Mathematical Biosciences xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Mathematical Biosciences

journal homepage: www.elsevier.com/locate/mbs

Fluctuation analysis on mutation models with birth-date dependence

Adrien Mazoyer[1]

Laboratoire Jean Kuntzmann, Bâtiment IMAG, 700 avenue centrale, Saint Martin d'Hères 38401, France

ARTICLE INFO

Keywords:
Branching process
Fluctuation analysis
Luria–Delbrück distribution
Non-exponential growth
Parametric estimation
Probability generating function

MSC:
92D25
60J80
62P10

ABSTRACT

The classic Luria–Delbrück model can be interpreted as a Poisson compound (number of mutations) of exponential mixtures (developing time of mutant clones) of geometric distributions (size of a clone in a given time). This "three-ingredients" approach is generalized in this paper to the case where the split instant distributions of cells are not i.i.d. : the lifetime of each cell is assumed to depend on its birth date. This model takes also into account cell deaths and non-exponentially distributed lifetimes. Previous results on the convergence of the distribution of mutant counts are recovered. The particular case where the instantaneous division rates of normal and mutant cells are proportional is studied. The classic Luria–Delbrück and Haldane models are recovered. Probability computations and simulation algorithms are provided. Robust estimation methods developed for the classic mutation models are adapted to the new model: their properties of consistency and asymptotic normality remain true; their asymptotic variances are computed. Finally, the estimation biases induced by considering classic mutation models instead of an inhomogeneous model are studied with simulation experiments.

1. Introduction

Mutation models are probabilistic descriptions of the growth of a population of cells, in which scarce mutations randomly occur. Data are samples of integers, interpreted as final numbers of mutant cells. The frequent appearance in the data of very large mutant counts, usually called "jackpots", evidences heavy-tailed probability distributions. Mutation models have two objectives: study the distribution of the number of mutant cells at the end of the growth process; perform fluctuation analysis on data to estimate the probability for a mutant to appear at any division.

Any classic mutation model can be interpreted as the result of the three following ingredients [10]:

1. a random number of mutations occurring with small probability among a large number of cell divisions. Due to the law of small numbers, the number of mutations approximately follows a Poisson distribution. The expectation of that distribution is the product of the mutation probability by the total number of divisions;
2. from each mutation, a clone of mutant cells growing during a random time. Due to exponential growth, most mutations occur close to the end of the process, and the developing time of a random clone has exponential distribution. The rate of that distribution is the *relative fitness*, i.e., the ratio of the growth rate of normal cells to that of mutants;
3. the number of mutant cells that any clone developing for a given time will produce. The distribution of this number depends on the modeling assumptions, in particular the lifetimes of mutants.

This approach leads to a family of probability distributions which depend on the expected number of mutations and the relative fitness. One of the most used mutation models is the well known Luria–Delbrück model [20]. A review on the Luria–Delbrück distribution for the second half of the twentieth century can be found in [43]. Here we try to give a historical summary of previous works, including last decades. Mathematical descriptions were introduced by Lea and Coulson [18], followed by Armitage [3] and Bartlett [5]. In that model, division times of mutant cells were supposed to be exponentially distributed. Thus a clone develops according to a Yule process [42, p. 35]; [4, p. 109], and its size at any given time follows a geometric distribution. The distribution of final mutant counts is also explicit when lifetimes of mutant cells are supposed to be constant. This latter model is called Haldane model by Sarkar [30]; a first practical algorithm for computing the Haldane distribution was proposed much later by Zheng [45]; an explicit form of the asymptotic distribution is finally given by Ycart [39]. General lifetimes have also been studied in [39], but no explicit distribution is available apart from the exponential and constant lifetimes. Other extensions of the Luria–Delbrück model take into account the case where cells have a certain probability to die rather than divide [2, sec. 3.1]; [6,14,40], where final number of cells are

E-mail address: mazoyer.adrien@courrier.uqam.ca.
[1] Present address: Universitè du Quèbec á Montrèal.

https://doi.org/10.1016/j.mbs.2018.06.006
Received 20 November 2017; Received in revised form 13 June 2018; Accepted 18 June 2018

random [2,14,41,46].

As mentioned above, the main statistical objective of mutation models is the estimation of the probability for a mutant cell to appear upon any given cell division. It is computed dividing estimate of the mean number of mutations by the mean final number of cells. Computing robust estimates is of crucial importance in medical applications, like cancer tumor relapse or multidrug resistance of *Mycobacterium Tuberculosis* for instance. Estimates are realizations of an estimator which is a random variable depending on the considered sample. A robust estimator satisfies two properties: consistency, and explicit asymptotic distribution. Thus confidence intervals and p-values can be computed.

