Applied Mathematics Senior Thesis Final Draft 1

Tim Stutz, Advisor Jo Hardin

April 6, 2012

Modeling the Evolution of Sexual Diploid Populations via a Stochastic Moran Process

1 Introduction

Here we mathematically model the evolution of the genetic composition of a finite population of diplod sexual organisms acting under natural selection. In particular, we restrict our study to one genetic loci with two distinct alleles as a base case application of our methods. We create a model that describes the evolution of the population's probability function; this gives the probability of seeing a certain genetic composition of the population at time t.

This formulation incorporates the Moran Process population model wherein individuals die randomly and are immediately replaced by an offspring of another randomly chosen set of individuals. This process has several important properties, the first being that it maintains a constant population size. Second, it allows for easy derivation of the transition rates, the rate at which an individual of a certain genotype will die off and be replaced with an individual of another certain genotype. Third, it is a continuous-time Markov process, which guarantees several highly useful properties, mainly that there exists a stationary probability distribution that the population will evolve under over an extended time frame. This population model has also been shown to have similar behaviour to other common approximations of population dynamics, mainly the Wright Fisher model and Kimura's diffusion equation for population genetics [1, 2].

We will be basing our model off of previous work that has been done on creating and analyzing a Moran Process applied to a finite population of haploid individuals [2]. The authors assumed that the dynamics of the population contained only Poisson events which allowed for the generation of transition rates from a given genotype composition to the next; these rates were then used to precisely describe the partial differential equation of the probability function that describes the composition of the fixed population in terms of the number of mutant alleles at a given time. Their model included natural selection and genetic drift, and was analyzed primarily via methods of probability generating functions. It also provides a straightforward and precise method for quantifying the probability function governing the population and the fixation probabilities of different alleles under differing initial conditions; we aim to produce similar results for the sexual diploid case.

Additionally, our approach allows for the study of populations of organisms with both multiple genetic loci and with more complex reproductive regime. In particular, it is our long-term goal to examine the disconnect that exists between the evolutionary forces of mutation and natural selection in ciliates due to their unique genome architecture [3].

Previous work relating to ciliates includes altered Hardy-Weinberg derivations, specifically for analyzing mutation-selection balance. These derivations include several assumptions about the ciliate population in question: the population is infinite, all sexual reproduction is done via random mating, and no additional genetic inflow is allowed [4, 5]. Further refinement of the Hardy-Weinberg model requires the removal of the infinite population assumption; this will allow for the creation of a stochastic process that can then be used to provide information on how the effects of mutation, natural selection, and genetic drift are altered by ciliates' unique genomic architecture [6]. In particular, we wish to study the effects that ciliate biology have on fixation probabilities for beneficial alleles via describing an ideal ciliate population under the methods used by [2].

Therefore our goals are to first to expand the previous Moran Process from haploid to diploid sexual organisms, then to include mutational pressures. This model will serve as a baseline comparison, a standard, for the expanded model that describes a ciliate population, again with mutational pressures. This serves two purposes: one, to precisely quantify any genetic advantages or disadvantages that are conferred to ciliates via their biology, and two, to further extend the applications of stochastic partial differential equations to population genetics, specifically as a more precise and tractable alternative to Kimura's diffusion equation.

2 Formulation of the model

2.1 Derivation of the diploid sexual master equation for the recessive case

Consider a diploid, sexual population of fixed size N containing two alleles at a given locus. We denote the wild type allele as A and the mutant allele as B and give allele A dominant fitness 1 and allele B recessive fitness $1+s$, $s>0$. We also denote n as the number of individuals with genotype BB and m the number with AB, hence $N-m-n$ is the number of individuals with AA.

Let X_t be a continuous birth-death stochastic process with Poissonian events for the population such that X_t takes on values $n, m \in \mathbb{Z}^+$. Note that the fixed population size also limits $n+m \leq N$. Therefore X_t describes the genotype composition of the population as above, specifically $X_t = (n, m)$ denotes that the population has n individuals of type BB and m individuals of type AB at time t.

We then let the transition rates, given by the function W, specifically as the probability density of the process changing from a given composition (n, m) to (n', m') in an infinitesimal time period $\Delta t \to 0$ be denoted as

$$
\begin{aligned}
W(n \to n+1, m \to m) &= W^{+0}(n,m), \\
W(n \to n, m \to m+1) &= W^{0+}(n,m), \\
W(n \to n-1, m \to m) &= W^{-0}(n,m), \\
W(n \to n, m \to m-1) &= W^{0-}(n,m), \\
W(n \to n+1, m \to m-1) &= W^{+-}(n,m), \\
W(n \to n-1, m \to m+1) &= W^{-+}(n,m), \\
W(n \to n, m \to m) &= W^{00}(n,m), \\
W(n \to n+k, m \to m+q) &= 0 \text{ if } |k| > 1 \text{ or } |q| > 1 \text{ or } k = q \neq 0. \qquad (1)
\end{aligned}
$$

The transition rates are limited because the population can only gain and lose one individual from each genotype in a given event; this rules out both W^{++} and W^{--}, as losing or gaining both a AB and BB individual would change the overall population size. The actual values of the transition rates will be derived later. In addition, we will later show that the term W^{00} becomes unnecessary.

