Theoretical Population Biology 103 (2015) 27–37

Contents lists available at ScienceDirect

Theoretical Population Biology

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

# Looking down in the ancestral selection graph: A probabilistic approach to the common ancestor type distribution

Ute Lenz [a], Sandra Kluth [b,1], Ellen Baake [b,*], Anton Wakolbinger [a]

[a] Institut für Mathematik, Goethe-Universität Frankfurt, Box 111932, 60054 Frankfurt am Main, Germany
[b] Faculty of Technology, Bielefeld University, Box 100131, 33501 Bielefeld, Germany

ARTICLE INFO

*Article history:*
Received 7 August 2014
Available online 16 April 2015

*Keywords:*
Common ancestor type distribution
Ancestral selection graph
Lookdown graph
Pruning
Wright–Fisher diffusion with selection and mutation

ABSTRACT

In a (two-type) Wright–Fisher diffusion with directional selection and two-way mutation, let $x$ denote today's frequency of the beneficial type, and given $x$, let $h(x)$ be the probability that, among all individuals of today's population, the individual whose progeny will eventually take over in the population is of the beneficial type. Fearnhead (2002) and Taylor (2007) obtained a series representation for $h(x)$. We develop a construction that contains elements of both the ancestral selection graph and the lookdown construction and includes pruning of certain lines upon mutation. Besides being interesting in its own right, this construction allows a transparent derivation of the series coefficients of $h(x)$ and gives them a probabilistic meaning.



## 1. Introduction

The understanding of ancestral processes under selection and mutation is among the fundamental challenges in population genetics. Two central concepts are the ancestral selection graph (ASG) and the lookdown (LD) construction. The ancestral selection graph (Krone and Neuhauser, 1997; Neuhauser and Krone, 1997; see also Shiga and Uchiyama, 1986 for an analogous construction in a diffusion model with spatial structure) describes the set of lines that are potential ancestors of a sample of individuals taken from a present population. In contrast, the lookdown construction (Donnelly and Kurtz, 1999a,b) is an integrated representation that makes all individual lines in a population explicit, together with the genealogies of arbitrary samples. See Etheridge (2011, Chapter 5) for an excellent overview of the area.

Both the ASG and the LD are important theoretical concepts as well as valuable tools in applications. Interest is usually directed towards the genealogy of a sample, backwards in time until the most recent common ancestor (MRCA). However, the ancestral line that continues beyond the MRCA into the distant past is of considerable interest on its own, not least because it links the genealogy (of a sample from a population) to the longer time scale of phylogenetic trees. The extended time horizon then shifts attention to the asymptotic properties of the ancestral process. The stationary type distribution on the ancestral line may differ substantially from the stationary type distribution in the population. This mirrors the fact that the ancestral line consists of those individuals that are successful in the long run; thus, its type distribution is expected to be biased towards the favourable types.

When looking at the evolution of the system in (forward) time $[0, \infty)$, one may ask for properties of the so-called *immortal line*, which is the line of descent of those individuals whose offspring eventually takes over the entire population. In other words, the immortal line restricted to any time interval $[0, t]$ is the common ancestral line of the population back from the far future. It then makes sense to consider the type of the immortal line at time 0. To be specific, let us consider a Wright–Fisher diffusion with two types of which one is more and one is less fit. The *common ancestor type (CAT) distribution* at time 0, conditional on the type frequencies $(x, 1-x)$, then has weights $(h(x), 1-h(x))$, where $h(x)$ is the probability that the population ultimately consists of offspring of an individual of the beneficial type, when starting with a frequency $x$ of beneficial individuals at time 0.

