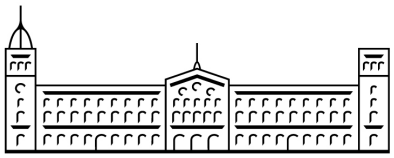

Treball final de màster

MÀSTER DE
MATEMÀTICA AVANÇADA

Facultat de Matemàtiques
Universitat de Barcelona

Biological interactions between multiple genetic mutations

Autor: Adrià Màdico Ferrer

Directors:	Dr. Luca Ferretti
	Dr. Sebastián Ramos-Onsins
Tutor:	Dr. Carles Simó
Realitzat a:	Departament de
	Matemàtica Aplicada i Anàlisi,
	Centre de Recerca en Agrigenomica

Barcelona, June 30, 2014

Contents

Acknowledgements

First of all, I want to remember my family to support me in every moment that I need it.

I would like to thank the Master's mates with whom we have done really hard work and exciting parties. I also want to remember my CRAG officemates for these two amazing months discussing about Biology, Maths, live and football.

Finally, I want to acknowledge personally Luca Ferretti for showing me that surprising ways of solving problems, Sebastián Ramos for introducing me to that world of Biology and Carles Simó for encouraging all the students to improve our ambitions and results.

Abstract

The aim of this project is to provide mathematical results for the effect of selection and environmental changes on the distribution of mutations in DNA sequences.

This project is divided in two parts. The first part is devoted to the stationary solutions of the Kolmogorov equations for the distribution of mutation frequencies in a population. We generalize the well-known diffusion equations for a single mutation to pairs of mutations, and we give an explicit stationary solution in the neutral case.

In the second part, the evolution of the frequencies of pairs of mutations is studied in the case of a sudden environmental change. In particular, we assume that the mutations are evolving neutrally (therefore their distribution follows the expression found in the first part) until a change occurs and one of the mutations becomes selected. We derive the final distribution of frequencies for this scenario.

Goals and motivation

Evolution in living organisms proceeds through natural selection. Selection acts on mutations in the genetic material of individuals, increasing the frequency of some of these mutations in the population after each generation. Modern DNA sequencing methods allow to obtain the DNA sequences of many individuals collected from a population at a given time. Using these data, selection can be inferred by the distribution of mutations in the population. However, detecting selection is not easy. Even for neutral (i.e., non-selected) mutations, random changes in the frequencies of the mutations result in non-trivial patterns in their distribution.

In recent years, it has become clear that natural selection is not constant, but depends on the external environment, therefore selection changes in time following environmental fluctuations. Environmental changes are widespread at different scales and some of these changes are caused or amplified by the impact of human activities. It is therefore of great interest to understand the pattern of time-varying selection on mutation frequencies.

The goal of this project is to provide mathematical expressions for some quantities related to the distribution of mutations in a population, both without selection and with selection changing in time. These results will be useful to detect episodes of time-dependent selection from DNA sequence data.

The objectives of this project are twofold.

One of the simplest and most used statistics of mutation patterns is the frequency spectrum, i.e. the count or the distribution of the frequency of different mutations in a population [2, 1].

The first objective is to derive simple expressions for the frequency spectrum of pairs of mutations, which corresponds to the stationary solution of a set of diffusion equations. Existing approaches in terms of polynomial expansions [5, 4] led to complicated expressions for the solution [6], while recent results from coalescent theory [3] suggest that there is a simple solution, at least in the neutral case. We write the diffusion equations for pairs of mutations and prove that the solution is a stationary solution of these equations.

The second objective is to characterize the frequency spectrum after a recent environmental change. The change triggers a selection pressure on a mutation in a sequence that was evolving neutrally. In the limit of strong selection, we obtain an exact formula for the frequency spectrum. For intermediate selection, we obtain the frequency spectrum as an expansion in powers of the inverse selection coefficient.

Chapter 1

Biological introduction

This is a project of applied mathematics that models a biological problem, then an introduction to the biological background is needed. In this chapter I explain the bases of the project.

1.1 Basic concepts

Definitions 1.1.1. *Here I define some relevant topics:*

- *A **chromosome** is a sequence of DNA bases, which contains the genetic information of an individual.*
- ***Haploid** is the term used for individuals that only have one copy of each chromosome. In that project we consider that our population are N_e haploid individuals, despite that it is equivalent to work with diploid individuals (i.e., with two copies) and a population of $N_e/2$ individuals.*
- *A **gene** is a small part of a chromosome with usually an specific function.*
- *A **locus** is a concrete position in a chromosome, the place where a gene is located.*
- *An **allele** is one of the several forms of a gene.*

  As an example, in the Figure 1.1 we can see genes (a, b, c, d) and a number of individuals (five, in vertical) with different number of alleles, for example, in the first generation the gene a has 3 alleles and the others have 2.
- *A **mutation** is a modification in a gene of a single individual that occurs randomly in the nature. If a mutation occurs, then it is created a new allele, the derived allele. In some cases, when the model has more than one mutation, we will denote as **focus mutation** or **ancestral** to the original one.*

  In this project, we will focus on biallelic mutations, i.e. only 2 alleles are possible at each locus.

- *The **offspring** of an individual is the set of the next generation individuals that comes from it.*

- *The **fitness advantage**, s_i, of a particular allele i is a term that expresses the potential advantage of an allele in the immediate future generation. It is also called selection coefficient. A usual notation in biology for the absolute fitness (which corresponds roughly to the average relative number of offspring) is $1 + s_i \geq 0$, considering that $s_i = 0$ corresponds to neutral selection, $s_i < 0$ a negative selection and $s_i > 0$ positive selection. If an allele has fitness advantage -1, then it does not have any possibility of being expressed in the future generations.*

- *We are going to say that a mutation is been **fixed** if it is finally present in all the population and **lost** if it does not appear in any individual.*

  In the Figure 1.1 if we consider that $\times$ are the mutated alleles and $\bigcirc$ the ancestral ones, we will say that in the second generation the mutation of c is lost and the one of d is fixed.

- ***Recombination** is a process where the genetic information of two individuals is mixed, i.e. some parts of the genetic information of the new individual will come from one of the ancestor and other part from another, creating a new possible combination of alleles in the new generation. Our models are without recombination, so all the information pass from one individual directly to the the new one.*

- *We say that a mutation is **nested** in another if it was born in a sequence containing the first one. Equivalently we are going to say that a mutation A is **containing** B if B is nested in A. If the two mutations have the same frequency in the population, we call them **co-occurring**. Otherwise, we say that is **exclusive** if it was born in a sequence with the original allele. If the sum of the frequencies reaches 1, we call the mutations **complementary**.*

  As an example in the same Figure 1.1 we have all the relevant cases; in the first generation and considering that $\times$ are the mutated alleles and $\bigcirc$ the ancestral ones, the mutation of b is nested in the one of d, or d is containing b, and the mutation of c is exclusive with respect to the mutation in b and both mutations of a.

1.2 Wright-Fisher model and diffusion approximation

We are going to work with the following assumptions:

- Large haploid population N_e constant in time.

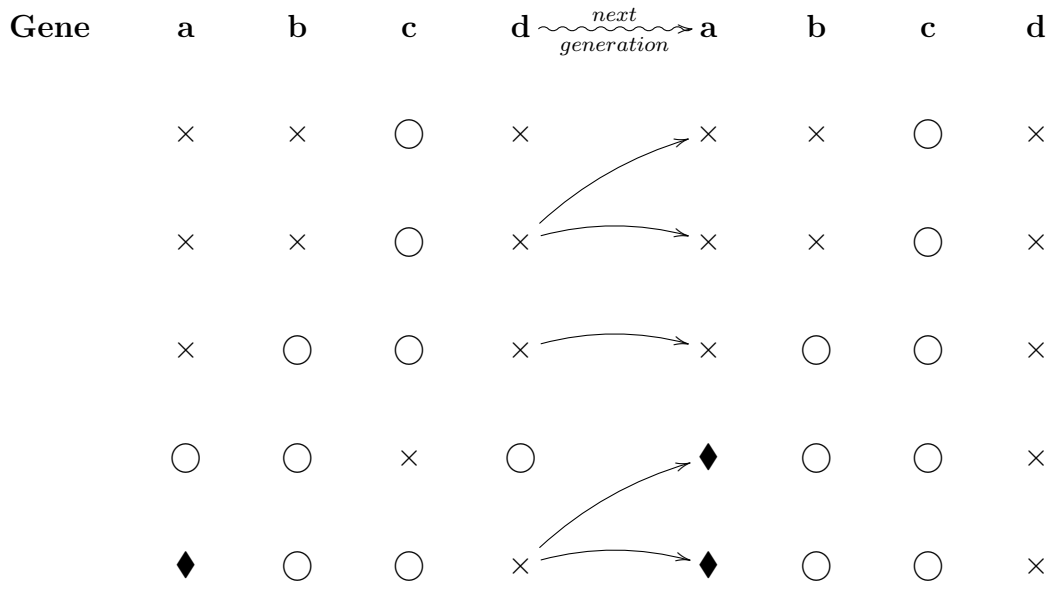

Figure 1.1: Evolution of some mutations. This diagram contents 2 different generations of a constant population of five individuals (columns) with 4 genes (a, b, c and d). It is widely uses in the initial definitions in order to give several examples.

- $K-1$ different mutations that create K different alleles combinations. We will focus on the case $K=3$, i.e. mutations at two loci.
- Generations does not overlap one to each other.
- They don't have recombination, so each copy of the gene found in the new generation is drawn independently at random from all copies of the gene in the old generation.
- All alleles have the same fitness, i.e. evolution is neutral. (In the selected case considered later, the new generation will be drawn with relative weight $1+s_i$ on the probability of choosing the ith copy of the gene as parent.)

The Wright-Fisher Model is based on these assumptions. It is the Markov model that we are going to use to study the evolution of the frequencies and probabilities of fixation of our alleles. We are actually going to consider a more sophisticated problem, with non-neutral fitness in general (i.e. $s_i \neq 0$).

Notation 1.2.1. From now on, we denote n_i as the number of individuals with the ith combination of alleles, $x_i := n_i/N_e$ their frequency and s_i their fitness. Notice that $\sum_{i=0}^{K} n_i = N_e$, and consequently, $\sum_{i=0}^{K} x_i = 1$.

Under the conditions of evolution considered before, the distribution for $N = (n_1, \ldots, n_K)$ conditioned on the previous frequencies $x_i(t)$ is clearly a multinomial distribution with probabilities $p = (p_1 \ldots p_n)$:

$$p_i = \frac{x_i(t)(1+s_i)}{\sum_{j=1}^{K} x_j(t)(1+s_j)}. \tag{1.2.1}$$

We will use the following Lemma to prove the case $K = 2$ of the theorem 1.2.3.

Lemma 1.2.2. *Let $(x_1(t+\Delta t), x_2(t+\Delta t)) = (n_1, n_2)/N_e$ two random variables. $x_1(t+\Delta t)$ follows a binomial distribution divided by a constant $N_e = n_1 + n_2$ with probability p_1 as defined above in formula 1.2.1 with $K = 2$.*

We define $\Delta x_1 = x_1(t+\Delta t) - x_1(t)$. Then, if s_i and $1/N_e$ are sufficiently small the following expressions are good approximations at first order in s_i and $1/N_e$:

$$\begin{aligned} \mathrm{E}[\Delta x_1|\mathcal{F}_t] &= x_1(s_1 - s_2)(1 - x_1), \\ \mathrm{E}[(\Delta x_1)^2|\mathcal{F}_t] &= \frac{x_1(1-x_1)}{N_e}, \\ \mathrm{E}[(\Delta x_1)^j|\mathcal{F}_t] &= o(1/N_e) \qquad\qquad \text{for } j \geq 3; \end{aligned}$$

where $\mathcal{F}_t = \sigma(x_1(t') : t' \leq t)$ is the sigma algebra generated by the previous values of x_1.

