Journal of Theoretical Biology 402 (2016) 158–170

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

# An informational transition in conditioned Markov chains: Applied to genetics and evolution

Lei Zhao [a], Martin Lascoux [a,b], David Waxman [a,*]

[a] Centre for Computational Systems Biology, Fudan University, 220 Handan Road, Shanghai 200433, PR China
[b] Evolutionary Biology Center, Department of Ecology and Genetics, Uppsala University, Uppsala 75236, Sweden

## HIGHLIGHTS

- We reveal an informational transition in conditioned Markov chains.
- Two observations may not determine a population's behaviour at all intermediate times.
- This occurs when the interval between observations exceeds a characteristic time, $T_c$.
- We determine $T_c$ for genetic examples involving Wright–Fisher models.

## ARTICLE INFO

*Article history:*
Received 1 August 2015
Received in revised form
19 February 2016
Accepted 17 April 2016
Available online 19 April 2016

*Keywords:*
Random genetic drift
Population genetics theory
Frequency trajectories
Conditional distribution
Ancient DNA

## ABSTRACT

In this work we assume that we have some knowledge about the state of a population at two known times, when the dynamics is governed by a Markov chain such as a Wright–Fisher model. Such knowledge could be obtained, for example, from observations made on ancient and contemporary DNA, or during laboratory experiments involving long term evolution. A natural assumption is that the behaviour of the population, between observations, is related to (or constrained by) what was actually observed. The present work shows that this assumption has limited validity. When the time interval between observations is larger than a characteristic value, which is a property of the population under consideration, there is a range of intermediate times where the behaviour of the population has reduced or no dependence on what was observed and an equilibrium-like distribution applies. Thus, for example, if the frequency of an allele is observed at two different times, then for a large enough time interval between observations, the population has reduced or no dependence on the two observed frequencies for a range of intermediate times. Given observations of a population at two times, we provide a general theoretical analysis of the behaviour of the population at all intermediate times, and determine an expression for the characteristic time interval, beyond which the observations do not constrain the population's behaviour over a range of intermediate times. The findings of this work relate to what can be meaningfully inferred about a population at intermediate times, given knowledge of terminal states.



## 1. Introduction

Two major activities of population genetics involve reconstructing an extant population's history and inferring the evolutionary forces that influenced a population's behaviour. Until recently, the only approach to these activities was via contemporary DNA polymorphism data. However, the development of ancient DNA (aDNA) techniques has allowed new perspectives. It is now possible to obtain temporal samples on deep time-scales. In some cases, these techniques have led to important new insights. For example, the sequencing of Neanderthal fossils (Green et al., 2010; Sankararaman et al., 2014) lends strong support to mixing between Neanderthals and modern humans, while previous analysis of contemporary DNA had only suggested this (Plagnol and Wall, 2006; Wall et al., 2009). The availability of aDNA data has also led to a renewal of population genetic studies of temporal samples (e.g. Malaspinas et al., 2012; Sjödin et al., 2014; Skoglund et al., 2014; Sams et al., 2015). Another area of evolutionary genetics where temporal samples are playing and important role is the blooming field of Long Term Evolutionary Experiments (see Long et al., 2015 for a recent review and Tenaillon et al., 2016 for the longest experiment so far, namely 50,000 generations).

* Corresponding author.
*E-mail address:* davidwaxman@fudan.edu.cn (D. Waxman).

http://dx.doi.org/10.1016/j.jtbi.2016.04.021
0022-5193/

In the present work we consider what can be inferred about a population's behaviour, when the dynamics is governed by a Markov chain (see e.g., Kemeny and Snell, 1976), for example a Wright–Fisher model (Gale, 1990; Ewens, 2004). We work under the assumption that we have knowledge about the state of the population at two different times, from two separate observations of the population. This knowledge may be highly detailed, for example, an accurate estimate of the frequency of a focal allele at both times, or less detailed, such as an estimate of the frequency of the allele at one time, and simply that the allele is segregating at the other time. Such knowledge about the state of the system could come, for example, from the analysis of samples of ancient and contemporary DNA.

A very natural assumption, given observations of a population at two different times, is that what occurs in the population, between the two times, is related to (or constrained by) what is actually observed at the two times. It turns out that this is not invariably the case. We shall provide examples where, when the time interval is larger than a characteristic value, $T_c$, *there is a range of intermediate times, over which the behaviour of the population is independent of the observations made*. We shall investigate this phenomenon by exploring the influence of the size of the time interval on properties of the population at intermediate times.

