Theoretical Population Biology 111 (2016) 16–27

Contents lists available at ScienceDirect

Theoretical Population Biology

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

Inferring the demographic history from DNA sequences: An importance sampling approach based on non-homogeneous processes

S. Ait Kaci Azzou *, F. Larribe, S. Froda

ÉMoStA, Département de Mathématiques, Université du Québec à Montréal, Canada

ARTICLE INFO

Article history:
Received 13 December 2015
Available online 27 May 2016

Keywords:
Non-homogeneous processes
Importance sampling
Effective population size
Calibrated skywis plot
Coalescent process

ABSTRACT

In Ait Kaci Azzou et al. (2015) we introduced an Importance Sampling (IS) approach for estimating the demographic history of a sample of DNA sequences, the *skywis plot*. More precisely, we proposed a new nonparametric estimate of a population size that changes over time. We showed on simulated data that the *skywis plot* can work well in typical situations where the effective population size does not undergo very steep changes. In this paper, we introduce an iterative procedure which extends the previous method and gives good estimates under such rapid variations. In the *iterative calibrated skywis plot* we approximate the effective population size by a piecewise constant function, whose values are re-estimated at each step. These piecewise constant functions are used to generate the waiting times of non homogeneous Poisson processes related to a coalescent process with mutation under a variable population size model. Moreover, the present IS procedure is based on a modified version of the Stephens and Donnelly (2000) proposal distribution. Finally, we apply the *iterative calibrated skywis plot* method to a simulated data set from a rapidly expanding exponential model, and we show that the method based on this new IS strategy correctly reconstructs the demographic history.

1. Introduction

The methods for estimating the demographic history from gene sequences using coalescent theory can be classified into two categories: parametric and nonparametric. The parametric methods require a definite analytic demographic model which describes the changes in the population size. Such methods are typically based on importance sampling or Markov Chain Monte Carlo (MCMC) sampling and infer the demographic history by estimating the demographic parameters (Slatkin and Hudson, 1991; Kuhner et al., 1998; Drummond et al., 2002). Because it is not known in advance which demographic model fits the sampled gene sequences, nonparametric and semi-parametric methods for inferring the demographic history from sequence data or from an estimated genealogy have been developed; see for example Pybus et al. (2000), and Fu (1994). In practice, the result of a nonparametric method can be used as a preliminary estimate that could be supplemented by a parametric analysis.

Among the most known nonparametric methods using coalescence theory to estimate the effective population size, we mention the family of *skyline plot* methods that was introduced by Pybus et al. (2000), referred in the literature as the *classical skyline plot*. This first method produces a piecewise reconstruction of the demographic history which is considered as quite noisy. Therefore, several extensions have been proposed in order to improve the performance of the estimator of the demographic history. In what follows, we mention the most popular among them.

Strimmer and Pybus (2001) developed the *generalized skyline plot* estimate based on the Akaike Information Criterion correction (AIC) by allowing to cumulate multiple coalescent events. Later, important developments were obtained in a Bayesian framework. Thus, Drummond et al. (2005) and Opgen-Rhein and Fahrmeir (2005) use multiple change-point (MCP) models to estimate population size dynamics. Also, in order to achieve temporal smoothing of the effective population size, Minin et al. (2008) propose an alternative to change-point modeling that resorts to Gaussian Markov random fields. This last method does not require to set a prior total number of change points. Finally, Heled and Drummond (2008) introduce the *extended Bayesian skyline plot*, which permits the analysis of multiple unlinked loci, leading to an improvement in the reliability of the demographic inference and a substantial reduction in estimation error.

* Corresponding author.
E-mail addresses: ait_kaci_azzou.sadoune@uqam.ca (S. Ait Kaci Azzou), larribe.fabrice@uqam.ca (F. Larribe), froda.sorana@uqam.ca (S. Froda).

http://dx.doi.org/10.1016/j.tpb.2016.05.004
0040-5809/

Ait Kaci Azzou et al. (2015) developed a new method (*skywis plot*) based on coalescent theory in order to estimate the effective population size. They are using an efficient importance sampling scheme where the estimate comes to an average over a large number of simulated genealogies. In this approach, one computes a weighted average of the effective population sizes on specific time intervals (*epochs*), where the genealogies that better agree with the data are given more weight. Moreover, Ait Kaci Azzou et al. (2015) illustrated by simulation that the *skywis plot* correctly reconstructs the recent demographic history under scenarios where the slope of the population size is not too steep.

The *skywis plot* is essentially a data driven approach where the genealogies are simulated according to a constant population size scenario. By contrast, in this paper, we propose to improve the *skywis plot* method by simulating genealogies according to an approximating piecewise constant function. First, we propose to resort to some prior information about the effective population size at specific times in the past (*calibrated skywis plot*) and further we introduce an iterative procedure (*iterative calibrated skywis plot*) which requires no additional information. The assumption of having prior information about the effective population size at different times is realistic in practice. Indeed, it is possible to use the structure of heterochronous sequences (serial sampling) by introducing some information at the sampling times. A good example is given by Maretty et al. (2013) who used the information about the Viral Load[1] (V) obtained at the time of sequencing in the case of serial sampling from patients infected with rapidly evolving viruses like HIV. Maretty et al. (2013) assumed a linear relationship[2] between the Viral Load V, and the effective population size N_e at the sampling times $t_0, t_1, \ldots, t_S$ ($N_e = \lambda V$). Thus, between two successive sampling times, it is possible to estimate N_e so that $N_{t_i} = (V_{t_{i-1}} + V_{t_i})/2$ assuming neutral evolution. The *skywis plot* can be refined by using this type of additional information in the case of serial sampling.