Proof. In order to simplify the notation, we do specify the dependence on t of the variable x_1.

Before starting notice that $x_1(t) + x_2(t) = 1$ and the moment-generating function for a binomial distribution is $(1 - p + pe^t)^n$.

$$\begin{aligned} \mathrm{E}[\Delta x_1|\mathcal{F}_t] = \mathrm{E}[x_1(t+\Delta t)|\mathcal{F}_t] - \mathrm{E}[x_1(t)|\mathcal{F}_t] &= \frac{N_e}{N_e}\frac{x_1(1+s_1)}{x_1(1+s_1)+x_2(1+s_2)} - x_1 \\ &= \frac{x_1(1+s_1)}{x_1+x_2+x_1s_1+x_2s_2} - x_1 = \frac{x_1(1+s_1)}{1+s_2+x_1(s_1-s_2)} - x_1 \\ &= \frac{x_1(s_1-(s_2+x_1(s_1-s_2)))}{1+s_2+x_1(s_1-s_2)} = \frac{x_1(1-x_1)(s_1-s_2)}{1+O(s_1)+O(s_2)}. \end{aligned}$$

$$\begin{aligned} \mathrm{E}[(\Delta x_1)^2|\mathcal{F}_t] &= \frac{N_e}{N_e^2}\, p_1(1-p_1) \\ &= \frac{1}{N_e}\frac{x_1(1+s_1)}{x_1(1+s_1)+x_2(1+s_2)}\left(1 - \frac{x_1(1+s_1)}{x_1(1+s_1)+x_2(1+s_2)}\right) \\ &= \frac{1}{N_e}\frac{x_1(1+s_1)x_2(1+s_2)}{(x_1(1+s_1)+x_2(1+s_2))^2} = \frac{1}{N_e}\frac{x_1(1-x_1)+O(s_1)+O(s_2)}{1+O(s_1)+O(s_2)}. \end{aligned}$$

If we consider the values of selection, s_i, and $1/N_e$ sufficiently small:

$$\mathrm{E}[\Delta x_1|\mathcal{F}_t] = x_1(s_1 - s_2)(1 - x_1),$$
$$\mathrm{E}[(\Delta x_1)^2|\mathcal{F}_t] = \frac{x_1(1 - x_1)}{N_e}.$$

Finally, in the limit of small $1/N_e$, the estimate $\mathrm{E}[(\Delta x_1)^j|\mathcal{F}_t] = o(1/N_e)$ for $j \geq 3$ is a known result from the theory of Central Limit Theorems.

□

The object that we want to study is the *allele frequency spectrum*, i.e. the mean density of mutations of frequency f in the population, with $0 < f < 1$. It is denoted by $\xi(f)$. This quantity is closely related to the Green function $G(x, y, t)$ for the evolution of the frequency of a mutation, i.e. the probability density of finding the mutation in x after a time t given the initial frequency y. In fact, we have

$$\xi(f) = \frac{\theta}{2} \int_0^{+\infty} G(f, 1/N_e, t)dt,$$

where $\theta/2$ is the rate of new mutations in the population per unit time.

Theorem 1.2.3. *The evolution equations that model our problem defined above are the forward and backward Kolmogorov equations, defined respectively as:*

$$\frac{\partial G((x_1, \ldots, x_K), (y_1, \ldots, y_K), t)}{\partial t} = \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{1}{2} \frac{\partial^2}{\partial x_i \partial x_j} \left[b(x_i, x_j, t) G \right] - \sum_{i=1}^{K} \frac{\partial}{\partial x_i} \left[a(x_i, t) G \right],$$
$$\frac{\partial G((x_1, \ldots, x_K), (y_1, \ldots, y_K), t)}{\partial t} = \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{1}{2} b(y_i, y_j, 0) \frac{\partial^2 G}{\partial y_i \partial y_j} + \sum_{i=1}^{K} a(y_i, 0) \frac{\partial G}{\partial y_i};$$

where we assume that $b(x_i, x_j, t) = Cov(\Delta x_i, \Delta x_j|\mathcal{F}(t))$ *and* $a(x_i, t) = \mathrm{E}[\Delta x_i|\mathcal{F}_t]$ *and* $\mathcal{F}_t = \sigma(x_1(t') : t' \leq t)$.

Proof. We are going to prove that Theorem for $K = 2$. The arguments for general K are similar but involve more calculus.

Consider x the frequency of the derived allele and $y = 1 - x$ the frequency of the original allele. Define the function $G(x(t_1), x(t_0), t_1 - t_0)$ as the probability distribution for the derived allele to pass from frequency $x(t_0)$ to $x(t_1)$ in $t_1 - t_0$ time if $t_0 < t_1$. Then consider an intermediate time m, $t_0 < m < t_1$ and we have the following equality:

$$G(x(t_1), x(t_0), t_1 - t_0) = \int_0^1 G(x(m), x(t_0), t - t_0) G(x(t_1), x(m), t_1 - t) dx(m),$$

which is called the Chapman-Kolmogorov equation.

Forward Kolmogorov Equation

In other to find the Forward Equation, we consider $m = t$ and $t_1 = t + \Delta t$.

In the second equality, we change $x(t)$ for $z-\varepsilon$, $x(t+\Delta t)$ for $z-\varepsilon+\varepsilon'$ and $x(t+\Delta t)$ for y. Notice that $\varepsilon=\varepsilon'$, but we differentiate it in order to clarify the Taylor expansion where the variable $z-\varepsilon$ is z.

$$\begin{aligned}
&G(x(t+\Delta t),x(t_0),t+\Delta t-t_0)\\
=&\int_0^1 G(x(t),x(t_0),t-t_0)G(x(t+\Delta t),x(t),\Delta t)dx(t)\\
=&\int_0^1 \left[G(z-\varepsilon,y,t-t_0)G(z-\varepsilon+\varepsilon',z-\varepsilon,\Delta t)\right]_{\varepsilon'=\varepsilon} d(z-\varepsilon)\\
=&\int_{z-1}^{z}\sum_{j\geq 0}\left(\frac{(z-\varepsilon-z)^j}{j!}\left[\frac{\partial^j}{\partial z^j}[G(z,y,t-t_0)G(z+\varepsilon',z,\Delta t)]\right]_{\varepsilon'=\varepsilon}\right)d\varepsilon\\
=&\int_{z-1}^{z}\sum_{j\geq 0}\frac{(-1)^j}{j!}\frac{\partial^j}{\partial z^j}[G(z,y,t-t_0)G(z+\varepsilon,z,\Delta t)\varepsilon^j]d\varepsilon\\
=&G(z,y,t-t_0)\int_{1-z}^{-z}G(z+\varepsilon,z,\Delta t)d\varepsilon\\
&+\sum_{j\geq 1}\frac{(-1)^j}{j!}\frac{\partial^j}{\partial z^j}[G(z,y,t-t_0)\int_{z-1}^{z}\varepsilon^j G(z+\varepsilon,z,\Delta t)d\varepsilon].
\end{aligned}$$

Now, we subtract in both sides $G(z,y,t-t_0)\int_{z-1}^{z}G(z+\varepsilon,z,\Delta t)d\varepsilon$ and we divide it by Δt taking the limit when Δt is going to 0. Hence, using that $\int_{1-z}^{-z}G(z+\varepsilon,z,\Delta t)d\varepsilon$ is one, so in the left hand side we have:

$$\lim_{\Delta t\to 0}\frac{G(z,y,t+\Delta t-t_0)-G(z,y,t-t_0)}{\Delta t}=\frac{\partial G}{\partial t}.$$

For the right hand side:

$$\mathrm{E}[(x(t+\Delta t)-x(t))^j|\mathcal{F}_t]=\mathrm{E}[(\Delta x)^j|\mathcal{F}_t]=\lim_{\Delta t\to 0}\frac{1}{\Delta t}\int_{z-1}^{z}\varepsilon^j G(z+\varepsilon,z,\Delta t)d\varepsilon,$$

where $\mathcal{F}_t=\sigma(x(t'):t'\leq t)$. Hence, using the Lemma 1.2.2 we know that we can truncate the equation in the second two. Despite that it is noticeable that the equation with all the orders is called Kramers-Moyal.

$$\sum_{j\geq 0}\frac{(-1)^j}{j!}\frac{\partial^j}{\partial z^j}[G(z,y,t-t_0)\,\mathrm{E}[(\Delta x)^j|\mathcal{F}_t]=-\frac{\partial^j}{\partial x^j}\,\mathrm{E}[\Delta x|\mathcal{F}_t]G+\frac{1}{2}\frac{\partial^j}{\partial x^j}\,\mathrm{Var}(\Delta x|\mathcal{F}(t))G.$$

Hence,

$$\frac{\partial G}{\partial t}=-\frac{\partial}{\partial x}\,\mathrm{E}[\Delta x|\mathcal{F}_t]G+\frac{1}{2}\frac{\partial^2}{\partial x^2}\,\mathrm{Var}(\Delta x|\mathcal{F}(t))G. \tag{1.2.2}$$

Considering the equation that we want to get and applying that change of variables $x=x_1$, $r=x_1+x_2$ to them, we found the immediate above equation 1.2.2.

Recall that $b(x,y) = \mathrm{Cov}(\Delta x, \Delta y|\mathcal{F}(t))$.

$$\begin{aligned}
\frac{\partial G}{\partial t} =& \sum_{i=1}^{K}\sum_{j=1}^{K}\frac{1}{2}\frac{\partial^2}{\partial x_i \partial x_j}\left[b(x_i,x_j)G\right] - \sum_{i=1}^{K}\frac{\partial}{\partial x_i}\left[\mathrm{E}[\Delta x_i|\mathcal{F}_t]G\right] \\
=& \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[b(x,x)G\right] + \frac{1}{2}\frac{\partial^2}{\partial x \partial r}\left[2b(x,x)G\right] + \frac{1}{2}\frac{\partial^2}{\partial r^2}\left[b(x,x)G\right] + \\
&+ 2\left(\frac{1}{2}\frac{\partial^2}{\partial x \partial r}\left[b(x,1-x)G\right] + \frac{1}{2}\frac{\partial^2}{\partial r^2}\left[b(x,1-x)G\right]\right) \\
&+ \frac{1}{2}\frac{\partial^2}{\partial r^2}\left[b(1-x,1-x)G\right] - \frac{\partial}{\partial x}\left[\mathrm{E}[\Delta x|\mathcal{F}_t]G\right] - \frac{\partial}{\partial r}\left[\mathrm{E}[\Delta x|\mathcal{F}_t]G\right] \\
&- \frac{\partial}{\partial r}\left[\mathrm{E}[1-\Delta x|\mathcal{F}_t]G\right] \\
=& \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[\mathrm{Var}(\Delta x|\mathcal{F}(t))G\right] - \frac{\partial}{\partial x}\left[\mathrm{E}[\Delta x|\mathcal{F}_t]G\right],
\end{aligned}$$

using that

$$\begin{aligned}
&\mathrm{E}[\Delta x|\mathcal{F}_t] = -\mathrm{E}[1-\Delta x|\mathcal{F}_t], \\
&\mathrm{Var}(\Delta x|\mathcal{F}(t)) = b(1-x,1-x) = -b(x,1-x).
\end{aligned}$$

Backward Kolmogorov equation

In that case, we define $m = t_0 + \Delta t$ and $t_1 = t + \Delta t$.

