Journal of Theoretical Biology 366 (2015) 1–7

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Fast maximum likelihood estimation of mutation rates using a birth–death process

Xiaowei Wu*, Hongxiao Zhu

Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, United States

HIGHLIGHTS

- Maximum likelihood estimator of mutation rates encounters computational difficulty.
- We propose a new maximum likelihood estimator based on a birth–death process model.
- The proposed method shows substantial improvement on computational speed.
- The estimation is accurate, and applicable to arbitrarily large number of mutants.

ARTICLE INFO

Article history:
Received 21 June 2014
Received in revised form
31 October 2014
Accepted 12 November 2014
Available online 20 November 2014

Keywords:
Fluctuation analysis
Spontaneous mutation
Markov branching process

ABSTRACT

Since fluctuation analysis was first introduced by Luria and Delbrück in 1943, it has been widely used to make inference about spontaneous mutation rates in cultured cells. Under certain model assumptions, the probability distribution of the number of mutants that appear in a fluctuation experiment can be derived explicitly, which provides the basis of mutation rate estimation. It has been shown that, among various existing estimators, the maximum likelihood estimator usually demonstrates some desirable properties such as consistency and lower mean squared error. However, its application in real experimental data is often hindered by slow computation of likelihood due to the recursive form of the mutant-count distribution. We propose a fast maximum likelihood estimator of mutation rates, MLE-BD, based on a birth–death process model with non-differential growth assumption. Simulation studies demonstrate that, compared with the conventional maximum likelihood estimator derived from the Luria–Delbrück distribution, MLE-BD achieves substantial improvement on computational speed and is applicable to arbitrarily large number of mutants. In addition, it still retains good accuracy on point estimation.

Published by Elsevier Ltd.

1. Introduction

It has been more than half a century since Luria and Delbrück first introduced fluctuation analysis and showed that mutations in bacteria arise spontaneously, rather than being a response to selection (Luria and Delbrück, 1943). In a typical fluctuation experiment, a number of separate cultures (called parallel cultures), each started with a small inoculum of cells from the same strain, are plated on a selective medium for the test of resistance. The number of mutants in parallel cultures are counted and then used for statistical analysis. One important task of fluctuation analysis is to estimate the mutation rate using the mutant counts. Under several model assumptions, e.g., cells grow at a constant rate and mutations occur at a rate proportional to the total number of cells, one can explicitly derive the probability distribution of the number of mutants that appear in a fluctuation experiment, known as the Luria–Delbrück (LD) distribution. This distribution constitutes the basis of mutation rate estimation (Zheng, 1999; Foster, 2006).

Various mutation rate estimators have been developed based on the LD distribution, including the P_0 estimator (Luria and Delbrück, 1943), the mean estimator (Luria and Delbrück, 1943), the median estimator (Lea and Coulson, 1949; Drake, 1991; Jones et al., 1994) and the maximum likelihood estimator (MLE) (Sarkar et al., 1992; Jones et al., 1993). Extensive simulations have shown that the MLE outperforms the other estimators in terms of bias and variance (Stewart, 1994; Rosche and Foster, 2000) since it incorporates all information from the mutant-count distribution (rather than just moments or quantiles). Efficient algorithms for computing the MLE have been proposed in the literature (Sarkar et al., 1992; Ma et al., 1992). To quantify the uncertainty of the MLE, confidence interval

* Corresponding author
E-mail address: xwwu@vt.edu (X. Wu).

http://dx.doi.org/10.1016/j.jtbi.2014.11.009
0022-5193/Published by Elsevier Ltd.

can also be obtained (Stewart, 1994; Zheng, 2002). In addition, several computer packages have been developed to facilitate the use of the MLE on real fluctuation experimental data (Zheng, 2002; Hall et al., 2009).

Despite the advantages of the MLE, a major obstacle for its practical use lies in the computational complexity due to the recursive form of the LD distribution, especially when the observed number of mutants per culture is large (for example, greater than 5000) in real fluctuation experiments. To overcome this limitation, we propose an alternative maximum likelihood-based estimator, MLE-BD, using a birth–death process model. In the proposed method we assume binary fission and non-differential growth for the cell population. Under this assumption, MLE-BD is straightward to derive and calculate. The main advantages of the proposed approach include: (1) It improves the computational speed substantially as compared to the traditional MLE. (2) It can easily handle practical issues arising in real fluctuation experiments, such as (arbitrarily) large number of mutants per culture, large number of parallel cultures, and divergent culture sizes. (3) It provides an alternative interpretation to the mutation rate from a "mutation probability per division" point of view, which may be of special interest to biologists since the mutation probability is disengaged from the growth rate. Through simulation studies, we demonstrate that MLE-BD is computationally superior to the existing MLE in several orders of magnitude while retaining good estimation accuracy.