Now let $P(n,m,t) = \mathbb{P}\{X_t = (n,m)\}$, the probability that the population is in state (n,m) at time t. We wish to derive how this distribution changes over time, particularly when any equilibria may exist. This approach also allows for the derivation of the fixation probabilities. We can then calculate the partial derivative of P with respect to time from the basic definition of the derivative,

$$
\begin{aligned}
\frac{\partial P(n,m,t)}{\partial t} &= \lim_{\Delta t \to 0} \frac{1}{\Delta t}\Big(P(n,m,t+\Delta t) - P(n,m,t)\Big) \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta t}\Big(\mathbb{P}\{X_{t+\Delta t} = (n,m)\} - \mathbb{P}\{X_t = (n,m)\}\Big) \qquad (2)
\end{aligned}
$$

We know via the previous arguments about the transition rates in (1) that only certain events are allowed to occur in the infinitesimally small time period $\Delta t \to 0$. In particular, these events are the death of a single individual and its replacement by the offspring of a randomly chosen pair in the population; these transitions are those with non-zero $W^{\pm\pm}$ terms in (1).

Knowing which events have positive probability in Δt, we can then expand $\mathbb{P}\{X_{t+\Delta t} = (n,m)\}$ via its conditional probabilities. The probability $\mathbb{P}\{X_{t+\Delta t} = (n,m)\}$ can also be thought of as the summation of all the conditional probabilities of moving from a given state (j,k) at time t to state (n,m) during the time period Δt times the probability of being in state (j,k) at time t. Formally, this is given as

$$\mathbb{P}\{X_{t+\Delta t} = (n,m)\} = \sum_{j,k\in\mathbb{Z}^+} \mathbb{P}\{X_{t+\Delta t} = (n,m) | X_t = (j,k)\}\mathbb{P}\{X_t = (j,k)\} \tag{3}$$

Note that the events that have positive probability of occurring in Δt, ie the events given in (1), are the only events that will contribute positive probability to the above summation. We then substitute (3) into (2) to get the expansion

$$\begin{aligned}\frac{\partial P(n,m,t)}{\partial t} = \lim_{\Delta t\to 0}\frac{1}{\Delta t}\Big(&\mathbb{P}\{X_{t+\Delta t} = (n,m)|X_t = (n,m+1)\}\mathbb{P}\{X_t = (n,m+1)\}\\ &+ \mathbb{P}\{X_{t+\Delta t} = (n,m)|X_t = (n+1,m)\}\mathbb{P}\{X_t = (n+1,m)\}\\ &+ \mathbb{P}\{X_{t+\Delta t} = (n,m)|X_t = (n+1,m-1)\}\mathbb{P}\{X_t = (n+1,m-1)\}\\ &+ \mathbb{P}\{X_{t+\Delta t} = (n,m)|X_t = (n-1,m+1)\}\mathbb{P}\{X_t = (n-1,m+1)\}\\ &+ \mathbb{P}\{X_{t+\Delta t} = (n,m)|X_t = (n-1,m)\}\mathbb{P}\{X_t = (n-1,m)\}\\ &+ \mathbb{P}\{X_{t+\Delta t} = (n,m)|X_t = (n,m-1)\}\mathbb{P}\{X_t = (n,m-1)\}\\ &+ \mathbb{P}\{X_{t+\Delta t} = (n,m)|X_t = (n,m)\}\mathbb{P}\{X_t = (n,m)\}\\ &- \mathbb{P}\{X_t = (n,m)\}\Big).\end{aligned}$$