As in the Forward proof, in the second inequality we change $x(t+\Delta t)$ for z, $x(t_0)$ for y and $x(t_0+\Delta t)$ for $y+\varepsilon$. Then, we perform a Taylor expansion in the variable $y+\varepsilon$ around y.

$$\begin{aligned}
&G(x(t+\Delta t), x(t_0), t+\Delta t - t_0) \\
=& \int_0^1 G(x(t_0+\Delta t), x(t_0), \Delta t) G(x(t+\Delta t), x(t_0+\Delta t), t-t_0) dx(t_0+\Delta t) \\
=& \int_0^1 G(y+\varepsilon, y, \Delta t) G(z, y+\varepsilon, t-t_0) d(y+\varepsilon) \\
=& \int_{-y}^{1-y} \sum_{j\geq 0} \frac{\varepsilon^j}{j!} G(y+\varepsilon, y, \Delta t) \frac{\partial^j}{\partial y^j}[G(z,y,t-t_0)] d\varepsilon \\
=& G(z,y,t-t_0)\int_{-y}^{1-y} G(y+\varepsilon, y, \Delta t) d\varepsilon \\
&+ \sum_{j\geq 1}\frac{1}{j!}\left[\int_{-y}^{1-y} G(y+\varepsilon, y, \Delta t)\varepsilon^j d\varepsilon\right]\frac{\partial^j}{\partial y^j}[G(z,y,t-t_0)].
\end{aligned}$$

For the right hand side, using the same than in the forward equation and we have:

$$\sum_{j\geq 1}\frac{1}{j!}\frac{\partial^j G}{\partial y^j}\mathrm{E}[(\Delta x)^j|\mathcal{F}_t] = \mathrm{E}[\Delta x|\mathcal{F}_t]\frac{\partial G}{\partial y} + \mathrm{Var}(\Delta x|\mathcal{F}(t))\frac{1}{2}\frac{\partial^2 G}{\partial y^2}.$$

Hence,

$$\frac{\partial G}{\partial t} = \mathrm{E}[\Delta x|\mathcal{F}_t]\frac{\partial G}{\partial y} + \mathrm{Var}(\Delta x|\mathcal{F}(t))\frac{1}{2}\frac{\partial^2 G}{\partial y^2}. \tag{1.2.3}$$

Now, doing the change of variables $x = x_1$, $s = x_1 + x_2$ as in the forward equation we found the equation 1.2.3.

□

Corollary 1.2.4. *As we have seen in the proof of the Theorem1.2.3, for the case of one mutation, $K = 2$, we can use the following two differential equations*

$$\begin{aligned}\frac{\partial G}{\partial t} &= -\frac{\partial}{\partial x}[\mathrm{E}[\Delta x|\mathcal{F}_t]G] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[\mathrm{Var}(\Delta x|\mathcal{F}(t))G],\\ \frac{\partial G}{\partial t} &= \mathrm{E}[\Delta x|\mathcal{F}_t]\frac{\partial G}{\partial y} + \mathrm{Var}(\Delta x|\mathcal{F}(t))\frac{1}{2}\frac{\partial^2 G}{\partial y^2}\end{aligned}$$

or the original ones from the Theorem because they are the equivalent.

Proposition 1.2.5. *Let $X = (x_1, \ldots, x_n) = (n_1 \ldots, n_K)/N_e$ random variables that follow a multinomial distribution divided by a constant $N_e = \sum_{j=1}^{K} n_j$ with probabilities $p = (p_1 \ldots p_n)$, where p_i is defined as in 1.2.1. Then, if s_i and $1/N_e$ are sufficiently small:*

$$\begin{aligned}\mathrm{E}[\Delta x_i|\mathcal{F}_t] &= x_i\sum_{j=1}^{K}(s_i - s_j)x_j,\\ Var(\Delta x_i|\mathcal{F}(t)) &= \frac{x_i(1-x_i)}{N_e},\\ Cov(\Delta x_i, \Delta x_j|\mathcal{F}(t)) &= \frac{-x_i x_j}{N_e}, \quad i \neq j.\end{aligned}$$

Proof. In order to simplify the notation, we do specify the dependence on t of the variable x_i.

Before starting notice that $\sum_{j=1}^{K} x_j(t) = 1$.

$$\begin{aligned}\mathrm{E}[\Delta x_i|\mathcal{F}_t] = \mathrm{E}[x_i(t+\Delta t)|\mathcal{F}_t] - \mathrm{E}[x_i(t)|\mathcal{F}_t] &= \frac{N_e p_i}{N_e} - x_i = \frac{x_i(1+s_i)}{\sum_{j=1}^{K} x_j(1+s_j)} - x_i\\ &= \frac{x_i(1+s_i)}{\sum_{j=1}^{K} x_j + \sum_{j=1}^{K} s_j x_j} - x_i = \frac{x_i s_i - x_i(\sum_{j=1}^{K} s_j x_j)}{1 + \sum_{j=1}^{K} s_j x_j}\\ &= x_i\frac{\sum_{j=1}^{K} x_j s_i - \sum_{j=1}^{K} s_j x_j}{1 + \sum_{j=1}^{K} s_j x_j} = x_i\frac{\sum_{j=1}^{K}(s_i - s_j)x_j}{1 + \sum_{j=1}^{K} s_j x_j}.\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(\Delta x_i|\mathcal{F}(t)) &= \frac{N_e p_i(1-p_i)}{N_e^2} = \frac{1}{N_e}\frac{x_i(1+s_i)}{\sum_{j=1}^K x_j(1+s_j)}\left(1-\frac{x_i(1+s_i)}{\sum_{j=1}^K x_j(1+s_j)}\right)\\ &= \frac{1}{N_e}\frac{x_i(1+s_i)}{1+\sum_{j=1}^K s_j x_j}\left(\frac{1+\sum_{j=1}^K s_j x_j - x_i(1+s_i)}{1+\sum_{j=1}^K s_j x_j}\right)\\ &= \frac{1}{N_e}\frac{x_i(1+s_i)+x_i(1+s_i)\sum_{j=1}^K s_j x_j - (x_i(1+s_i))^2}{(1+\sum_{j=1}^K s_j x_j)^2}\\ &= \frac{1}{N_e}\frac{x_i - x_i^2 + \sum_{j=1}^K O(s_j)}{1+\sum_{j=1}^K O(s_j)}.\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}(\Delta x_i,\Delta x_j|\mathcal{F}(t)) &= \frac{N_e p_i p_j}{N_e^2} = \frac{1}{N_e}\frac{x_i(1+s_i)}{\sum_{j=1}^K x_j(1+s_j)}\left(1-\frac{x_i(1+s_i)}{\sum_{j=1}^K x_j(1+s_j)}\right)\\ &= \frac{1}{N_e}\frac{x_i(1+s_i)}{1+\sum_{j=1}^K s_j x_j}\left(\frac{1+\sum_{j=1}^K s_j x_j - x_i(1+s_i)}{1+\sum_{j=1}^K s_j x_j}\right)\\ &= \frac{1}{N_e}\frac{x_i(1+s_i)+x_i(1+s_i)\sum_{j=1}^K s_j x_j - (x_i(1+s_i))^2}{(1+\sum_{j=1}^K s_j x_j)^2}\\ &= \frac{1}{N_e}\frac{x_i - x_i^2 + \sum_{j=1}^K O(s_j)}{1+\sum_{j=1}^K O(s_j)}.\end{aligned}$$

If we consider the values of selection, s_i, sufficiently small the statement follows. □

1.3 Relation between Forward and Backward

Consider both differential operators of the Theorem 1.2.3:

$$\begin{aligned}L^+(\xi) &= \sum_{i=1}^K\sum_{j=1}^K \frac{1}{2}\frac{\partial^2\left[\mathrm{Cov}(\Delta x_i,\Delta x_j|\mathcal{F}(t))\xi\right]}{\partial x_i\partial x_j} - \sum_{i=1}^K \frac{\partial\left[\mathrm{E}[\Delta x_i|\mathcal{F}_t]\xi\right]}{\partial x_i},\\ L^-(\xi) &= \sum_{i=1}^K\sum_{j=1}^K \frac{1}{2}\mathrm{Cov}(\Delta x_i,\Delta x_j|\mathcal{F}(t))\frac{\partial^2\xi}{\partial x_i\partial x_j}\sum_{i=1}^K \mathrm{E}[\Delta x_i|\mathcal{F}_t]\frac{\partial\xi}{\partial x_i}.\end{aligned}$$

Proposition 1.3.1. *Consider $f, g \in \mathcal{C}^\infty$ with compact support contained in $(0,1)$, then L^+, L^- are adjoint operators, $< f, L^+ g >=< L^- f, g >$.*

Proof. The proof of the statement is only given for $K = 2$. Using the Corollary 1.2.4:

$$\begin{aligned}L^+(\xi) &= -s\frac{\partial}{\partial x}[x(1-x)\xi] + \frac{1}{2N_e}\frac{\partial^2}{\partial x^2}[x(1-x)\xi]\\ L^-(\xi) &= sx(1-x)\frac{\partial\xi}{\partial x} + \frac{x(1-x)}{2N_e}\frac{\partial^2\xi}{\partial x^2}\end{aligned}$$

where $s = s_1 - s_2$.

$$
\begin{aligned}
< f, L^+ g > &= \int_0^1 f(x) \left(-s\frac{\partial}{\partial x}[x(1-x)g(x)] + \frac{1}{2N_e}\frac{\partial^2}{\partial x^2}[x(1-x)g(x)] \right) dx \\
&= -s\int_0^1 f(x)\frac{\partial}{\partial x}[x(1-x)g(x)]dx + \frac{1}{2N_e}\int_0^1 f(x)\frac{\partial^2}{\partial x^2}[x(1-x)g(x)]dx \\
&=^3 -s\left([f(x)x(1-x)g(x)]_0^1 - \int_0^1 \frac{\partial f(x)}{\partial x}x(1-x)g(x)dx \right) \\
&+ \frac{1}{2N_e}\left(f(x)\frac{\partial}{\partial x}[x(1-x)g(x)] - \int_0^1 \frac{\partial f(x)}{\partial x}\frac{\partial}{\partial x}[x(1-x)g(x)]dx \right) \\
&=^4 \int_0^1 sx(1-x)\frac{\partial f(x)}{\partial x}g(x)dx + \frac{\left[f(x)\left(x(1-x)\frac{\partial g(x)}{\partial x} + g(x)(1-2x) \right) \right]_0^1}{2N_e} \\
&- \frac{1}{2N_e}\left(\left[\frac{\partial f(x)}{\partial x}x(1-x)g(x) \right]_0^1 - \int_0^1 x(1-x)\frac{\partial^2 f(x)}{\partial x^2}g(x)dx \right) \\
&= \int_0^1 sx(1-x)\frac{\partial f(x)}{\partial x}g(x)dx + \frac{1}{2N_e}\int_0^1 x(1-x)\frac{\partial^2 f(x)}{\partial x^2}g(x)dx \\
&= \int_0^1 \left(sx(1-x)\frac{\partial f(x)}{\partial x}dx + \frac{1}{2N_e}x(1-x)\frac{\partial^2 f(x)}{\partial x^2} \right) g(x)dx \\
&=< L^- f, g > .
\end{aligned}
$$

Doing the first two integrations by parts in equality 3 and the other in the 4:

$$
\begin{aligned}
u &= f(x), & dv &= \frac{\partial}{\partial x}[x(1-x)g(x)]dx; \\
u &= f(x), & dv &= \frac{\partial^2}{\partial x^2}[x(1-x)g(x)]dx; \\
u &= \frac{\partial f(x)}{\partial x}, & dv &= \frac{\partial}{\partial x}[x(1-x)g(x)]dx.
\end{aligned}
$$

□

Chapter 2

Stationary solution for two mutations equation

In this chapter we are going to describe the evolution of certain mutations in a specific ambient using the forward Kolmogorov equations. Principally, we want study the stationary solutions in the case of 2 different alleles and the neutral stationary solutions in the case of 3 different alleles. In this case we are going to use the definitions of nested, co-occurring, containing, complementary and exclusive that we remind in the Figure 2.1.