The quantity $h(x)$ can also be understood as the limiting probability (as $s \to \infty$) that the ancestor at time 0 of an individual sampled from the population at the future time $s$ is of the beneficial type, given that the frequency of the beneficial type at time 0 is $x$. Equivalently, $h(x)$ is the limiting probability (as $s \to \infty$) that the ancestor at the past time $-s$ of an individual sampled from

* Corresponding author.
*E-mail addresses:* lenz@math.uni-frankfurt.de (U. Lenz), sandra.kluth@uni-bielefeld.de (S. Kluth), ebaake@techfak.uni-bielefeld.de (E. Baake), wakolbin@math.uni-frankfurt.de (A. Wakolbinger).
[1] Present address: Department of Internal Medicine I and Center for Integrated Oncology, University Hospital Cologne, Kerpener Strasse 62, 50937 Köln, Germany.

http://dx.doi.org/10.1016/j.tpb.2015.01.005
0040-5809/

the population at time 0 is of the beneficial type, given that the frequency of the beneficial type at time $-s$ was $x$.

Fearnhead (2002) computed the common ancestor type distribution for time-stationary type frequencies, representing it in the form $\int_0^1 (h(x),\, 1-h(x))\pi\,(dx)$ (where $\pi$ is Wright's equilibrium distribution) and calculating a recursion for the coefficients of a series representation of $h(x)$. Later, $h(x)$ has been represented in terms of a boundary value problem (Taylor, 2007; Kluth et al., 2013), see also Section 7.

In the case without mutations (in which $h(x)$ coincides with the classical fixation probability of the beneficial type starting from frequency $x$), Mano (2009) and Pokalyuk and Pfaffelhuber (2013) have represented $h(x)$ in terms of the *equilibrium ASG*, making use of a time reversal argument (see Section 2.2). However, the generalisation to the case with mutation is anything but obvious. One purpose of this article is to solve this problem. A key ingredient will be a combination of the ASG with elements of the lookdown construction, which also seems of interest in its own right.

The paper is organised as follows. In Section 2, we start by briefly recapitulating the ASG (starting from the Moran model for definiteness). We then recall the Fearnhead–Taylor representation of $h(x)$ and give its explanation in terms of the equilibrium ASG in the case without mutations, inspired by Pokalyuk and Pfaffelhuber (2013). In Section 3, we prepare the scene by ordering the lines of the ASG in a specific way; in Section 4, we then represent the ordered ASG in terms of a fixed arrangement of levels, akin to a lookdown construction. In Section 5, a pruning procedure is described that reduces the number of lines upon mutation. The stationary number of lines in the resulting pruned LD–ASG will provide the desired connection to the (conditional) common ancestor type distribution. Namely, the tail probabilities of the number of lines appear as the coefficients in the series representation. In Section 6, the graphical approach will directly reveal various monotonicity properties of the tail probabilities as functions of the model parameters, which translate into monotonicity properties of the common ancestor type distribution. Section 7 is an add-on, which makes the connection to Taylor's boundary value problem for $h(x)$ explicit; Section 8 contains some concluding remarks.

## 2. Concepts and models

### 2.1. The Moran model and its diffusion limit

Let us consider a haploid population of fixed size $N \in \mathbb{N}$ in which each individual is characterised by a type $i \in S := \{0, 1\}$. An individual of type $i$ may, at any instant in continuous time, do either of two things: it may reproduce, which happens at rate 1 if $i = 1$ and at rate $1+s_N$, $s_N \geq 0$, if $i = 0$; or it may mutate to type $j$ at rate $u_N \nu_j$, $u_N \geq 0$, $0 \leq \nu_j \leq 1$, $\nu_0 + \nu_1 = 1$. If an individual reproduces, its single offspring inherits the parent's type and replaces a randomly chosen individual, maybe its own parent. Concerning mutations, $u_N$ is the total mutation rate and $\nu_j$ the probability of a mutation to type $j$. Note that the possibility of silent mutations from type $j$ to type $j$ is included.

The Moran model has a well-known graphical illustration as an interacting particle system (cf. Fig. 1). The individuals are represented by horizontal line pieces, with forward time running from left to right in the figure. Arrows indicate reproduction events with the parent at its tail and the offspring at its head. For later use, we decompose reproduction events into neutral and selective ones. Neutral arrows appear at rate $1/N$, selective arrows (those with a star-shaped arrowhead in Fig. 1) at rate $s_N/N$ per ordered pair of lines, irrespective of their types. The rates specified above are obtained by the convention that neutral arrows may be used by all individuals, whereas selective arrows may only be used by type-0 individuals and are ignored otherwise. Mutations to type 0 are marked by circles, mutations to type 1 by crosses.