We then swap $\mathbb{P}\{X_{t+\Delta t} = (n,m)|X_t = (n,m)\}$ in the above with its complement, $1 - \mathbb{P}\{X_{t+\Delta t} \neq (n,m)|X_t = (n,m)\}$; this gives

$$\begin{aligned}\frac{\partial P(n,m,t)}{\partial t} = \lim_{\Delta t\to 0}\frac{1}{\Delta t}\Big(&\mathbb{P}\{X_{t+\Delta t}=(n,m)|X_t=(n,m+1)\}\mathbb{P}\{X_t=(n,m+1)\}\\ &+\mathbb{P}\{X_{t+\Delta t}=(n,m)|X_t=(n+1,m)\}\mathbb{P}\{X_t=(n+1,m)\}\\ &+\mathbb{P}\{X_{t+\Delta t}=(n,m)|X_t=(n+1,m-1)\}\mathbb{P}\{X_t=(n+1,m-1)\}\\ &+\mathbb{P}\{X_{t+\Delta t}=(n,m)|X_t=(n-1,m+1)\}\mathbb{P}\{X_t=(n-1,m+1)\}\\ &+\mathbb{P}\{X_{t+\Delta t}=(n,m)|X_t=(n-1,m)\}\mathbb{P}\{X_t=(n-1,m)\}\\ &+\mathbb{P}\{X_{t+\Delta t}=(n,m)|X_t=(n,m-1)\}\mathbb{P}\{X_t=(n,m-1)\}\\ &+\big(1-\mathbb{P}\{X_{t+\Delta t}\neq(n,m)|X_t=(n,m)\}\big)\mathbb{P}\{X_t=(n,m)\}\\ &-\mathbb{P}\{X_t=(n,m)\}\Big). \qquad (4)\end{aligned}$$

We now see that $\mathbb{P}\{X_{t+\Delta t}\neq(n,m)|X_t=(n,m)\}$ has limited possible states for $X_{t+\Delta t}\neq(n,m)$ precisely because $X_t=(n,m)$. Again, these possible states are those given by the transition rates in (1) excluding the transition W^{00}, as this results in $X_{t+\Delta t}=X_t=(n,m)$, which is not allowed. We then expand $\mathbb{P}\{X_{t+\Delta t}\neq(n,m)|X_t=(n,m)\}$ to all possible states, given as

$$\begin{aligned}&\mathbb{P}\{X_{t+\Delta t}\neq(n,m)|X_t=(n,m)\}=\\ &\quad\mathbb{P}\{X_{t+\Delta t}=(n+1,m)|X_t=(n,m)\}+\mathbb{P}\{X_{t+\Delta t}=(n,m+1)|X_t=(n,m)\}\\ &\quad+\mathbb{P}\{X_{t+\Delta t}=(n+1,m-1)|X_t=(n,m)\}+\mathbb{P}\{X_{t+\Delta t}=(n-1,m+1)|X_t=(n,m)\}\\ &\quad+\mathbb{P}\{X_{t+\Delta t}=(n-1,m)|X_t=(n,m)\}+\mathbb{P}\{X_{t+\Delta t}=(n,m-1)|X_t=(n,m)\}. \qquad (5)\end{aligned}$$

We can then subsitute (5) into (4) which results in

$$
\begin{aligned}
\frac{\partial P(n,m,t)}{\partial t} = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \Big(& \mathbb{P}\{X_{t+\Delta t} = (n,m) | X_t = (n,m+1)\} \mathbb{P}\{X_t = (n,m+1)\} \\
& + \mathbb{P}\{X_{t+\Delta t} = (n,m) | X_t = (n+1,m)\} \mathbb{P}\{X_t = (n+1,m)\} \\
& + \mathbb{P}\{X_{t+\Delta t} = (n,m) | X_t = (n+1,m-1)\} \mathbb{P}\{X_t = (n+1,m-1)\} \\
& + \mathbb{P}\{X_{t+\Delta t} = (n,m) | X_t = (n-1,m+1)\} \mathbb{P}\{X_t = (n-1,m+1)\} \\
& + \mathbb{P}\{X_{t+\Delta t} = (n,m) | X_t = (n-1,m)\} \mathbb{P}\{X_t = (n-1,m)\} \\
& + \mathbb{P}\{X_{t+\Delta t} = (n,m) | X_t = (n,m-1)\} \mathbb{P}\{X_t = (n,m-1)\} \\
& + \big(1 - \mathbb{P}\{X_{t+\Delta t} = (n+1,m) | X_t = (n,m)\} + \mathbb{P}\{X_{t+\Delta t} = (n,m+1) | X_t = (n,m)\} \\
& + \mathbb{P}\{X_{t+\Delta t} = (n+1,m-1) | X_t = (n,m)\} + \mathbb{P}\{X_{t+\Delta t} = (n-1,m+1) | X_t = (n,m)\} \\
& + \mathbb{P}\{X_{t+\Delta t} = (n-1,m) | X_t = (n,m)\} + \mathbb{P}\{X_{t+\Delta t} = (n,m-1) | X_t = (n,m)\}\big) \\
& * \mathbb{P}\{X_t = (n,m)\} - \mathbb{P}\{X_t = (n,m)\}\Big). \qquad (6)
\end{aligned}
$$

This equation is precisely composed of the sum of all the probabilities of moving to the state (n, m) minus the sum of all the probabilities of moving out of the state (n, m) during some infinitesimal time period $\Delta t \to 0$ divided by Δt.