2.1 One mutation

Thanks to the Corollary 1.2.4, we know that to study the dynamic when there is only one mutation it is sufficient to use the following equation:

$$\frac{\partial \xi}{\partial t} = -s\frac{\partial}{\partial f}[f(1-f)\xi] + \frac{1}{2N_e}\frac{\partial^2}{\partial f^2}[f(1-f)\xi], \tag{2.1.1}$$

where $s = (s_1 - s_2)$.

It is easily checked that the following function is the most general stationary solution for the differential equation:

$$\xi = \frac{ke^{2sN_e f} + k'}{f(1-f)}.$$

Despite that, our solution is generally divergent in 0 and 1, and modifying the constants we only can impose convergence in one of the point, non in both. However, this equation describes the evolution of a single mutation. In order to solve this problem, we add a flow of new mutations born at uniform rate μN_e at low frequency $f = 1/N_e = \varepsilon$, which is biologically consistent because such mutations occur in many loci along the chromosome, and the allele frequency spectrum counts their number. Formally, we represent that adding a Dirac delta in the formula 2.1.1:

$$\frac{\partial \xi}{\partial t} = -s\frac{\partial}{\partial f}[f(1-f)\xi] + \frac{1}{2N_e}\frac{\partial^2}{\partial f^2}[f(1-f)\xi] + \mu N_e \delta(f-\varepsilon), \tag{2.1.2}$$

where μ is a constant that can be interpreted as the average number of mutations born in unit time in a single individual.

Therefore, we will have two solutions, ξ_0 for the values of $f \in [0, \varepsilon)$ convergent at zero and ξ_1 for the values of $f \in (\varepsilon, 1]$ convergent at one. In addition, we want our solution to be continuous, i.e. that both solutions coincide at ε and that the only discontinuity in the derivative coincide with the one generated by the Dirac δ, that we can find integrating the equation 2.1.2 around ε.

Lemma 2.1.1. *The discontinuity of the first derivative of $f(1-f)\xi$ is*

$$\left[\frac{\partial[f(1-f)\xi]}{\partial f}\right]_{\varepsilon^-}^{\varepsilon^+} = -2\mu N_e^2.$$

Proof. What we have to do is integrate the equation 2.1.2 and evaluate it between ε^+ and ε^- using that f and ξ are continuous.

$$\begin{aligned}
&\lim_{\varepsilon' \to 0} \int_{\varepsilon-\varepsilon'}^{\varepsilon+\varepsilon'} \left(-s\frac{\partial}{\partial f}[f(1-f)\xi] + \frac{1}{2N_e}\frac{\partial^2}{\partial f^2}[f(1-f)\xi] + \mu N_e \delta(f-\varepsilon)\right) df \\
=&\left[-sf(1-f)\xi + \frac{1}{2N_e}\frac{\partial}{\partial f}[f(1-f)\xi]\right]_{\varepsilon^-}^{\varepsilon^+} + \mu N_e \\
=&\frac{1}{2N_e}\left[\frac{\partial}{\partial f}[f(1-f)\xi]\right]_{\varepsilon^-}^{\varepsilon^+} + \mu N_e.
\end{aligned}$$

Then, since we are considering that we are in a stationary solution, we finally find the expected result.

□

Remark that we have calculated the first derivative of $f(1-f)\xi$ instead of the first derivative of ξ because in the following theorem now it will be easier to compute the values of k_0, k_0', k_1 and k_1'.

Theorem 2.1.2. *The equation that solves the differential equation 2.1.2 is $\xi(f) = \xi_0(f)\mathbb{1}_{[0,1/N_e]}(f) + \xi_1(f)\mathbb{1}_{(1/N_e,1]}(f)$, where*

$$\begin{aligned}
\xi_1(f) &= \frac{\mu N_e}{s(1-e^{2sN_e})}\frac{(1-e^{-2s})e^{2sN_e f} - (e^{2sN_e} - e^{2s(N_e-1)})}{f(1-f)}, \\
\xi_0(f) &= \frac{\mu N_e}{s(1-e^{2sN_e})}\frac{(1-e^{2s(N_e-1)})e^{2sN_e f} - (1-e^{2s(N_e-1)})}{f(1-f)}.
\end{aligned}$$

Proof. First of all we define ξ_0 and ξ_1 as:

$$\xi_0 = \frac{k_0 e^{2sN_e f} - k_0'}{f(1-f)}, \qquad \xi_1 = \frac{k_1 e^{2sN_e f} - k_1'}{f(1-f)}. \tag{2.1.3}$$

Since we impose that ξ_0 must converge at 0, we have the following equality $k_0 e^{2sN_e 0} - k_0' = 0 \Rightarrow k_0' = k_0$, and since ξ_1 must converge at 1, we have $k_1 e^{2sN_e 1} - k_1' = 0 \Rightarrow k_1' = e^{2sN_e} k_1$. Applying these two equalities to the condition of continuity of ξ in $\varepsilon = 1/N_e$:

$$\frac{k_0 e^{2sN_e\varepsilon} - k_0}{\varepsilon(1-\varepsilon)} = \frac{k_1 e^{2sN_e\varepsilon} - e^{2sN_e} k_1}{\varepsilon(1-\varepsilon)} \Rightarrow k_0(e^{2s}-1) = k_1(e^{2s} - e^{2sN_e})$$

and by the lemma 2.1.1 we found the fourth equality:

$$\begin{aligned} -2\mu N_e^2 &= \frac{\partial[f(1-f)\xi_1]}{\partial f}(1/N_e) - \frac{\partial[f(1-f)\xi_0]}{\partial f}(1/N_e) \\ &= \left[(k_1-k_0)2sN_e e^{2sN_e f}\right]_{f=1/N_e} \\ &= (k_1-k_0)2sN_e e^{2s}, \end{aligned}$$

then,

$$-\mu N_e = e^{2s}(k_1 - k_0)s.$$

Finally using the two last equations:

$$\begin{aligned} &k_1\left(1 - \frac{e^{2s} - e^{2sN_e}}{e^{2s}-1}\right) = \frac{-\mu N_e e^{-2s}}{s} \Rightarrow k_1\left(\frac{e^{2sN_e}-1}{e^{2s}-1}\right) = \frac{-\mu N_e e^{-2s}}{s} \\ &\Rightarrow k_1 = \frac{\mu N_e(1-e^{-2s})}{s(1-e^{2sN_e})} \\ &\Rightarrow k_0 = \frac{\mu N_e(1-e^{-2s})}{s(1-e^{2sN_e})}\ \frac{e^{2s}-e^{2sN_e}}{e^{2s}-1} = \frac{\mu N_e(1-e^{2s(N_e-1)})}{s(1-e^{2sN_e})}, \end{aligned}$$

therefore, applying to the initial equalities:

$$\begin{aligned} k_1' &= e^{2sN_e}\frac{\mu N_e(1-e^{-2s})}{s(1-e^{2sN_e})} = \frac{\mu N_e(e^{2sN_e} - e^{2s(N_e-1)})}{s(1-e^{2sN_e})}, \\ k_0' &= \frac{\mu N_e(1-e^{2s(N_e-1)})}{s(1-e^{2sN_e})} \end{aligned}$$

we found the result. □

Finally, in the limit of large N_e, with $N_e s$ and $\theta = 2\mu N_e$ fixed, we obtain the classical result by Wright and Kimura for the frequency spectrum with selection:

$$\xi(f) = \frac{\theta(1-e^{-2sN_e(1-f)})}{(1-e^{-2sN_e})f(1-f)}$$

that converges, in the limit $s \to 0$, to the neutral frequency spectrum

$$\xi(f) = \frac{\theta}{f}$$

2.2 Two mutations

This case is similar to the case of one mutation. Also in this case we can reduce the number of variables in the evolution equation.

Two mutations without recombination give rise to three possible variants, as we will see later.

Proposition 2.2.1. *The forward Kolmogorov equation for three alleles satisfies the following equality:*

$$\begin{aligned}\frac{\partial \xi}{\partial t} =& \frac{1}{2N_e}\left(\frac{\partial^2}{\partial f_1^2}\left[f_1(1-f_1)\xi\right] + \frac{\partial^2}{\partial f_1 f_2}\left[-2f_1f_2\xi\right] + \frac{\partial^2}{\partial f_2^2}\left[f_2(1-f_2)\xi\right]\right) \\ &- \frac{\partial}{\partial f_1}\left[(s_1f_1(1-f_1) - s_2f_1f_2 - s_3f_1(1-f_1-f_2))\xi\right] \\ &- \frac{\partial}{\partial f_2}\left[(s_2f_2(1-f_2) - s_1f_1f_2 - s_3f_2(1-f_1-f_2))\xi\right],\end{aligned} \tag{2.2.1}$$

where f_i is the frequency of the allele i.

Proof. Doing the following change of variables $f_1 = x_1$, $f_2 = x_2$, $r = x_1 + x_2 + x_3$ we have that $\frac{\partial}{\partial x_1} = \frac{\partial}{\partial f_1} + \frac{\partial}{\partial r}$, $\frac{\partial}{\partial x_2} = \frac{\partial}{\partial f_2} + \frac{\partial}{\partial r}$, $\frac{\partial}{\partial x_3} = \frac{\partial}{\partial r}$. Doing all the calculus and a reordering of the terms:

$$\begin{aligned}\frac{\partial \xi}{\partial t} =& \sum_{i=1}^{3}\sum_{j=1}^{3}\frac{1}{2}\frac{\partial^2}{\partial x_i \partial x_j}\left[\frac{x_i(1-x_i)}{N_e}\xi\right] - \sum_{i=1}^{3}\frac{\partial}{\partial x_i}\left[x_i\sum_{j=1}^{K}(s_i - s_j)x_j\xi\right] \\ =& \frac{1}{2N_e}\Bigg(\frac{\partial^2}{\partial f_1 f_2}\left[-2f_1f_2\xi\right] + \frac{\partial^2}{\partial f_1 r}\left[2f_1(-f_2 - (r-f_1-f_2) + 1 - f_1)\xi\right] \\ &+ \frac{\partial^2}{\partial f_2 r}\left[2f_2(-f_1 - (r-f_1-f_2) + 1 - f_2)\xi\right] + \frac{\partial^2}{\partial f_1^2}\left[f_1(1-f_1)\xi\right] \\ &+ \frac{\partial^2}{\partial f_2^2}\left[f_2(1-f_2)\xi\right] + \frac{\partial^2}{\partial r^2}[(-2f_1f_2 - 2f_1(r-f_1-f_2) - 2f_2(r-f_1-f_2) \\ &+ f_1(1-f_1) + f_2(1-f_2) + (r-f_1-f_2)(1-(r-f_1-f_2)))\xi]\Bigg) \\ &- \frac{\partial}{\partial f_1}\left[f_1((s_1-s_2)f_2 + (s_1-s_3)(r-f_1-f_2))\xi\right] \\ &- \frac{\partial}{\partial f_2}\left[f_2((s_2-s_1)f_1 + (s_2-s_3)(r-f_1-f_2))\xi\right].\end{aligned}$$

Now, using the fact that $r = 1$ all the terms with $\frac{\partial}{\partial r}$ disappear and the equation is reduced to the final expression.

