.: The kin number problem in a multitype galton-watson population. *J. Appl. Probab.*, **22**, 37–47 (1985)

[9] Joffe, A., Waugh, W.A.O'N.: Exact distribution of kin number in multitype galton-watson process. *Semi-Markov models*, 397–405 (1986)

[10] Krawczak and Reiss.: Polymerase chain reaction replication errors and reliability of gene diagnosis. *Nucleic Acids Research*, **17**(6), 2197–2201 (1989)

[11] Mode, C.: *Multitype Branching Processes*. American Elsevier, New York, 1971

[12] Mullis, K.B., Faloona, F.A.: Specific synthesis of DNA invitro via a a polymerase-catalyzed chain-reaction. *Methods in Enzymology*, **155**, 335–351 (1987)

[13] Navidi, W., Tavaré, S., Arnheim, N.: The role of the mutation rate and selective pressures on observed levels of the human mitochondrial DNA deletion mtDNA^{4977}. Unpublished manuscript. 1996

[14] Shaw, C.: *Genealogical Methods for Multitype Branching Processes with Applications in Biology*. PhD thesis, Rice University, June 2000

[15] Shenkar, R., Navidi, W., Tavaré, S., Dang, M.H., Chomyn, A., Attardi, G., Cortopassi, G., Arnheim, N.: The mutation rate of the human mtDNA deletion mtDNA^{4977}. *Am. J. Hum. Genet.*, **59**, 749–755 (1996)

[16] Soong, N-W., Hinton, D.R., Cortopassi, G., Arnheim, N.: Mosaicism for a specific somatic mitochondrial DNA mutation in adult human brain. *Nature Genetics*, **2**, 318–323 (1992)

[17] Waugh: Application of the Galton-Watson process to the kin number problem. *Adv. Appl. Probab.*, **13**, 631–649 (1981)

[18] Weiss, G., von Haeseler, A.: A genealogical approach to the polymerase chain reaction. *Nucleic Acids Research*, **25**(15), 3082–3087 (1997)