Now we note that the final row of (6) cancels via the 1 in the expansion from (5). We then note that the transition probabilities, given above as $\mathbb{P}\{X_{t+\Delta t} = (n+j, m+k)|X_t = (n, m)\}$, become the transition rates in the limit $\Delta t \to 0$. We can then substitute all the transition probabilities with their corresponding transition rates in (6). Additionally, for notational simplicity, we will also substitute the $\mathbb{P}$ probability notation in (6) with the previously definted probability function P. Note that all the conditional probabilities in (6) have been replaced with the W transition rates, so there will be no confusion in notation by changing to $P(n, m, t)$. After some rearrangement to group together like terms based on the transition rates, this results in what we will henceforth term the master equation:

$$\begin{aligned}\frac{\partial P(n,m,t)}{\partial t} &= W^{-0}(n+1,m)P(n+1,m,t) - W^{-0}(n,m)P(n,m,t)\\ &+ W^{0-}(n,m+1)P(n,m+1,t) - W^{0-}(n,m)P(n,m,t)\\ &+ W^{+-}(n-1,m+1)P(n-1,m+1,t) - W^{+-}(n,m)P(n,m,t)\\ &+ W^{-+}(n+1,m-1)P(n+1,m-1,t) - W^{-+}(n,m)P(n,m,t)\\ &+ W^{+0}(n-1,m)P(n-1,m,t) - W^{+0}(n,m)P(n,m,t)\\ &+ W^{0+}(n,m-1)P(n,m-1,t) - W^{0+}(n,m)P(n,m,t). \qquad (7)\end{aligned}$$

2.2 Derivation of the Transition Rates

We shall now describe the transition rates W in terms of $n, m,$ and N in the same fashion as [2]. Note that because W^{00} is not included in the master equation (7), we do not define it below.

We begin with an example derivation of one of the two most complicated transition rates. W^{-+} is the probability density for an individual of type BB dying and being replaced by an individual of type AB in an infinitesimal time period $\Delta t \to 0$. Hence $W^{-+}(n,m)$ is equivalent to $\mathbb{P}\{X_{t+\Delta t} = (n-1, m+1) | X_t = (n,m)\}$ as $\Delta t \to 0$. Note that this W^{-+}, and indeed all the other transition rates W, is dependent only on the current state (n,m) and is independent of time t, making this process memoryless and hence Markovian.

We further characterize $W^{-+}(n,m)$ by its event: type BB dies and is replaced with type AB. Hence we have a rate μ at which individuals die, and therefore μn is the probability of a BB individual dying per unit time [2]. Since any death event is accompanied by a birth event, we then partition each death event by the birth of each of the three different genotypes. Hence we calculate the probability of an AB individual being born in the state (n,m) as being the sum of the probabilities that each different pairing of individuals results in an AB offspring times the probability of that pairing occurring. Hence using the notation where $AAxAB \to AB$ is the event that an AA-AB mating occurs and produces a AB offspring, we can calculate the individual probabilities of each mating occurring and producing an AB offspring as being

$$\begin{aligned}
\mathbb{P}\{AAxAB \to AB\} &= \frac{(N-m-n)m}{N^2} * \frac{1}{2}, \quad & \mathbb{P}\{ABxAA \to AB\} &= \frac{m(N-m-n)}{N^2} * \frac{1}{2} \\
\mathbb{P}\{AAxBB \to AB\} &= \frac{(N-m-n)n}{N^2}, \quad & \mathbb{P}\{BBxAA \to AB\} &= \frac{n(N-m-n)}{N^2} \\
\mathbb{P}\{BBxAB \to AB\} &= \frac{nm}{N^2} * \frac{1}{2}, \quad & \mathbb{P}\{ABxBB \to AB\} &= \frac{mn}{N^2} * \frac{1}{2} \\
\mathbb{P}\{ABxAB \to AB\} &= \frac{m^2}{N^2} * \frac{1}{2}
\end{aligned}$$