□

Notation 2.2.2. From now on, we denote f_0 as the frequency of the focal mutation and f the frequency of the second mutation. It is important to not confuse these frequencies with f_1, f_2 and f_3, the frequencies of the alleles.

From the results in the article [3], the stationary solution, $\xi(f|f_0)$, for the frequency spectrum of the second mutation can be broken into two different components, one component ξ^S containing mutations that share some individuals with the focal mutation, the other ξ^E that includes mutation that are mutually exclusive with the focal one. The spectrum is given by $\xi = \xi^S + \xi^E$. These component could be further broken into different subspectra such that $\xi^S = \xi^{(n)} + \xi^{(co)} + \xi^{(c)}$ and $\xi^E = \xi^{(cm)} + \xi^{(e)}$, where:

- $\xi^{(n)}$: nested mutations, i.e. mutations occurring in a subset of the individuals carrying the focal mutation;
- $\xi^{(co)}$: co-occurring mutations, i.e. occurring on the same individuals as the focal mutation;
- $\xi^{(c)}$: containing mutations, i.e. the focal mutation occurs in a subset of this;
- $\xi^{(cm)}$: complementary mutations, i.e. each individual has either this mutation or the focal one;
- $\xi^{(e)}$: exclusive mutations, i.e. involving a set of individuals that is non-overlapping and not complementary with the focal one.

It can be seen that these are the only possible cases without recombination, since either two mutations are born in different backgrounds (then they are complementary or exclusive), or one is born inside the other (in this case they are nested/containing or co-occurring).

Now it is sufficient to find the components of the spectrum. This has already been done through polynomial expansions [6], but simpler solutions are possible, at least in the neutral case. We are going to use an unpublished result obtained by Luca Ferretti using methods from coalescent theory. The principal steps to find these components in the neutral case are applying the limit $n \to \infty$ to the equations (14-16) of the article [3] divided then by the full spectrum θ/l and multiplying by the length L of the sequence. Up to a global multiplicative factor θL, the result is

$$
\begin{aligned}
\xi^{(n)}(f|f_0) =& \frac{f_0}{(1-f)^2}\left(1+\frac{1}{f}+\frac{2\ln(f)}{1-f}\right)\\
\xi^{(co)}(f|f_0) =& \xi^{(co)}(f_0) = \frac{2f_0}{1-f_0}\left(-\frac{\ln(f_0)}{1-f_0}-1\right)\\
\xi^{(c)}(f|f_0) =& \frac{f_0}{(1-f_0)^2}\left(1+\frac{1}{f_0}+\frac{2\ln(f_0)}{1-f_0}\right)\\
\xi^{(cm)}(f|f_0) =& \xi^{(cm)}(f_0) = \left[\frac{1-f_0}{f_0}\log(1-f_0)+\left(\frac{f_0}{1-f_0}\right)^2\log(f_0)+\frac{1}{1-f_0}\right]\\
\xi^{(e)}(f|f_0) =& \left[\frac{1}{f}-\frac{f_0}{(1-f)^2}\left(1+\frac{1}{f}+\frac{2\ln(f)}{1-f}\right)-\frac{f_0}{(1-f_0)^2}\left(1+\frac{1}{f_0}+\frac{2\ln(f_0)}{1-f_0}\right)\right]
\end{aligned}
$$

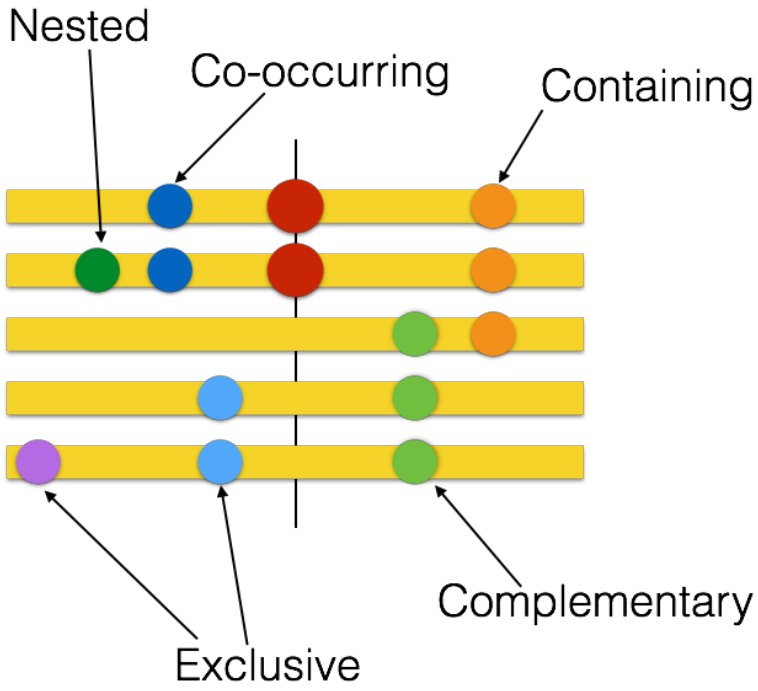

Figure 2.1: Sequence variants and classification of the possible types of mutation with respect to the focal mutation (in red).

Since we have $\xi^{(i)}(f, f_0) \propto \xi^{(i)}(f|f_0)\xi(f_0)$, all $\xi^{(i)}(f|f_0)/f_0$ for $i \in \{n, co, c, cm, e\}$ are solutions of the differential equation. We are going to study the cases nested and exclusive because the nested case is the symmetric of the containing case, while co-occurring and complementary are border cases that are difficult to study.

Proposition 2.2.3. *The equations $\xi^{(n)}$ and $\xi^{(e)}$ are stationary solutions in the case of neutral selection of the equation 2.2.1.*

Proof. In nested case the appropriate change of variables is $f = f_1, f_0 = f_1 + f_2$. Then dividing $\xi^{(n)}(f|f_0)$ by $f_1 + f_2$ we found an stationary solution:

$$\xi^{(n)}(f, f_0) = \frac{1}{(1-f_1)^2}\left(1 + \frac{1}{f_1} + \frac{2\ln(f_1)}{1-f_1}\right).$$

We know that it is a stationary solution because using the second part of the equation 2.2.1 we found that:

$$\frac{\partial \xi^{(n)}(f, f_0)}{\partial t} = 0.$$

In the exclusive case the change of variables is $f = f_1, f_0 = 1 - f_1 - f_2$ and

dividing $\xi^{(e)}(f|f_0)$ between $1-f_1-f_2$ we found another stationary solution:

$$\xi^{(e)}(f,f_0) = \frac{1}{f_1(1-f_1-f_2)} - \frac{1}{(1-f_1)^2}\left(1+\frac{1}{f_1}+\frac{2\ln(f_1)}{1-f_1}\right) - \frac{1}{(f_1+f_2)^2}\left(1+\frac{1}{1-f_1-f_2}+\frac{2\ln(1-f_1-f_2)}{f_1+f_2}\right).$$

And as in the previous case:

$$\frac{\partial \xi^{(n)}(f,f_0)}{\partial t} = 0.$$

□

As in the case of two mutations, given a f_0, we want to study the function behaviour near the border, $f=0$ and $f=1$, that is where we can have problems using our definition of the solution. For the values of f near 1 there is not any problem because the limits exist in both cases:

$$\begin{aligned}\lim_{f\to 1^-}\xi^{(n)}(f|f_0) &= \lim_{f\to 1^-}\frac{f_0}{(1-f)^2}\left(1+\frac{1}{f}+\frac{2\ln(f)}{1-f}\right) = f_0\lim_{f\to 1^-}\frac{1-f^2+2f\ln(f)}{f-3f^2+3f^3-f^4}\\ &= f_0\lim_{f\to 1^-}\frac{-2f+2\ln(f)+2ff^{-1}}{1-6f+9f^2-4f^3} = f_0\lim_{f\to 1^-}\frac{-2+2f^{-1}}{-6+18f-12f^2}\\ &= f_0\lim_{f\to 1^-}\frac{-2f^{-2}}{18-24f} = \frac{f_0}{3}.\end{aligned}$$

$$\lim_{f\to 1^-}\xi^{(e)}(f|f_0) = \lim_{f\to 1^-}\frac{1}{f} - \frac{f_0}{(1-f)^2}\left(1+\frac{1}{f}+\frac{2\ln(f)}{1-f}\right) + g(f_0) = 1-\frac{f_0}{3}+g(f_0),$$

where $g(f_0)$ is a function that does not affect because do not have dependence on f.

Despite that, when f goes to zero the solution tends to infinity in both cases. In order to solve it, we find another solution not divergent for that values of $f\in(0,\varepsilon)$ such that the values of the two solution coincide in $f=\varepsilon$. We formalize that adding in our differential equation 2.2.1 a Dirac delta of $f-\varepsilon$ multiplied by an adequate coefficient as we can see in the equation 2.2.2. As in the case with one mutation that is biologically consistent because it is like if our mutation was lost and another appears and replaces it.

$$\begin{aligned}\frac{\partial\xi}{\partial t} =& \frac{1}{2N_e}\left(\frac{\partial^2}{\partial f_1^2}\left[f_1(1-f_1)\xi\right] + \frac{\partial^2}{\partial f_1 f_2}\left[-2f_1f_2\xi\right] + \frac{\partial^2}{\partial f_2^2}\left[f_2(1-f_2)\xi\right]\right)\\ &- \frac{\partial}{\partial f_1}\left[(s_1f_1(1-f_1) - s_2f_1f_2 - s_3f_1(1-f_1-f_2))\xi\right]\\ &- \frac{\partial}{\partial f_2}\left[(s_2f_2(1-f_2) - s_1f_1f_2 - s_3f_2(1-f_1-f_2))\xi\right]\\ &+ \frac{\theta}{f_0}\mu N_e\mu(f_0)\delta(f-\varepsilon).\end{aligned} \tag{2.2.2}$$

$+\frac{\theta}{f_0}\mu N_e f_0 \delta(f-1/N_e) = \frac{\theta^2}{2}\delta(f-1/N_e)$ where $\frac{\theta}{f_0}$ is the density of mutations with frequency f_0 and $\mu N_e \mu(f_0)$ is the number of mutations born inside $\mu(f_0)$. For example, in the nested case $\mu(f_0) = f_0$ because the second mutation is nested in f_0 and in the exclusive case $\mu(f_0) = 1 - f_0$ because of the opposite reason.

As in the case of one mutation, here we want a continuous solution with a discontinuity in the first derivative of ξ differential f in the point $f = \varepsilon$.

Unfortunately, we only are able to find the value of the discontinuity in the derivative.