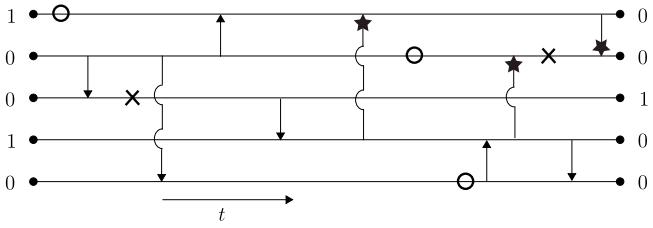


**Fig. 1.** The Moran model with two-way mutation and selection. The types are indicated for the initial population (left) and the final one (right). Crosses represent mutations to type 1, circles mutations to type 0. Selective events are depicted as arrows with star-shaped heads.

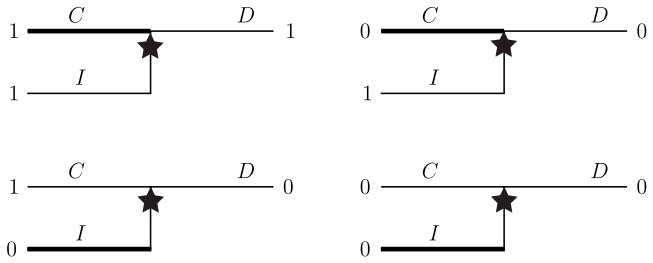


**Fig. 2.** Incoming branch ($I$), continuing branch ($C$), and descendant ($D$). The ancestor is marked bold.

The usual diffusion rescaling in population genetics is applied, i.e. rates are rescaled such that $\lim_{N\to\infty} Ns_N = \sigma$ and $\lim_{N\to\infty} Nu_N = \theta$, $0 \leq \sigma, \theta < \infty$, and time is sped up by a factor of $N$. Let $X_t$ be the frequency of type-0 individuals at time $t$ in this diffusion limit. Then, the process $(X_t)_{t\in\mathbb{R}}$ is a Wright–Fisher diffusion which is characterised by the drift coefficient $a(x) = (1-x)\theta\nu_0 - x\theta\nu_1 + x(1-x)\sigma$ and the diffusion coefficient $b(x) = 2x(1-x)$. The stationary density $\pi$ is given by $\pi(x) = C(1-x)^{\theta\nu_1-1}x^{\theta\nu_0-1}\exp(\sigma x)$, where $C$ is a normalising constant (cf. Durrett, 2008, Chapters 7, 8 or Ewens, 2004, Chapters 4, 5).

### 2.2. The ancestral selection graph

The *ancestral selection graph* was introduced by Krone and Neuhauser (1997) and Neuhauser and Krone (1997) to construct samples from a present population, together with their ancestries, in the diffusion limit of the Moran model with mutation and selection. The basic idea is to understand selective arrows as unresolved reproduction events backwards in time: the descendant has two *potential ancestors*, the *incoming branch* (at the tail) and the *continuing branch* (at the tip), see also Fig. 2. The incoming branch is the ancestor if it is of type 0, otherwise the continuing one is ancestral. For a hands-on exposition, see Wakeley (2009, Chapter 7.1).

The ASG is constructed by starting from the (as yet untyped) sample and tracing back the lines of all potential ancestors. In the finite graphical representation, a neutral arrow that joins two potential ancestral lines appears at rate $2/N$ per currently extant pair of potential ancestral lines, then giving rise to a *coalescence event*, i.e. the two lines merge into a single one. In the same finite setting, a selective arrow that emanates from outside the current set of $n$ potential ancestral lines and hits this set appears at rate $n(N-n)s_N/N$. This gives rise to a *branching event*, i.e., viewed backwards in time, the line that is hit by the selective arrow splits into an incoming and continuing branch as described above. Thus, in the diffusion limit, since $N - n \sim N$ as $N \to \infty$, the process $(K_r)_{r\in\mathbb{R}}$, where $K_r$ is the number of lines in the ASG at time $r = -t$,