Note that the event $ABxAA \to AB$ is not the same as the event $AAxAB \to AB$ even though they have the same probability of occurring. Therefore special attention must be paid to the ordering of the pairings to obtain the proper total probabilities. We then sum the above terms and multiply the result by the original probability of a BB individual being removed to get $W^{-+}(n,m)$. Performing the same derivation for each possible genotype combination given by (1) gives the following transition rates:

$$\begin{aligned}
W^{0-}(n,m) &= \mu m\Big[\Big(\frac{N-n-m}{N}\Big)^2 + \frac{1}{4}\Big(\frac{m}{N}\Big)^2 + \frac{m(N-m-n)}{N^2}\Big], \\
W^{+-}(n,m) &= \mu m\Big[\Big(\frac{n}{N}\Big)^2 + \frac{1}{4}\Big(\frac{m}{N}\Big)^2 + \frac{mn}{N^2}\Big](1+s), \\
W^{-0}(n,m) &= \mu n\Big[\Big(\frac{N-n-m}{N}\Big)^2 + \frac{1}{4}\Big(\frac{m}{N}\Big)^2 + \frac{m(N-m-n)}{N^2}\Big], \\
W^{+0}(n,m) &= \mu(N-m-n)\Big[\Big(\frac{n}{N}\Big)^2 + \frac{1}{4}\Big(\frac{m}{N}\Big)^2 + \frac{mn}{N^2}\Big](1+s), \\
W^{-+}(n,m) &= \mu n\Big[\frac{1}{2}\Big(\frac{m}{N}\Big)^2 + \frac{2n(N-m-n)}{N^2} + \frac{nm}{N^2} + \frac{m(N-m-n)}{N^2}\Big], \\
W^{0+}(n,m) &= \mu(N-m-n)\Big[\frac{1}{2}\Big(\frac{m}{N}\Big)^2 + \frac{2n(N-m-n)}{N^2} + \frac{nm}{N^2} + \frac{m(N-m-n)}{N^2}\Big]
\end{aligned}$$

The exact derivations of the replacement terms are similar to those shown in [7] for the Hardy-Weinberg laws. Also note that the term $1 + s$ denotes the additional fitness gained by replicating individuals of genotype BB; this can intuitively be thought of as each transition that results in a BB individual being born occurring at faster rate than usual. Note that we will be henceforth measuring time in units N^2/μ, so without loss of generality we set $\mu/N^2 = 1$. This simplifies the above rates to:

$$\begin{aligned}
W^{0-}(n,m) &= m\Big[\,(N-n-m)^2 + \frac{m^2}{4} + m(N-m-n)\Big],\\
W^{+-}(n,m) &= m\Big[n^2 + \frac{m^2}{4} + mn\Big](1+s),\\
W^{-0}(n,m) &= n\Big[\,(N-n-m)^2 + \frac{m^2}{4} + m(N-m-n)\Big],\\
W^{+0}(n,m) &= (N-m-n)\Big[n^2 + \frac{m^2}{4} + mn\Big](1+s),\\
W^{-+}(n,m) &= n\Big[\frac{m^2}{2} + 2n(N-m-n) + nm + m(N-m-n\Big],\\
W^{0+}(n,m) &= (N-m-n)\Big[\frac{m^2}{2} + 2n(N-m-n) + nm + m(N-m-n)\Big]
\end{aligned} \tag{8}$$

2.3 The multidimensional generating function and its time derivative

We now wish to obtain an analytical solution to (7) so that we may have a full description of the probability distribution for the process at any time t. However, (7) is not easily solved due to the mixed values of the independent variables in the probability function, eg both $P(n+1, m, t)$ and $P(n, m, t)$ being in (7). Hence we resort to other methods that give analytical solutions for the steady state probability distribution, ie solutions for (7) equal to 0.

One strategy used in [2, 8] involves working with the probability generating function that governs our process. This probability generating function, hence termed the PGF, is given explicitly as

$$\phi(w, z, t) = \sum_n \sum_m w^n z^m P(n, m, t) \tag{9}$$

Also note that the above summation is recognizable as the expectation of $w^n z^m$, henceforth denoted $\mathbb{E}[w^n z^m]$.

PGFs are important tools in stochastic processes as they uniquely define the probability distribution that governs the process [8]. Intuitively, we transform (9) by $w = e^{t_1}$ and $z = e^{t_2}$, which gives the two dimensional moment generating function that also uniquely determines the distribution:

$$M_X(t_1, t_2, t) = \sum_n \sum_m e^{t_1 n + t_2 m} P(n, m, t)$$

The goal behind using the PGF is to transform (7) into a partial derivative with respect to time of the PGF, henceforth denoted as the dPGF. We know that there exists a steady-state PGF that corresponds to the steady-state solution of (7). We wish to find an analytical solution to the dPGF equation; this will correspond to the steady-state PGF that governs the steady-state solution to (7). This steady state PGF will then allow for the easy derivation of the fixation probabilities for any allele with selection coefficient s in a population of size N.