Proposition 2.2.4. *In the nested case, using that $\varepsilon = 1/N_e$, the discontinuity of the first derivative of ξ is*

$$\left[\frac{\partial \xi}{\partial f}\right]_{\frac{1}{N_e}^-}^{\frac{1}{N_e}^+} \simeq -2(1-f_0)\left[\frac{\partial \xi}{\partial f_0}\right]_{\varepsilon^-}^{\varepsilon^+} - (\theta N_e)^2.$$

Proof. First of all, we do the change of variables for the nested case change of variable $f = f_1$ and $f_0 = f_1 + f_2$. Then, we integrate the equation 2.2.2 with $\nu(f_0) = f_0$ and $\theta = 2\mu N_e$. In the evaluation of the integral we using that the unique components no-continuous in $f = \varepsilon$ are the $\frac{\partial \xi}{\partial f}$ and $\frac{\partial \xi}{\partial f}$.

$$\begin{aligned}
&\lim_{\varepsilon'\to 0}\int_{\varepsilon-\varepsilon'}^{\varepsilon+\varepsilon'} \frac{1}{2}\left(\frac{\partial^2}{\partial f_1^2}\left[\frac{f_1(1-f_1)}{N_e}\xi\right] + \frac{\partial^2}{\partial f_1 f_2}\left[\frac{-2f_1 f_2}{N_e}\xi\right] + \frac{\partial^2}{\partial f_2^2}\left[\frac{f_2(1-f_2)}{N_e}\xi\right]\right)\\
&- \frac{\partial}{\partial f_1}\left[(s_1 f_1(1-f_1) - s_2 f_1 f_2 - s_3 f_1(1-f_1-f_2))\xi\right]\\
&- \frac{\partial}{\partial f_2}\left[(s_2 f_2(1-f_2) - s_1 f_1 f_2 - s_3 f_2(1-f_1-f_2))\xi\right] + \frac{\theta^2}{2}\delta(f-\varepsilon)df\\
=&\lim_{\varepsilon'\to 0}\int_{\varepsilon-\varepsilon'}^{\varepsilon+\varepsilon'} \frac{1}{2}\left(\frac{\partial^2}{\partial f^2}\left[\frac{f(1-f)}{N_e}\xi\right] + \frac{\partial^2}{\partial f f_0}\left[\frac{2f(1-f_0)}{N_e}\xi\right]\right.\\
&\left.+\frac{\partial^2}{\partial f_0^2}\left[\frac{(f-f_0)(-1+f+f_0)}{N_e}\xi\right]\right) - \frac{\partial}{\partial f}\left[(s_1 f(1-f) - s_2 f(f_0-f)\right.\\
&\left.-s_3 f(1-f_0))\xi\right] - \frac{\partial}{\partial f_0}\left[(s_1 f(1-f_0) - s_2(f_0-f)(1-f_0+2f)\right.\\
&\left.-s_3(2f-f_0)(1-f_0))\xi\right] + \frac{\theta^2}{2}\delta(f-\varepsilon)df\\
=&\left[\frac{1}{2}\left(\frac{\partial}{\partial f}\left[\frac{f(1-f)}{N_e}\xi\right] + \frac{\partial}{\partial f_0}\left[\frac{2f(1-f_0)}{N_e}\xi\right]\right)\right]_{\varepsilon^-}^{\varepsilon^+} + \frac{\theta^2}{2}\\
=&\left[\frac{1}{2N_e}\left((-2f+1)\xi + f(1-f)\frac{\partial \xi}{\partial f} - 2f\xi + 2f(1-f_0)\frac{\partial \xi}{\partial f_0}\right)\right]_{\varepsilon^-}^{\varepsilon^+} + \frac{\theta^2}{2}\\
=&\left[\frac{1}{2N_e}\left(f(1-f)\frac{\partial \xi}{\partial f} + 2f(1-f_0)\frac{\partial \xi}{\partial f_0}\right)\right]_{\varepsilon^-}^{\varepsilon^+} + \frac{\theta^2}{2}.
\end{aligned}$$

Then, since we are considering that we are in a stationary solution:

$$\left[f(1-f)\frac{\partial\xi}{\partial f}+2f(1-f_0)\frac{\partial\xi}{\partial f_0}\right]_{\varepsilon^-}^{\varepsilon^+}=-\theta^2N_e$$
$$\varepsilon(1-\varepsilon)\left[\frac{\partial\xi}{\partial f}\right]_{\varepsilon^-}^{\varepsilon^+}+2\varepsilon(1-f_0)\left[\frac{\partial\xi}{\partial f_0}\right]_{\varepsilon^-}^{\varepsilon^+}=-\theta^2N_e$$
$$\left[\frac{\partial\xi}{\partial f}\right]_{\varepsilon^-}^{\varepsilon^+}=\frac{1}{\varepsilon(1-\varepsilon)}\left(-2\varepsilon(1-f_0)\left[\frac{\partial\xi}{\partial f_0}\right]_{\varepsilon^-}^{\varepsilon^+}-\theta^2N_e\right).$$

Using $\varepsilon=1/N_e$ we find the searched value.

□

Proposition 2.2.5. *In the exclusive case, using that $\varepsilon=1/N_e$, the discontinuity of the first derivative of ξ is*

$$\left[\frac{\partial\xi}{\partial f}\right]_{\varepsilon^-}^{\varepsilon^+}=2f_0\left[\frac{\partial\xi}{\partial f_0}\right]_{\varepsilon^-}^{\varepsilon^+}-\frac{\theta^2(1-f_0)N_e^2}{f_0}.$$

Proof. As in the nested case, first of all, we do adequate change of variable $f=f_1$ and $f_0=1-f_1-f_2$. Then, we integrate the equation 2.2.2 $\nu(f_0)=1-f_0$ and $\theta=2\mu N_e$. Then, we use that unique components no-continuous in $f=\varepsilon$ are the $\frac{\partial\xi}{\partial f}$ and $\frac{\partial\xi}{\partial f}$.

$$\begin{aligned}
&\lim_{\varepsilon'\to0}\int_{\varepsilon-\varepsilon'}^{\varepsilon+\varepsilon'}\frac{1}{2}\left(\frac{\partial^2}{\partial f_1^2}\left[\frac{f_1(1-f_1)}{N_e}\xi\right]+\frac{\partial^2}{\partial f_1f_2}\left[\frac{-2f_1f_2}{N_e}\xi\right]+\frac{\partial^2}{\partial f_2^2}\left[\frac{f_2(1-f_2)}{N_e}\xi\right]\right)\\
&-\frac{\partial}{\partial f_1}\left[(s_1f_1(1-f_1)-s_2f_1f_2-s_3f_1(1-f_1-f_2))\xi\right]\\
&-\frac{\partial}{\partial f_2}\left[(s_2f_2(1-f_2)-s_1f_1f_2-s_3f_2(1-f_1-f_2))\xi\right]+\frac{\theta^2(1-f_0)}{2f_0}\delta(f-\varepsilon)df\\
=&\lim_{\varepsilon'\to0}\int_{\varepsilon-\varepsilon'}^{\varepsilon+\varepsilon'}\frac{1}{2}\left(\frac{\partial^2}{\partial f^2}\left[\frac{f(1-f)}{N_e}\xi\right]+\frac{\partial^2}{\partial ff_0}\left[\frac{-2ff_0}{N_e}\xi\right]-\frac{\partial^2}{\partial f_0^2}\left[\frac{f_0(1-f_0)}{N_e}\xi\right]\right)\\
&-\frac{\partial}{\partial f}\left[(s_1f(1-f)-s_2f(1-f-f_0)-s_3f(1-f-(1-f-f_0)))\xi\right]\\
&+\frac{\partial}{\partial f_0}\left[(s_2f_0(1-f-f0)+s_1ff_0+s_3f_0(-2+f+f_0))\xi\right]+\frac{\theta^2(1-f_0)}{2f_0}\delta(f-\varepsilon)df\\
=&\left[\frac{1}{2}\left(\frac{\partial}{\partial f}\left[\frac{f(1-f)}{N_e}\xi\right]+\frac{\partial}{\partial f_0}\left[\frac{-2ff_0}{N_e}\xi\right]\right)\right]_{\varepsilon^-}^{\varepsilon^+}+\frac{\theta^2(1-f_0)}{2f_0}\\
=&\left[\frac{1}{2N_e}\left((-2f+1)\xi+f(1-f)\frac{\partial\xi}{\partial f}-2f\xi-2ff_0\frac{\partial\xi}{\partial f_0}\right)\right]_{\varepsilon^-}^{\varepsilon^+}+\frac{\theta^2(1-f_0)}{2f_0}\\
=&\left[\frac{1}{2N_e}\left(f(1-f)\frac{\partial\xi}{\partial f}-2ff_0\frac{\partial\xi}{\partial f_0}\right)\right]_{\varepsilon^-}^{\varepsilon^+}+\frac{\theta^2(1-f_0)}{2f_0}.
\end{aligned}$$

Then, since we are considering that we are in a stationary solution:

$$\begin{aligned}
\left[f(1-f)\frac{\partial \xi}{\partial f} - 2f f_0 \frac{\partial \xi}{\partial f_0}\right]_{\varepsilon^-}^{\varepsilon^+} &= -\frac{\theta^2(1-f_0)N_e}{f_0} \\
\varepsilon(1-\varepsilon)\left[\frac{\partial \xi}{\partial f}\right]_{\varepsilon^-}^{\varepsilon^+} - 2\varepsilon f_0 \left[\frac{\partial \xi}{\partial f_0}\right]_{\varepsilon^-}^{\varepsilon^+} &= -\frac{\theta^2(1-f_0)N_e}{f_0} \\
\left[\frac{\partial \xi}{\partial f}\right]_{\varepsilon^-}^{\varepsilon^+} &= \frac{1}{\varepsilon(1-\varepsilon)}\left(2\varepsilon f_0 \left[\frac{\partial \xi}{\partial f_0}\right]_{\varepsilon^-}^{\varepsilon^+} - \frac{\theta^2(1-f_0)N_e}{f_0}\right).
\end{aligned}$$

Using $\varepsilon = 1/N_e$ we find the searched value.

□

Chapter 3

Evolution of mutations after an environmental change

3.1 Introduction to the problem

In this chapter we are going to describe the evolution of certain mutations in a specific environment. Mainly, we want to study the result of a sudden change in the fitness of a derived allele in a population with different mutations, evolving neutrally up to this moment. The evolution of the selected allele changes the frequency of the alleles at different positions in the sequence. We are going to differentiate between three cases: nested, containing and exclusive mutations, because each of them presents a different evolution of the neutral mutation. The evolution in the other two cases - co-occurring and complementary - is directly related to the evolution of the selected mutation.

The assumptions for the model are the same as in the previous chapters of the project, but we have to add/change some notation:

- f_F will denote the frequency of the selected mutation whereas f will denote the frequency of the other one.
- $\xi_N(f|f_F(0))$ is the stationary neutral spectrum defined in Chapter 2.

First of all, we are going to consider the trajectories of the selected mutation. We assume a deterministic dynamics in a rescaled time that makes equations easier. This new time definition comes from $t = st'$ where t' is the previous time and s is the allele fitness.