We can summarise our basic findings (in a genetics context) as follows, assuming that the state of a population has been observed at two different times (we shall refer to these states as terminal states).

1. When the time-interval between observations is smaller than a characteristic time, $T_c$, which is a property of the particular population under consideration, the distribution of the state of a population depends on the observed states, for all intermediate times.
2. When the time-interval between observations is larger than a characteristic time, $T_c$, there is a range of intermediate times where the distribution of the state of the population contains reduced or no dependence on the observed terminal states. During this range of intermediate times, an equilibrium-like distribution of the population applies, independent of the observed states.

These findings lead to the following picture. When the time-interval between observations is increased, from below a characteristic value, to above it, an *informational transition* occurs: the population changes from containing information about the terminal states for all intermediate times, to having restricted or no information about the terminal states *for a range* of intermediate times.

## 2. Choice of examples

The informational transition involves time, and we consider populations with discrete generations, labelled by $t = 0, 1, 2, \ldots$ There are a number of possible examples that can illustrate the transition; we shall devote most attention to one of the simplest cases, to show the basic phenomenon, unencumbered by too many details, yet sufficiently realistic to be of biological interest in its own right.

We consider a population of haploid organisms with two alleles, and assume that mutation can be neglected between two times of observation of the population. We further assume that the frequency of the focal allele is recorded at both observation times and that the observations are informative about the state of the population at the final time, in the sense that the frequencies at both observations are segregating. It turns out that under neutrality, or weak genic/additive selection of the focal allele, there is only a small probability that the focal allele is segregating after a time interval that exceeds the characteristic value, $T_c$. Thus to explore the informational transition we assume a mechanism that enhances the chance a population stays polymorphic, so that after a time of $T_c$ there is an appreciable chance the focal allele is still segregating.

Negative frequency-dependent selection is a possible mechanism that enhances the time of segregation (Robertson, 1962; Zhao and Waxman, 2016), and we have adopted this form of selection as the main example in this work. This form of selection is important in its own right since "…frequency-dependent selection might often be the selective mechanism underlying balanced polymorphisms" (Mitchell-Olds et al., 2007).

Other possible mechanisms that provide such an enhancement of the time of segregation include some cases of spatial population structure and overdominant selection.[1] Later in the paper we shall use populations with these mechanisms as additional examples where the informational transition may occur. Given that the transition may occur with both frequency-dependent and overdominant selection, the results we present have relevance, more generally, to balancing selection, and work by Andrés et al. (2009) provides an indication of how common such selection is within the human genome. Indeed recent genome scans in humans and chimpanzees have revealed the presence of trans-species polymorphisms, and there may be other cases of balancing selection, maintaining variation over millions of years, that remain unrecognised (Gao et al., 2015).

## 3. Illustrating the informational transition

We shall illustrate the informational transition, described above, by focussing on a highly indicative feature of the focal allele in a haploid model, namely its *mean frequency trajectory*. To proceed, we first fully define the properties of the model we shall use for this purpose.

### 3.1. Wright–Fisher model with frequency-dependent selection

Consider a finite population of one locus haploid organisms with two alleles, one of which we designate the focal allele. We make the assumption that mutation can be neglected during the finite time interval separating two observations.[2]

We shall incorporate negative frequency-dependent selection into the model by taking the selection coefficient of the focal allele, when at a frequency of $x$, to be $s(x)$. The relative fitness of the focal allele is then $1 + s(x)$ while that of the other allele is 1. We assume that the selection coefficient $s(x)$ is positive at low frequencies, and as the frequency is increased, $s(x)$ decreases in value, passes through zero at an intermediate frequency ($0 < x < 1$), and then becomes negative at yet higher frequencies. Thus the focal allele has a selective advantage at low frequencies, a selective disadvantage at high frequencies, and, in a very large population, has a balanced polymorphism at the intermediate frequency where $s(x)$ vanishes (cf. Zhao and Waxman, 2016). We adopt the very simplest form for the selection coefficient with these properties,

[1] We note that with overdominance, there is a slowing down or "retardation" of the rate at which both fixation and loss can occur (Robertson, 1962). Furthermore, some cases of overdominance are mathematically equivalent to frequency-dependent selection (Takahata and Nei, 1990), and hence retardation may also apply here. Some closely related issues are considered by Zhao and Waxman (2016).

[2] The neglect of mutation is not an essential assumption, as we discuss later. The neglect of mutation is useful in that it forces us to address a common complication, namely the role of absorbing states.