Theoretical Population Biology 108 (2016) 24–35

Contents lists available at ScienceDirect

Theoretical Population Biology

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

Isolation-by-distance-and-time in a stepping-stone model

Nicolas Duforet-Frebourg *, Montgomery Slatkin

Department of Integrative Biology, University of California Berkeley, Berkeley, CA 94720, United States

ARTICLE INFO

Article history:
Received 7 August 2015
Available online 21 November 2015

Keywords:
Isolation-by-distance
Ancient DNA
Coalescence times
Principal component analysis

ABSTRACT

With the great advances in ancient DNA extraction, genetic data are now obtained from geographically separated individuals from both present and past. However, population genetics theory about the joint effect of space and time has not been thoroughly studied. Based on the classical stepping-stone model, we develop the theory of Isolation by distance and time. We derive the correlation of allele frequencies between demes in the case where ancient samples are present, and investigate the impact of edge effects with forward-in-time simulations. We also derive results about coalescent times in circular and toroidal models. As one of the most common ways to investigate population structure is principal components analysis (PCA), we evaluate the impact of our theory on PCA plots. Our results demonstrate that time between samples is an important factor. Ancient samples tend to be drawn to the center of a PCA plot.

Published by Elsevier Inc.

1. Introduction

Geography plays a central role in the pattern of genetic differentiation within a species. Seminal work on describing the evolution of continuous populations was done by Wright and Malécot. They studied genetic differentiation and inbreeding in continuously distributed populations (Wright, 1943; Malécot, 1948). The resulting idea is that, under the assumption of local dispersion, genetic differentiation accumulates with distance. This pattern of genetic structure is called Isolation-By-Distance (IBD), which is detected by computing measures of differentiation such as F_{ST} (Wright, 1943; Nei, 1973; Weir and Cockerham, 1984), or correlation coefficients (Malécot, 1955; Kimura and Weiss, 1964). Understanding the effect of geographic distance on population structure is an important task for population geneticists, as it is a source of neutral genetic variation (Slatkin, 1985; Rousset, 1997). Furthermore, IBD has been observed in humans and many other species (Sharbel et al., 2000; Castric and Bernatchez, 2003; Ramachandran et al., 2005; Hellberg, 2009; Karakachoff et al., 2015).

The role of geography in neutral genetic variation has been widely studied partly because of the many population genetic studies of individuals sampled from different locations in present-day populations. Because of the development of methods for sequencing DNA from fossils, genomes of individuals alive at previous times are now available to bring new information about the evolutionary processes that affected a species in the past. Since the first studies of ancient DNA (aDNA) three decades ago (Higuchi et al., 1984; Pääbo, 1985), techniques to retrieve DNA molecules from ancient bones have tremendously developed Pääbo et al. (2004).

In modern evolutionary biology, the similarity of differentiation in space and time has been recognized (Depaulis et al., 2009; Andrello et al., 2011; Teacher et al., 2011). Theoretical developments predict the effect of time on F_{ST} and related quantities (Skoglund et al., 2014). Epperson (2000) studied patterns of isolation by distance and time in ecology by using stochastic spatial time series and identity by descent probabilities. However such theoretical studies remain scarce.

The effect of separation in time can be studied using classical statistical methods in population genetics, such as principal component analysis (PCA) (Patterson et al., 2006). PCA is widely used to determine relatedness between individuals, and is a convenient way to represent geographic patterns (Novembre et al., 2008). But PCA can also capture the differentiation between ancient and modern samples: the percentage of variance explained by time can be expressed on the same scale as the percentage of variance explained by geography (Skoglund et al., 2014). Unfortunately, PCA does not give a complete picture of how quantities such as F_{st} and correlation coefficients evolve in time and space.

In this article we generalize the theory of IBD to allow for difference in the times at which different individuals are sampled. We call this the theory of isolation by distance and time (IBDT). We base our work on the stepping-stone model of Kimura (1953) and add to the theoretical results already derived for this model

* Corresponding author.
E-mail address: duforetn@berkeley.edu (N. Duforet-Frebourg).

http://dx.doi.org/10.1016/j.tpb.2015.11.003
0040-5809/Published by Elsevier Inc.

(Kimura and Weiss, 1964; Weiss and Kimura, 1965; Maruyama, 1971a; Nagylaki, 1983; Cox and Durrett et al., 2002; De and Durrett, 2007). We start by briefly reviewing the original results for the infinite stepping-stone model at equilibrium and the decay of correlation of allele frequencies with distance. Then, we extend the original work to derive the correlation between individuals separated by distance and time. We perform simulations that show the validity of the analytic results, even in the case of a finite number of populations where some demes are subject to edge effect. We also derive the expected coalescence times between samples separated by time and space in circular and toroidal models (Slatkin, 1991, 1993). Finally we consider the consequences of IBDT on PCA in the common case of a dataset made up of a large proportion genomes from present-day individuals and few ancient genomes.

2. The stepping-stone model

The stepping-stone model describes the distribution of allele frequencies in an infinite set of demes in different locations of the space represented by Cartesian coordinates. We start by describing the 1-dimensional case. Let $p(k)$ be the frequency of one allele at a bi-allelic locus in population k and $\bar{p}$ be the average allele frequency. In each generation, $p(k)$ is updated with the following three steps (Crow and Kimura et al., 1970):

- Exchange a proportion m_i of migrants with demes at a distance i.
- Exchange a proportion m_∞ of migrants with a deme that has fixed allele frequency $\bar{p}$. The meaning of this step is discussed later.
- Sample gametes of the next generation in the population.