We begin to derive the dPGF by multiplying the master equation by $w^n z^m$ and summing over all $n, m \in \mathbb{Z}$. This transforms the left side of the master equation (7) into the dPGF:

$$\sum_n \sum_m w^n z^m \frac{\partial P(n, m, t)}{\partial t} = \frac{\partial}{\partial t} \sum_n \sum_m w^n z^m P(n, m, t) = \frac{\partial \phi(w, z, t)}{\partial t}.$$

Now consider as an example one of the particular terms in the right side of the master equation (7) after the same transformation: multiplication by $w^n z^m$ and summation over $n, m \in \mathbb{Z}^+$:

$$\sum_n \sum_m w^n z^m W^{-+}(n{+}1, m{-}1) P(n{+}1, m{-}1, t) = \sum_n \sum_m w^{n-1} z^{m+1} W^{-+}(n, m) P(n, m, t)$$

We can consider this equality to be given by a change of variables from n, m to $n-1$, $m+1$; note that this does not change the total summation because of the boundary conditions $P(n, m, t) = 0$ if $n, m < 0$ or $n + m > N$ and because the summation extends over all the integers. The above sum should also be recognizable as the expectation $\mathbb{E}[w^{n-1} z^{m+1} W^{-+}(n, m)]$

Performing these variable changes allows for the right side of the master equation (7) to be condensed, giving

$$\begin{aligned} \frac{\partial \phi(w, z, t)}{\partial t} &= \mathbb{E}[(w^n z^{m-1} - w^n z^m) W^{0-}(n, m)] \\ &+ \mathbb{E}[(w^{n-1} z^m - w^n z^m) W^{-0}(n, m)] \\ &+ \mathbb{E}[(w^{n+1} z^{m-1} - w^n z^m) W^{+-}(n, m)] \\ &+ \mathbb{E}[(w^{n-1} z^{m+1} - w^n z^m) W^{-+}(n, m)] \\ &+ \mathbb{E}[(w^{n+1} z^m - w^n z^m) W^{+0}(n, m)] \\ &+ \mathbb{E}[(w^n z^{m+1} - w^n z^m) W^{0+}(n, m)]. \end{aligned} \tag{10}$$

We will now substitute in the transition rates in (8) into the expectations in (10). First we take note of an important identity:

$$\mathbb{E}[n^p m^q w^n z^m] = (w\partial_w)^p (z\partial_z)^q \phi(w, z, t) \tag{11}$$

Here the term $(w\partial_w)^p$ denotes performing the following operation p times: take the derivative with respect to w and then multiply by w. Note that the multiplication and differentiation do not commute, hence particular care must be taken with explicitly stating the order of the operations when multiple iterations occur. This identity will in turn allow for (10) to be written in terms of partial derivatives of w and z instead of expectations. A similar derivation using a univariate version of (11) can be found in [2].

We begin the substitution with an example that will demonstrate using the identity (11) to transform the first line in (10) from an expectation to a series of sums of partial derivatives and multiples of w and z. We begin with considering the term

$$
\begin{aligned}
&\mathbb{E}[w^n z^{m-1} W^{0-}(n,m)] = \mathbb{E}[w^n z^{m-1} m(N - \frac{m}{2} - n)^2] = \\
&\quad = \mathbb{E}[w^n z^{m-1} m(N^2 - Nm - 2Nn + nm + \frac{m^2}{4} + n^2)] \\
&\quad = z^{-1}\mathbb{E}[N^2 m w^n z^m - Nm^2 w^n z^m - 2Nnm w^n z^m + nm^2 w^n z^m + \frac{m^3}{4} w^n z^m + n^2 m w^n z^m] \\
&\quad = z^{-1}[N^2 z\partial_z\phi - 2Nwz\partial_w\partial_z\phi + \frac{1}{4}(z\partial_z)^3\phi - N(z\partial_z)^2\phi + w\partial_w(z\partial_z)^2\phi + (w\partial_w)^2 z\partial_z\phi] \\
&\quad = \partial_z[N^2\phi - 2Nw\partial_w\phi + \frac{1}{4}(z\partial_z)^2\phi - Nz\partial_z\phi + wz\partial_w\partial_z\phi + (w\partial_w)^2\phi]
\end{aligned}
$$

We can then use the above derivation to get the full expansion of the first line of (10); this is best shown via rewriting the first line of (10) as

$$\mathbb{E}[(w^n z^{m-1} - w^n z^m)W^{0-}(n,m)] = \mathbb{E}[w^n z^{m-1} W^{0-}(n,m)](1-z)$$

This is because both z and w are deterministic variables and hence can be factored out of the expectation in the same fashion as constants.