Luria and Delbrück [20] have proposed two estimators. The first estimator, called p_0 estimator, is based on the relation between the probability to get a null count in the sample and the mean number of mutations: taking the negative logarithm of the relative frequency of zeros among the sample gives a robust estimate of the expected number of mutations [20, Eq. (5)]. Remark that if the sample does not contain null counts, the method cannot be applied. The second estimator proposed in the same article is based on the relation between the mean number of mutants, the sample size, and the final mutant of cells [20, Eq. (8)]. Since this estimator does not have expectation, it is not consistent and should not be used. A wide panel of estimation methods has been proposed since then [9,28]. Most of these methods are based on empirical median of the mutant count to reduce the heavy tail effects [18, Eq. (25)], [13, Eq. (6)]. Even if some median methods give good results in practice, the consistency property is not satisfied or cannot be checked: indeed the empirical median is not a robust estimator of the median for discrete distributions. Thus other methods which satisfy the properties of interest should be considered. Since the distribution of final numbers has an explicit form, the Maximum Likelihood (ML) seems to be an obvious optimal choice [21,32,44]. The computation of the likelihood and its derivatives can be numerically unstable in the sample contains large jackpots. One of the possible ways to reduce such tail effects is “Winsorization” of the sample [38, Section 2.2], which consists in replacing any value of the sample that pass a certain bound by the bound itself. However, since very large numbers of mutants are not countable in practice, such cases where the Winsorization is required are not common. The last method exposed here uses the probability generating function [10,29], and is called Generating Function (GF) method. This method is comparable in terms of efficiency to the ML method. Moreover, this method has a good numerical stability and a negligible computing time. However, this method depends on tuning parameters. These parameters should be set according to the data, which is not possible in practice. Therefore, the GF method is implemented as a trained semi-parametric method. The p_0, ML, and GF methods provide asymptotically normal estimators of the mean number of mutation. The estimation of the mutation probability can then be deduced dividing estimation of mean number of mutations by the mean final number of cells. The fluctuations of the final number of cells can also be taken into account [41], in order to get a more accurate estimation. Sometimes, data are samples of couples of integers, interpreted as final numbers of mutants and final numbers of cells. In that case, the Maximum Likelihood can be used to estimate directly the mutation probability [41]. Moreover, the relative fitness can also be estimated.

The lifetimes of the cells are supposed to be i.i.d. in the classic mutation models. Thus, the population grows exponentially or dies out [4,12]. However, the growth is in practice logistic [16]: it is exponential until an inflexion instant when the growth begins to slow and eventually tends to zero. Indeed, during an experiment, a colony of cells grows in an environment which contains a finite amount of resources. Then a cell born at a instant s_1 will complete its lifetime faster than a cell born at a instant $s_2 \ll s_1$. The Verhulst model [36] is one of the most known deterministic growth model which takes into account this limitation. Logistic-type stochastic models are described by Allen [1, sec. 9.4.2], and mathematically studied by several authors among which [17,34,35]. Stewart et al. [33] proposed an approach to take into account the decreasing rate of division as the cells run out of resources. Houchmandzadeh [11] described a discrete approach with a general growth model for mutant clones. A Luria–Delbrück model assuming that the replication number of any normal cell is limited whereas mutants are not has been exposed in [27]. However, none of these studies provide results for the non-i.i.d. lifetimes case, in particular on the distribution of final mutant count.

In a previous work [23], an extension of the classic mutation models to the case where the split instant of a cell depends on its birth date has been proposed. The results on the asymptotic distribution of the mutant count were very similar to that of classic Luria–Delbrück model. Therefore the methods of estimation described above should be directly adapted to the model with birth-date dependence. As for the homogeneous case, the three methods provide consistent and asymptotically normal estimator for the parameters of interest. However, fast simulation cannot be deduced from the approach exposed in [23]. Such algorithms are necessary to perform large scale simulation studies. Another approach of the model is proposed here: as for the homogeneous mutation models, the distribution of the final mutant count can still be interpreted as the result of three ingredients. As a direct consequence, a fast simulation algorithm can be deduced. The asymptotic results on the distribution of the mutant count of [23] are recovered and extended to the case where the death of normal cells are taken into account.

General modeling assumptions are described in Section 2. The three-ingredients approach is exposed and used to prove the convergence in distribution of the final mutant count in the Section 3. Probability computation and simulation algorithms are exposed in Section 4. The case where the hazard functions associated to the split instant distribution of normal and mutant cells are proportional is studied. In particular, the Luria–Delbrück distribution with cell deaths [40] is recovered. The Haldane model is also recovered, and extended to the case where mutant cell deaths are taken into account. The statistical question of estimation of the number of mutations and the relative fitness is studied in Section 5: assuming that the other parameters are known, the methods p_0, ML, and GF methods can be extended to inhomogeneous models. Estimation biases induced by considering classic mutation models instead of model with birth-date dependence are illustrated with simulation experiments in Section 6. In particular, simulations seem to show that the bias will be in practice negligible, which encourages to continue to use the classic model for estimation.

2. Hypotheses and models

Notations and hypotheses are described in this section. A rigorous definition of the probabilistic model as a tree-indexed process has already been given in [23, Section 2]. Thus, the dynamics are shortly described, and the modeling assumptions will be summarized at the end of this section.

Consider a normal cell born at a given instant s. At a random instant (called here a *final instant*) with cumulative distribution function (cdf) $F_\nu(s, \cdot)$, the cell produces one normal and one mutant cell with probability π (this event is called a *mutation*), two normal cells with probability $1 - \pi - \gamma$, or dies with probability γ. Consider now a mutant cell, born at a given time s. At a random instant with cdf $F_\mu(s, \cdot)$, the mutant produces two mutant cells with probability $1 - \delta$ or dies with probability δ. Starting from a single cell, whatever its nature, the set of all descendants constitutes a *clone*. Thus the *clone size* at a given time t denotes the number of cells alive at time t in the clone. Consider a given cell, the mutation or death events are independent from its final instant. Two cells are independent conditionally on their common ancestor. Therefore, the clones stemming from these cells are also independent conditionally on this ancestor. Remark that those dependence assumptions hold whatever the nature of the considered cells. At the beginning of the process, the population contains a given number n of normal cells and no mutants.