2. Materials and methods

2.1. An overview of mutation rate estimation using LD distribution

In order to model the cell growth and mutation process in fluctuation analysis, several mathematical formulations have been developed, as summarized in Zheng (1999). On one hand, these formulations share a few common assumptions, including (i) the cell population starts with nonmutant cells, (ii) mutations occur randomly at a rate proportional to the population size, and (iii) cell death and backward mutations are neglected. On the other hand, they differ mainly on the assumption of cell growth. For example, the Luria–Delbrück formulation assumes deterministic growth for both nonmutant and mutant cells, the Lea–Coulson formulation assumes deterministic growth for nonmutant cells but stochastic growth (i.e., Yule process) for mutant cells, and the Bartlett formulation assumes stochastic growth for both nonmutant and mutant cells (Bartlett, 1978; Zheng, 2008). Perhaps the most widely used formulation is the Lea–Coulson formulation. Under the specific assumptions of this formulation, one can easily see that the number of mutants at any time follows a Poisson-stopped-sum distribution. This provides us an elegant, well recognized probability mass function (pmf) for the number of mutants (Sarkar et al., 1992; Ma et al., 1992) (we limit the scope of this paper to non-differential growth, that is, mutants and nonmutants share the same growth rate):

$$p_0 = e^{-m}, \quad p_k = \frac{m}{k}\sum_{j=1}^{k} \phi^{j-1}\left(1-\frac{j\phi}{j+1}\right)p_{k-j}, \ k \geq 1, \tag{1}$$

where

$$m = \frac{\mu}{\beta_1}(n-n_0) = \mu_\beta(n-n_0), \quad \phi = 1-e^{-\beta_1 t} = 1-\frac{n_0}{n}. \tag{2}$$

Eq. (1) provides a recursive expression for pmf, where p_k is the probability of observing k mutants in the overall population that starts with n_0 nonmutant cells and grows to size n at time t. Here, m is a key intermediate parameter denoting the expected number of mutations at time t, β_1 is the growth rate of cells (for both nonmutants and mutants), μ is the mutation rate per unit time, and μ_β may be called the mutation rate per cell division (often scaled by log 2).

The above distribution, often referred to as the LD(m,ϕ) distribution, provides the basis of various mutation rate estimators. For example, for small mutation rates, the P_0 estimator is obtained by equating the zero probability, p_0, to the observed proportion of cultures containing only nonmutants. Moment-based and quantile-based methods, such as the mean estimator and the median estimator, equate explicit or numerical expressions of the mean (expected value) and the median (50% quantile) to their sample counterparts. With recent advances in computational technology, more accurate mutation rate estimators based on the maximum likelihood method become prevalent. The MLE aims to maximize the likelihood of obtaining the observed mutant counts in parallel cultures. Following convention, one may estimate m instead of μ or μ_β since this intermediate parameter is directly connected with the mutation rate through Eq. (2). Suppose that in a fluctuation experiment with J parallel cultures, the number of mutant cells at time t is counted as $k_1, \ldots, k_J$. The MLE of m can be written as

$$\hat{m} = \arg\max_m \ell(m, \phi | k_1, \ldots, k_J) = \arg\max_m \sum_{i=1}^{J} \log\ p_{k_i}(m, \phi). \tag{3}$$

In general, the MLE is more preferable in the estimation of mutation rates because its statistical properties, such as consistency and efficiency, can be easily assessed given large samples. However, due to its computational complexity, the MLE has not been practically applied on real experimental data until an explicit algorithm was given by Jones et al. (1994) and implemented in a computer package by Zheng (2002).

2.2. The birth–death process model and distribution of nonmutant counts

The above LD distribution-based MLE (herefrom referred to as MLE-LD) provides a sophisticated solution to mutation rate estimation. However, there still exist some impediments to its effective use in real experimental data. First and foremost, due to the recursive form of the LD pmf, the computation of the MLE often encounters difficulty when the observed number of mutants k in each culture is large. In practice, the point estimation becomes slow and the confidence interval calculation (even by the sequence convolution technique introduced in Zheng (2002)) appears intractable when k is on the magnitude of several thousands, say $k > 5000$. Strictly speaking, this issue cannot be simply bypassed through further dilution in experiment, because the dilution procedure will induce additional variability to the estimation (Wu et al., 2009). Second, in real fluctuation experiments, divergent culture sizes are always observed. In general, the observed data should include not only the number of mutant cells $k_1, \ldots, k_J$, but also the number of nonmutant cells $n_1, \ldots, n_J$ at time t. Although the log likelihood function can actually be rewritten as

$$\ell(m | k_1, \ldots, k_J, n_1, \ldots, n_J) = \sum_{i=1}^{J} \log\ p_{k_i}(m, \phi_i)$$

where $\phi_i = 1-n_0/n_i$ can be further approximated by 1 for each culture, the divergent culture sizes are in contradiction with the deterministic growth assumption for nonmutant cells using the Lea–Coulson formulation. Furthermore, it has been shown that the efficiency of estimation could be much reduced if assumed homogenous culture sizes actually diverge (Xiong et al., 2009).

To fill in the research gap, we propose an alternative maximum likelihood-based mutation rate estimator which can substantially improve the computational speed. Similar to the Bartlett formulation, we will assume stochastic growth for both nonmutant and