Further, the *iterative calibrated skywis plot* uses the *skywis plot* at a first estimation step, and at each iteration we approximate the effective population size by a piecewise constant function which is re-estimated using the *calibrated skywis plot*. These piecewise constant functions are used when generating the waiting times of non-homogeneous Poisson processes. In order to implement this procedure, we had to introduce a modified version of the proposal distribution of Stephens and Donnelly (2000). Thus, we take into account the specific time structure in the case of variable population size. This issue has not yet been considered in the literature on estimating the demographic history, and sets apart our paper.

2. Preliminaries

2.1. Coalescent theory: constant population size versus variable population size

Coalescent theory allows one to produce genealogies relating the sampled sequences according to a large class of population genetic models. In its simplest form, the coalescent process (Kingman, 1982) provides a model for the genealogy assuming a single, isolated and panmictic population (e.g. a Wright–Fisher model), and a constant population size; for more details see, for example, Nordborg (2007), Hein et al. (2005), and Wakeley (2008). This classical coalescent framework can be extended to include simple deviations from the idealized Wright–Fisher model, like recombination, fluctuating population size, population structure, and selection. In this paper, we focus on a single extension of the coalescent, namely variable population size.

The case of non-constant (variable) population size requires to introduce the concept of effective population size, $N_e(t), t > 0$; the time t is rescaled in units of N generations. If the population size is constant in time, N, then $N_e(t) \equiv N$. The effective population size reflects the number of individuals that contribute offsprings to the descendant generation and is almost always smaller than the census population size.

The variable population size coalescent model for contemporary gene sequences has been introduced by Griffiths and Tavaré (1994) and Donnelly and Tavaré (1995). In this case, the coalescence times $T_2, T_3, \ldots, T_n$ do not follow independent exponential distributions, which is a major difference with the constant population size model. In what follows we discuss the impact of variable population size on the time to the most recent ancestor.

Let $N_e(0) = N$ and we assume that, relative to the population size N at time 0, the size of the population time t units ago is $\nu(t) = N_e(t)/N$. Further, let $\Lambda(t)$ be the cumulative coalescent rate over time relatively to the rate at time $t = 0$:

$$\Lambda(t) = \int_0^t \frac{1}{\nu(u)} du, \tag{1}$$

and let $A_n(t)$ be the process that counts the number of ancestors at time t of a sample of size n in the case of constant population size. It is known that $\{A_n(t), t \geq 0\}$ is a pure death process that moves from state k to state $k-1$ at rate $k(k-1)/2$ (Griffiths and Tavaré, 1994). Further, let $\tilde{A_n}(t)$ be the process that counts the number of ancestors at time t of a sample of size n in the case of variable population size; then the process $\tilde{A_n}(t)$ jumps from $\tilde{A_n}(t) = k$ to $k-1$ at rate $k(k-1)/2\nu(t)$, and is a non-homogeneous death process. The non-homogeneous process $\{\tilde{A_n}(t), t \geq 0\}$ can be written as a function of $A_n(t)$ as follows (Tavaré and Zeitouni, 2004):

$$\tilde{A_n}(t) = A_n(\Lambda(t)), \quad t \geq 0.$$

For example, if the effective population size decreases from the present to the past, then $\nu(t) \leq 1$, $\Lambda(t) \geq t$, and $\tilde{A_n}(t) \leq A_n(t)$. As a result

- the total time required to find the most recent common ancestor in a small population is shorter than in a large one;
- the topology of the tree in the case of variable population size is same as in the constant population size, but its time scale has to be changed to account for the fluctuations in the population.

Then, in the case of a piecewise constant model, and in some interval Δt_c where the ratio $N_e(\Delta t_c)/N = \nu(\Delta t_c) = \delta_c$, the local coalescence rate is given by $k(k-1)/2\delta_c$ instead of $k(k-1)/2$; time is compressed or stretched. For example, if $\delta_c = 0.5$, in this interval the effective population size is half the one at time $t = 0$ and the sequences would coalesce twice as fast as in the case of constant population size. In other words, the coalescence time is a compressed version (factor 0.5) of the time one would have in the case of constant population size.

2.2. The skywis plot method

In this section, we remind the principle of the *skywis plot* estimate introduced in Ait Kaci Azzou et al. (2015), when n gene sequences are available at time $t = 0$. The main idea behind this approach is to simulate a large number of genealogies using

[1] For example, in the case of the HIV virus, the viral load is the amount of HIV in a sample of blood. When the viral load is high, a patient has more HIV in his body.

[2] For more information about the relationship between the viral load and the effective population size, see, for example, Gutierrez et al. (2012).