In the case considered by Kimura and Weiss (1964), migrants are exchanged only between neighboring locations in the first step, so that $m_i = 0, i > 1$. The second step consists of the exchange of migrants with an external population at rate m_∞. This event is equivalent to reversible mutation with equilibrium allele frequency m_∞. In general $m_1 \gg m_\infty$. Random sampling of step 3 is represented by a random change in the allele frequency $\epsilon(k)$, with $E[\epsilon(k)] = 0$, and $E[\epsilon(k)^2] = p(k)(1-p(k))/2N_e$, where N_e is the effective population size of a deme (Wright, 1940; Kimura and Crow, 1963).

Our interest is in the changes in allele frequency in one generation. We consider $\tilde{p}(k) = \bar{p} - p(k)$, the deviation from the average frequency. Given these three steps,

$$\tilde{p}'(k) = \left(1 - \sum_{i=1}^{\infty} m_i - m_\infty\right)\tilde{p}(k) + \frac{m_1}{2}(\tilde{p}(k-1) + \tilde{p}(k+1)) + \frac{m_2}{2}(\tilde{p}(k-2) + \tilde{p}(k+2)) + \cdots + \epsilon(k). \quad (1)$$

To simplify the notation, we define the operators S and L,

$$S\tilde{p}(k) = \tilde{p}(k+1), S^i\tilde{p}(k) = \tilde{p}(k+i), \quad i \in \mathbb{Z}, \quad (2)$$

$$L = m_0 S^0 + \sum_{i=1}^{\infty} \frac{m_i}{2}(S^i + S^{-i}), \quad (3)$$

where $m_0 = 1 - \sum_{i=1}^{\infty} m_i - m_\infty$, so that,

$$\tilde{p}'(k) = L\tilde{p}(k) + \epsilon(k). \quad (4)$$

The quantity of interest in this model is the correlation of allele frequencies between two demes at locations k_1 and k_2. Let $r(k)$ be the correlation coefficient of allele frequencies between populations that are k steps apart. Assuming equilibrium, we have

$$r(k) = \frac{\rho(k)}{\rho(0)} = \frac{E[\tilde{p}(k_1)\tilde{p}(k_2)]}{\rho(0)} = \frac{E[L\tilde{p}(k_1)L\tilde{p}(k_2)]}{\rho(0)}, \quad (5)$$

where $\rho(k)$ is the covariance in frequencies in demes k steps apart. The within-population variance of allele frequencies, $\rho(0)$, value is detailed in Weiss and Kimura (1965). The mathematical treatment of Eq. (5) by Weiss and Kimura (1965) using the spectral representation of a correlation (Doob, 1953) gives the general formula

$$r(k) = \frac{C}{2\pi}\int_0^{2\pi} \frac{\cos(k\theta)d\theta}{1 - \left[\sum_{i=0}^{\infty} m_i \cos(i\theta)\right]^2}, \quad (6)$$

where C is the normalizing constant chosen so that $r(0) = 1$.

In the case of a stepping-stone model where migrants are exchanged only between neighboring demes ($m_i = 0, i > 1$), r can be approximated by an exponential function of k:

$$r(k) = e^{-\sqrt{\frac{2m_\infty}{m_1}}k}, \quad (7)$$

as detailed in Kimura and Weiss (1964). This simple formula conveys the important idea that in one dimension, the correlation of allele frequencies between populations decays exponentially with distance. In the 2-dimensional and 3-dimensional cases, the correlation function is more difficult to approximate. Using modified Bessel function, it has been shown that correlation at a given distance is lower in these cases than in the 1-dimensional case (Weiss and Kimura, 1965).

3. Isolation-by-distance-and-time

3.1. 1-dimensional case

We are here interested in the case where genetic samples are collected from demes that are in different locations and at different times (measured in generations). Let $\rho(k, t)$ be the covariance between allele frequencies of two demes separated by k steps and t generations. We denote the coordinates of these demes by (k_1, t_1) and (k_2, t_2), and the deviations in allele frequencies $\tilde{p}(k_1)^{(t_1)}$ and $\tilde{p}(k_2)^{(t_2)}$. Since we assume the distribution of allele frequencies is stationary in both time (equilibrium distribution) and space (all migration rates are equal), we can consider these coordinates to be $(0, 0)$ and (k, t) with no loss of generality. Following previous notation

$$\rho(k, t) = E[\tilde{p}(k_1)^{(t_1)}\tilde{p}(k_2)^{(t_2)}] = E[\tilde{p}(k)^{(t)}\tilde{p}(0)^{(0)}]. \quad (8)$$

To characterize the evolution of the covariance between allele frequencies with respect to time t, we iteratively apply the operator L defined in Eq. (3). This operation describes the potential trajectories of an allele. This process leads to

$$\rho(k, t) = L^t\rho(k) \quad (9)$$

with $\rho(k) = \rho(k, 0)$ (see Appendix A).

Let $r(k, t)$ be the correlation between allele frequencies of two demes separated by k steps and t generations, Eqs. (5) and (9), combined with the general formula of Eq. (6) gives

$$r(k, t) = \frac{C}{2\pi}\int_0^{2\pi} \frac{\left[\sum_{i=0}^{\infty} m_i \cos(i\theta)\right]^t \cos(k\theta)d\theta}{1 - \left[\sum_{i=0}^{\infty} m_i \cos(i\theta)\right]^2} \quad (10)$$

and the constant C is set such that $r(0, 0) = 1$ (Appendix B).

This equation reduces to

$$r(k, t) = \frac{C}{2\pi}\int_0^{2\pi} \frac{[1 - m_1 - m_\infty + m_1\cos(\theta)]^t \cos(k\theta)d\theta}{1 - (1 - m_1 - m_\infty + m_1\cos(\theta))^2} \quad (11)$$