Now we perform the same series of substitutions using (11) to get the full expansion for every line in (10). Line by line, they are as follows:

$$
\begin{aligned}
\mathbb{E}[(w^n z^{m-1} - w^n z^m)W^{0-}(n,m)] &= \partial_z[N^2\phi - 2Nw\partial_w\phi + \frac{1}{4}(z\partial_z)^2\phi - Nz\partial_z\phi \\
&\quad + wz\partial_w\partial_z\phi + (w\partial_w)^2\phi](1-z), \\
\mathbb{E}[(w^{n-1} z^{m} - w^n z^m)W^{-0}(n,m)] &= \partial_w[N^2\phi - 2Nw\partial_w\phi + \frac{1}{4}(z\partial_z)^2\phi \\
&\quad - Nz\partial_z\phi + wz\partial_w\partial_z\phi + (w\partial_w)^2\phi](1-w), \\
\mathbb{E}[(w^{n+1} z^{m-1} - w^n z^m)W^{+-}(n,m)] &= w\partial_z[\frac{1}{4}(z\partial_z)^2\phi + wz\partial_w\partial_z\phi + (w\partial_w)^2\phi](1-w^{-1}z)(1+s), \\
\mathbb{E}[(w^{n-1} z^{m+1} - w^n z^m)W^{-+}(n,m)] &= z\partial_w[\frac{N}{2}z\partial_z\phi - \frac{1}{4}(z\partial_z)^2\phi - wz\partial_w\partial_z\phi \\
&\quad + Nw\partial_w\phi - (w\partial_w)^2\phi](1-wz^{-1}), \\
\mathbb{E}[(w^{n+1} z^{m} - w^n z^m)W^{+0}(n,m)] &= w[\frac{1}{4}N(z\partial_z)^2\phi + Nwz\partial_w\partial_z\phi + N(w\partial_w)^2\phi \\
&\quad - (w\partial_w)^3\phi - 2(w\partial_w)^2\partial_z\phi - \frac{5}{4}w\partial_w(z\partial_z)^2\phi - \frac{1}{4}(z\partial_z)^3\phi](1-w^{-1})(1+s), \\
\mathbb{E}[(w^{n} z^{m+1} - w^n z^m)W^{0+}(n,m)] &= z[\frac{1}{4}(z\partial_z)^3\phi + \frac{5}{4}w\partial_w(z\partial_z)^2\phi + 2(w\partial_w)^2 z\partial_z\phi \\
&\quad + (w\partial_w)^3\phi - \frac{3}{4}N(z\partial_z)^2 - \frac{5}{2}wz\partial_w\partial_z\phi - 2N(w\partial_w)^2\phi \\
&\quad + \frac{1}{2}N^2 z\partial_z\phi + N^2 w\partial_w\phi](1-z^{-1})
\end{aligned} \tag{12}
$$

Substituting each term in (12) back into the expectation form of the dPGF (10) gives proper partial differential equation (PDE) form of the dPGF; it is this equation that can be manipulated to find the steady-state PGF. This equation, which does not provide any significant insights once expanded through substitution, is the explicit representation of our dPGF.

2.4 Techniques for solving the dPGF

First, we note that in order to obtain a steady-state PGF and hence steady-state probability distribution for our process, we would need to find a solution to the dPGF set equal to 0. This may seem to be even more intractable than solving (7) set equal to 0, but there are several techniques for solving partial differential equations that we can utilize to simplify the dPGF. There are two main approaches available to us.

2.4.1 Method of separation of variables

We begin by assuming that the solution to the dPGF is of some nice form that allows for the separation of the independent variables w and z. This is best illustrated through a quick, elementary example,

$$\partial_t^2 u(x,t) = c^2 \partial_x^2 u(x,t).$$

We first assume that $u(x,t)$ can be separated into functions of the independent variables x and t. Here we'll assume $u(x,t) = X(x)T(t)$. From this we can then simplify the above equation to

$$[\partial_t^2 T(t)]X(x) = c^2[\partial_x^2 X(x)]T(t)$$

which we can then rearrange to

$$\frac{\partial_t^2 T(t)}{T(t)} = c^2 \frac{\partial_x^2 X(x)}{X(x)}$$

From this we note that for any x_0 contained within the boundary of the solution where $X(x_0) \neq 0$, $\partial_x^2 X(x_0)/X(x_0)$ will be equal to a constant, which we will term λ. Similarly for every t_0 contained within the boundary where $T(t_0) \neq 0$, $\partial_t^2 T(t_0)/T(t_0)$ will also be equal to a constant, and hence we get

$$\frac{\partial_t^2 T(t)}{T(t)} = c^2 \frac{\partial_x^2 X(x)}{X(x)} = \lambda$$

which is a simple series of second order ODEs that have easily obtainable solutions that will also apply to the original PDE for any constant λ. It is this separated series of ODEs that we wish to transform our dPGF into.