Proposition 3.1.1. *The deterministic dynamics of the selected mutation is:*

$$f_F(t) = \frac{f_F(0)}{f_F(0) + (1 - f_F(0))e^{-t}}.$$

Proof. Using the rescaling of time and the Proposition 1.2.5 with $K = 2$, we found that the equation for the selected mutation f_F is the solution of:

$$\frac{df_F}{dt} = \frac{1}{s} E\left[\Delta(f_F)|\mathcal{F}_t\right] = f_F(1 - f_F),$$

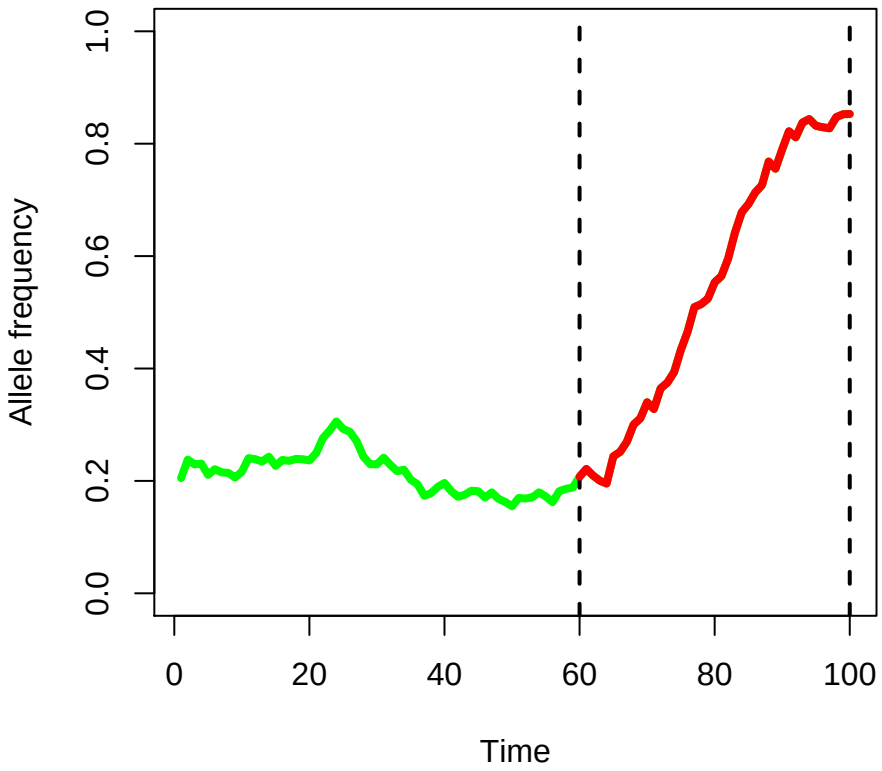

Figure 3.1: Example of the trajectory of the focal allele in the problem. From now on in pictures, in green the neutral evolution, in red the selective one.

with an arbitrary initial condition $f_F(0)$. The solution of the differential equation is:

$$f_F(t) = \frac{e^t}{e^t + k} = \frac{1}{1 + ke^{-t}} = \frac{f_F(0)}{f_F(0) + (1 - f_F(0))e^{-t}}.$$

In order to determine k, we apply that the initial condition is $f_F(0)$:

$$f_F(0) = \frac{1}{1+k} \Rightarrow k = \frac{1}{f_F(0)} - 1 = \frac{1 - f_F(0)}{f_F(0)}.$$

□

Then, notice that if $f_F(0) \neq 0$ and $f_F(0) \neq 1$ we have the following relations that we will use to solve equations in the future.

$$e^{-t} = \frac{f_F(0)(1 - f_F(t))}{f_F(t)(1 - f_F(0))} \qquad e^t = \frac{f_F(t)(1 - f_F(0))}{f_F(0)(1 - f_F(t))}. \tag{3.1.1}$$

Since we have an explicit formula for f_F, the differentials of f_F do not appear in the forward Kolmogorov equations as in the Chapter 2 problem. Now our rescaled differential equation (forward) is given by the following expression:

$$\begin{aligned}\frac{\partial \xi}{\partial t} &= \frac{1}{2N_e}\frac{\partial^2}{\partial f^2}\left[\mathrm{Var}\left(\Delta f|\mathcal{F}_t\right)\xi\right] - \frac{1}{s}\frac{\partial}{\partial f}\left[\mathrm{E}\left[\Delta f|\mathcal{F}_t\right]\xi\right] \;\Rightarrow \\ \frac{\partial \xi}{\partial t} &= \frac{1}{2N_e s}\frac{\partial^2}{\partial f^2}\left[f(1-f)\xi\right] - \frac{\partial}{\partial f}\left[E\left[\Delta f|\mathcal{F}_t\right]\xi\right]. \end{aligned} \tag{3.1.2}$$

where the formula for variance is given by the Proposition 1.2.5 and the formula for the expected value is different for the nested, exclusive and containing case. Despite that, every case $E\left[\Delta f|\mathcal{F}_t\right]$ has a linear dependence of s, therefore s appears only in the first term at the right-hand side as the combination $1/2N_e s$.

3.2 Nested Case

As the Figure 3.2 reflects, in that case the mutation with f is nested inside the selected mutation, i.e. all individuals with this mutation have also the selected mutation.

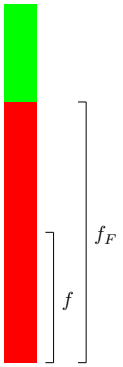

Figure 3.2: Explanation of the nested case.

Proposition 3.2.1. *In the nested case, the differential equation that we have to solve is*

$$\frac{\partial \xi(f|f_F(0),t)}{\partial t} = \frac{1}{2sN_e}\frac{\partial^2}{\partial f^2}\left[f(1-f)\xi(f|f_F(0),t)\right] - \frac{\partial}{\partial f}\left[f(1-f_F)\xi(f|f_F(0),t)\right].$$

Proof. What we only have to do is transforming the expected value of the equation 3.1.2 into an explicit formula. So, we apply the Proposition 1.2.5 with $K = 3$, $x_1 = f$, $x_2 = f_F - f$ and $x_3 = 1 - f_F$. Since $s_1 = s_2 = s$ and $s_3 = 0$:

$$E\left[\Delta f|\mathcal{F}_t\right] = E\left[\Delta x_1|\mathcal{F}_t\right] = x_1 s x_3 = sf(1-f_F).$$

□

Since we cannot find an explicit solution for this differential equation, we expand it in powers of $1/S = 1/2N_e s$ and we are going to find the solution for each order in a recursive way.

$$\xi(f|f_F(0),t) = \sum_{n=0}^{\infty}\left(\frac{1}{2N_e s}\right)^n \xi_n(f|f_F(0),t).$$

Hence, supposing that $1 \gg s \gg 1/N_e > 0$ we can solve it order by order starting at 0. The lowest orders are presumably more relevant than the higher ones. So, we are able to use the expression of ξ_i to solve ξ_{i+1}:

$$\begin{aligned}\frac{\partial \xi_0(f|f_F(0),t)}{\partial t} &= -\frac{\partial}{\partial f}\left[f(1-f_F)\xi_0(f|f_F(0),t)\right],\\ \frac{\partial \xi_{i+1}(f|f_F(0),t)}{\partial t} &= \frac{\partial^2}{\partial f^2}\left[f(1-f)\xi_i(f|f_F(0),t)\right] - \frac{\partial}{\partial f}\left[f(1-f_F)\xi_{i+1}(f|f_F(0),t)\right].\end{aligned} \tag{3.2.1}$$

The Green function of the system for $s = 0$ is going to solve both partial differential equations.

Theorem 3.2.2. *In the nested case the Green function for $s = 0$ is*

$$G(f,t|\tilde{f},\tilde{t}) = \theta\left(t-\tilde{t}\right)\delta\left(\tilde{f} - \frac{f(t)f_F(\tilde{t})}{f_F(t)}\right)\frac{f_F(\tilde{t})}{f_F(t)}.$$

Proof. Consider $\tilde{f}$ the initial frequency of f and $\tilde{t}$ the initial time, then the Green function is the function G such that satisfies:

$$\begin{aligned}&\frac{\partial G(f,t|\tilde{f},\tilde{t})}{\partial t} = \delta(f-\tilde{f})\delta(t-\tilde{t}) - \frac{\partial}{\partial f}\left[f(1-f_F)G(f,t|\tilde{f},\tilde{t})\right] \Rightarrow\\ &\frac{\partial G}{\partial t} + f(1-f_F)\frac{\partial G}{\partial f} = \delta(f-\tilde{f})\delta(t-\tilde{t}) - (1-f_F)G,\end{aligned}$$

with initial conditions $G = 0$ at $t \to -\infty$.

The solution is 0 for $t < \tilde{t}$, while the term $\delta(f-\tilde{f})\delta(t-\tilde{t})$ generates an initial condition $\delta(f-\tilde{f})$ at $t = \tilde{t}$ for the equation (valid at times $t > \tilde{t}$):

$$\frac{\partial G_2}{\partial t} + f(1-f_F)\frac{\partial G_2}{\partial f} = -(1-f_F)G_2.$$

Hence, applying the method of characteristics to the previous equation:

$$\begin{aligned}&\frac{\partial t}{\partial s} = 1 \Rightarrow t = s + \tilde{t},\\ &\frac{\partial f}{\partial s} = (1-f_F)f \Rightarrow\\ &f(s) = f(0)e^{\int_0^s (1-f_F(x))dx} = f(0)e^{x - \log(1+(e^x-1)f_F(0))]_0^s} \Rightarrow\\ &f(t) = f(\tilde{t})\frac{e^t}{1+(e^t-1)f_F(\tilde{t})} = f(\tilde{t})\frac{1}{f_F(\tilde{t}) + (1-f_F(\tilde{t}))e^{-t}} = f(\tilde{t})\frac{f_F(t)}{f_F(\tilde{t})},\\ &\frac{\partial G_2}{\partial s} = -(1-f_F)G_2 \Rightarrow G_2(f(t),t) = \ldots = G_2(f(\tilde{t}),\tilde{t})\frac{f_F(\tilde{t})}{f_F(t)}.\end{aligned}$$

In the previous equalities we have used several tricks in order not to carry too many notation and doing too much calculus. We have considered directly that the initial time is $\tilde{t}$ and s is equivalent to $s + \tilde{t}$, we have used the formulas from Proposition 3.1.1 and Equation 3.1.1 and the computation of $f_F(x)$ for $G_2(f(t), t)$.

Now, we complete the proof with the Lemma 3.2.3 proved below, with

$$\varphi(t) = \frac{f_F(t)}{f_F(\tilde{t})}, \qquad \phi(t) = 0, \qquad \psi(t) = \frac{f_F(\tilde{t})}{f_F(t)}.$$

and we find the Green function for the nested case.

□

Lemma 3.2.3. *Let $f(t) = f(\tilde{t})\varphi(t) + \phi(t)$ and $G_2(f(t), t) = G_2(f(\tilde{t}), \tilde{t})\psi(f(t), t)$ the solution for the appropriate differential equation for $t > \tilde{t}$ where $\tilde{t}$ is the initial fixed time and $\tilde{f}$ the initial frequency of f. Then, the Green function is*

$$G(f, t|\tilde{f}, \tilde{t}) = \theta\left(t - \tilde{t}\right) \delta\left(\tilde{f} - \frac{f - \phi(t)}{\varphi(t)}\right) \psi(f, t).$$

Proof.

$$\begin{aligned} &G_2\left(f(t), t\right) = G_2(f(\tilde{t}), \tilde{t})\psi(f(t), t) \Rightarrow \\ &G_2\left(f(t), t\right) = G_2\left(\frac{f(t) - \phi(t)}{\varphi(t)}, \tilde{t}\right) \psi(f(t), t). \end{aligned}$$

using the fact that $f(\tilde{t}) = \frac{f(t) - \phi(t)}{\varphi(t)}$. Then, since G_2 is the solution for $t > \tilde{t}$ and if the initial condition for $\tilde{f}$ is $\delta(f(\tilde{t}) - \tilde{f})$ we have the following Green function:

$$G(f, t|\tilde{f}, \tilde{t}) = \begin{cases} 0 & \text{if } t < \tilde{t}, \\ 0 & \text{if } \tilde{f} \neq \frac{f - \phi(t)}{\varphi(t)}, \\ \psi(f, t) & \text{otherwise.} \end{cases}$$

Therefore:

$$G(f, t|\tilde{f}, \tilde{t}) = \theta\left(t - \tilde{t}\right) \delta\left(\tilde{f} - \frac{f - \phi(t)}{\varphi(t)}\right) \psi(f, t).$$

□

Theorem 3.2.4. *Let $\gamma(f, t)$, $H(f, t)$ and $\chi(f, t)$ be well-behaved functions such that:*

$$\frac{\partial \gamma(f, t)}{\partial t} = H(f, t) + \frac{\partial}{\partial f}\left[\chi(f, t)\gamma(f, t)\right].$$

Then, if $G(f, t|\tilde{f}, \tilde{t})$ is the Green function corresponding to $H(f, t) = 0$:

$$\gamma(f, t) = \int_\Omega H(\tilde{t}, \tilde{f}) G(f, t|\tilde{f}, \tilde{t}) d\tilde{f} d\tilde{t}.$$