For our process, we have two intuitive options. First we examine the solution

$\phi(w,z) = h(w)g(z)$. Note that because we are solving for the steady state solution, our solutions do not require time components. Unfortunately, this formulation does not produce separation as we have terms that are functions of both ∂_z and ∂_w; these will prevent the separation of the dPGF into the ideal functions $H(w) = cG(z)$.

Hence based off of our previous issue of being unable to separate mixed partial derivatives, we could instead make the assumption that the solution is instead an additive composition of functions of the independent variables, ie $\phi(w,z) = h(w)+g(z)$. This has the benefit of removing all the terms in the dPGF that are combinations of both ∂_z and ∂_w which makes the terms fully separable in terms of the partial derivatives. However, even after this simplifying assumption is made, certain terms in the dPGF still have mixed w and z coefficients multiplying their partial derivatives of ϕ. This prevents full separation of the independent variables, hence we cannot find a function of purely w and $\partial_w^q \phi$ for different powers q that is equal to a similar function for z. This means we cannot find an analytical solution to our dPGF with separation of variables using simple and intuitive assumptions about the function decomposition of the solutions.

Additionally, symbolic computation via Mathematica has failed to produce an analytical solution both in the general case where we solve for the solution of the dPGF set equal to 0 and for both of the cases designed to produce separation of variables, $\phi(w,z) = h(w) + g(z)$ and $\phi(w,z) = h(w)g(z)$. This leaves the method of direct numerical computation given specified values for the parameters N and s.

2.4.2 Numerical solutions to the dPGF

3 Conclusion

This particular approach to modeling population genetics has several distinct advantages. First, as stated by [2], this method allows for an analytic solution without “ad hoc” solutions that rely on arbitrary boundary conditions, as outlined in [1]. Second, this method is highly flexible; it can be extended to include mutation, alternative reproductive modes, and even populations with spatial structure [9]. These are realized via appropriate modifications to the transitions rates and the state variables. In particular, we aim to extend this project to include the mutation case for a sexual, diploid population. Then we will have created a complete population model for mutation, selection, and genetic drift.

This model is significant in that it is without approximation; this is not the case for the more commonly used diffusion-based model created by Kimura [1]. Kimura's diffusion equation assumes that terms smaller than $1/N$ are negligible, and hence it does not provide adequate resolution for small populations. This adversely affects the study of genetic drift, defined to be the changes that occur in a population due to random sampling. For instance, if an individual has a 10% improved chance of reproducing due to a beneficial allele, then in a population of size $N = 4$ the allele has only a 0.29 probability of becoming fixed [9]. This is a near-negligible difference in the fixation probability for a highly significant fitness gain, thus demonstrating the vast impact that genetic drift can have on a small population. Thus it is significant that the diffusion equations do not allow for satisfactory study of this essential evolutionary phenomenon. While our model not only successfully incorporates genetic drift without approximation, it also includes natural selection and can be extended to other evolutionary forces.

Additionally, it is not immediately obvious that a "nice," analytic solution exists for a population with an arbitrary number of distinct alleles, distinct genetic loci, spatial distribtuion, and/or reproductive regime. However, if the transition rates can be completely described, i.e. can list a finite series of transition rates, each with finite terms, then it is possible to numerically solve for the fixation probabilities for any given selection and mutation coefficients for any initial condition. This may in turn prove to be computationally unviable for increasingly large and complex populations, but it is a very distinct advantage for our method and allows for the consideration of complicated models that would otherwise be considered intractable.

Thus this project demonstrates the versatility of stochastic modeling in addressing a pivotal problem in population genetics in a precise, analytical fashion [2] as well as potential numerical solutions for more complex scenarios ranging from spatial population organization to the bizarre and unique genetics of ciliates.

References

[1] James F. Crow and Motoo Kimura. *An Introduction to Population Genetics Theory*. Harper and Row, 1970.

[2] Bahram Houchmandzadeh and Marcel Vallade. Alternative to the diffusion equation in population genetics. *Physical Review E*, 82:051913, 2010.

[3] David M. Prescott. The dna of ciliated protozoa. *Microbiological Reviews*, 58:233–267, 1994.

[4] Warren J. Ewens. *Mathematical Population Genetics*. Springer-Verlag, 1979.

[5] Thomas Nagylaki. *Introduction to Theoretical Population Genetics*. Springer-Verlag, 1992.

[6] Crispin W. Gardiner. *Handbook of Stochastic Methods*. Springer-Verlag, 1983.

[7] Peter J. Russell. *iGenetics*. Pearson, New York, 2006.

[8] Norman T.J. Bailey. *The Elements of Stochastic Processes*. John Wiley and Sons, 1964.

[9] Bahram Houchmandzadeh and Marcel Vallade. The fixation probability of a beneficial mutation in a geographically structured population. *New Journal of Physics*, 13, 2011.