Proof. The Green function is the function G such that satisfies:

$$\frac{\partial G(f,t|\tilde{f},\tilde{t})}{\partial t} = \delta(f-\tilde{f})\delta(t-\tilde{t}) + \frac{\partial}{\partial f}\left[\chi(f,f_F)G(f,t|\tilde{f},\tilde{t})\right].$$

Therefore, multiplying in both sides for $H(\tilde{t},\tilde{f})$ and integrating by $\tilde{t}$ and $\tilde{f}$:

$$\int_\Omega H(\tilde{t},\tilde{f})\frac{\partial G(f,t|\tilde{f},\tilde{t})}{\partial t}d\tilde{f}d\tilde{t} = \frac{\partial}{\partial t}\left[\int_\Omega H(\tilde{t},\tilde{f})G(f,t|\tilde{f},\tilde{t})d\tilde{f}d\tilde{t}\right],$$
$$\int_\Omega H(\tilde{t},\tilde{f})\delta(f-\tilde{f})\delta(t-\tilde{t})d\tilde{f}d\tilde{t} = H(t,f),$$
$$\int_\Omega H(\tilde{t},\tilde{f})\frac{\partial}{\partial f}\left[\chi(f,f_F)G(f,t|\tilde{f},\tilde{t})\right]d\tilde{f}d\tilde{t} = \frac{\partial}{\partial f}\left[\chi(f,f_F)\int_\Omega H(\tilde{t},\tilde{f})G(f,t|\tilde{f},\tilde{t})d\tilde{f}d\tilde{t}\right].$$

we found that:

$$\gamma(f,t) = \int_\Omega H(\tilde{t},\tilde{f})G(f,t|\tilde{f},\tilde{t})d\tilde{f}d\tilde{t}.$$

□

Corollary 3.2.5. *Via the Green function, the 0th order is the following:*

$$\xi_0(f|f_F(0),t) = \xi_N\left(\frac{ff_F(0)}{f_F(t)}\middle| f_F(0)\right)\frac{f_F(0)}{f_F(t)}.$$

The higher orders can be solved recursively as follows:

$$\xi_{i+1}(f|f_F(0),t) = \int_0^t\left[\frac{\partial^2}{\partial x^2}\left[x(1-x)\xi_i(x|f_F(0),t)\right]\right]_{x=\frac{ff_F(u)}{f_F(t)}}\frac{f_F(u)}{f_F(t)}du.$$

Proof. For the higher order therms $i+1 \geq 1$, we use the Theorem 3.2.2 and Equation 3.2.1 for applying the Theorem 3.2.4:

$$G(f,t|\tilde{f},\tilde{t}) = \theta\left(t-\tilde{t}\right)\delta\left(\tilde{f}-\frac{ff_F(\tilde{t})}{f_F(t)}\right)\frac{f_F(\tilde{t})}{f_F(t)},$$
$$H(f,t) = \frac{\partial^2}{\partial f^2}\left[f(1-f)\xi_i(f|f_F(0),t)\right].$$

Therefore, changing f for x inside the derivative and $\tilde{t}$ for u in the second equality to clarify the notation:

$$\begin{aligned}\xi_{i+1}(f|f_F(0),t) &= \int_\Omega \frac{\partial^2}{\partial f^2}\left[f(1-f)\xi_i(f|f_F(0),t)\right]\theta\left(t-\tilde{t}\right)\delta\left(\tilde{f}-\frac{ff_F(\tilde{t})}{f_F(t)}\right)\frac{f_F(\tilde{t})}{f_F(t)}d\tilde{f}d\tilde{t}\\ &= \int_0^t\left[\frac{\partial^2}{\partial x^2}\left[x(1-x)\xi_i(x|f_F(0),t)\right]\right]_{x=\frac{ff_F(u)}{f_F(t)}}\frac{f_F(u)}{f_F(t)}du.\end{aligned}$$

The 0th order is the stationary solution when the initial frequency of $\tilde{f}$ is $\frac{ff_F(0)}{f_F(t)}$ because in that case $H(f,t)=0$. Hence,

$$\xi_0(f|f_F(0),t) = \xi_N\left(\frac{ff_F(0)}{f_F(t)}\middle| f_F(0)\right)\frac{f_F(0)}{f_F(t)}.$$

□

3.3 Exclusive Case

As the Figure 3.3 reflects, in this case there are no sequences with the two mutations.

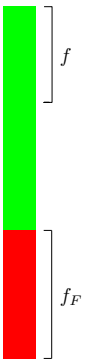

Figure 3.3: Explanation of the exclusive case.

Proposition 3.3.1. *In the case of exclusive mutations, the differential equation that we have to solve is*

$$\frac{\partial \xi(f|f_F(0),t)}{\partial t} = \frac{1}{2sN_e}\frac{\partial^2}{\partial f^2}\left[f(1-f)\xi(f|f_F(0),t)\right] + \frac{\partial}{\partial f}\left[f f_F \xi(f|f_F(0),t)\right].$$

Proof. As in the previous case, what we only have to do is transforming the expected value of the equation 3.1.2 into an explicit formula. So, we apply the Proposition 1.2.5 with $K = 3$, $x_1 = f_F$, $x_2 = 1 - f - f_F$ and $x_3 = f$. Since $s_1 = s$ and $s_2 = s_3 = 0$:

$$E\left[f(\Delta t)|\mathcal{F}_t\right] = E\left[x_3(\Delta t)|\mathcal{F}_t\right] = x_3(-s)x_1 = -s f f_F.$$

□

As above, we expand the solution in powers of $1/S = 1/2N_e s$:

$$\xi(f|f_F(0),t) = \sum_{n=0}^{\infty}\left(\frac{1}{2N_e s}\right)^n \xi_n(f|f_F(0),t).$$

Assuming that $1 \gg s \gg 1/N_e > 0$ we finally find the following equations:

$$\begin{aligned}
\frac{\partial \xi_0(f|f_F(0),t)}{\partial t} &= \frac{\partial}{\partial f}\left[f f_F \xi_0(f|f_F(0),t)\right], \\
\frac{\partial \xi_{i+1}(f|f_F(0),t)}{\partial t} &= \frac{\partial^2}{\partial f^2}\left[f(1-f)\xi_i(f|f_F(0),t)\right] + \frac{\partial}{\partial f}\left[f f_F \xi_{i+1}(f|f_F(0),t)\right].
\end{aligned} \tag{3.3.1}$$

The Green function of the system will be used to solve both partial differential equations.

Theorem 3.3.2. *In the exclusive case the Green function for $s = 0$ is*

$$G(f,t|\tilde{f},\tilde{t}) = \theta\left(t-\tilde{t}\right)\delta\left(\tilde{f} - \frac{f(1-f_F(\tilde{t}))}{1-f_F(t)}\right)\frac{1-f_F(\tilde{t})}{1-f_F(t)}.$$

Proof. Consider $\tilde{f}$ the initial frequency of f and $\tilde{t}$ the initial time, then the Green function is the function G such that satisfies:

$$\frac{\partial G(f,t|\tilde{f},\tilde{t})}{\partial t} = \delta(f-\tilde{f})\delta(t-\tilde{t}) + \frac{\partial}{\partial f}\left[f f_F G(f,t|\tilde{f},\tilde{t})\right] \Rightarrow$$
$$\frac{\partial G}{\partial t} - f f_F \frac{\partial G}{\partial f} = \delta(f-\tilde{f})\delta(t-\tilde{t}) + f_F G.$$

with initial conditions $G = 0$ at $t \to \infty$.

As in the nested case, the solution is 0 for $t < \tilde{t}$, while the term $\delta(f-\tilde{f})\delta(t-\tilde{t})$ generates the initial condition $\delta(f-\tilde{f})$ for the equation when $t > \tilde{t}$:

$$\frac{\partial G_2}{\partial t} - f f_F \frac{\partial G_2}{\partial f} = f_F G_2.$$

Hence, we apply the Characteristics method to the previous equation:

$$\begin{aligned}
\frac{\partial t}{\partial s} &= 1 \Rightarrow t = s + \tilde{t}, \\
\frac{\partial f}{\partial s} &= -f_F f \Rightarrow \\
f(s) &= f(0) e^{\int_0^s -f_F(x)dx} = f(0) e^{-\log(1+(e^x-1)f_F(0))]_0^s} \Rightarrow \\
f(t) &= f(\tilde{t}) \frac{1}{1+(e^t-1)f_F(\tilde{t})} = f(\tilde{t}) \frac{e^{-t}}{f_F(\tilde{t}) + (1-f_F(\tilde{t}))e^{-t}} \\
&= f(\tilde{t}) \frac{f_F(t)}{f_F(\tilde{t})} \frac{f_F(\tilde{t})(1-f_F(t))}{f_F(t)(1-f_F(\tilde{t}))} = f(\tilde{t}) \frac{1-f_F(t)}{1-f_F(\tilde{t})}, \\
\frac{\partial G_2}{\partial s} &= f_F G_2 \Rightarrow G_2(f(t),t) = \ldots = G_2(f(\tilde{t}),\tilde{t}) \frac{1-f_F(\tilde{t})}{1-f_F(t)}.
\end{aligned}$$

In the previous equalities we have used the same tricks than in the proof of the Theorem 3.2.2.

Now, using the Lemma 3.2.3 with

$$\varphi(t) = \frac{1-f_F(t)}{1-f_F(\tilde{t})}, \qquad \phi(t) = 0, \qquad \psi(t) = \frac{1-f_F(\tilde{t})}{1-f_F(t)}.$$

we found the Green function for the exclusive case.

□

Corollary 3.3.3. *Via the Green function, the 0th order is the following:*

$$\xi_0(f|f_F(0),t) = \xi_N\left(\frac{f(1-f_F(0))}{1-f_F(t)}\middle| f_F(0)\right)\frac{1-f_F(0)}{1-f_F(t)}.$$

The higher orders can be solved recursively as follows:

$$\xi_{i+1}(f|f_F(0),t) = \int_0^t \left[\frac{\partial^2}{\partial x^2}\left[x(1-x)\xi_i(x|f_F(0),t)\right]\right]_{x=\frac{f(1-f_F(u))}{1-f_F(t)}} \frac{1-f_F(u)}{1-f_F(t)} du.$$

Proof. The proof of that Corollary is the same than the one for Corollary 3.2.5 despite the fact that in that case we use the Theorem 3.3.2 and Equation 3.3.1. □

3.4 Containing Case

As the Figure 3.4 reflects, now the mutation with frequency f is present in all sequences containing the selected mutation.

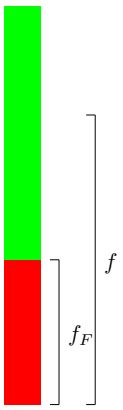

Figure 3.4: Explanation of the containing case.

Proposition 3.4.1. *In containing case, the differential equation that we have to solve is*

$$\frac{\partial \xi(f|f_F(0),t)}{\partial t} = \frac{1}{2sN_e}\frac{\partial^2}{\partial f^2}\left[f(1-f)\xi\right] - \frac{\partial}{\partial f}\left[f_F(1-f)\xi(f|f_F(0),t)\right].$$

Proof. As in the previous cases, what we only have to do is transforming the expected value of the equation 3.1.2 into an explicit formula. So, we apply the Proposition 1.2.5 with $K = 3$, $x_1 = f_F$, $x_2 = f - f_F$ and $x_3 = 1 - f$. Since $s_1 = s$ and $s_2 = s_3 = 0$:

$$\begin{aligned}[f(\Delta t)|\mathcal{F}_t] = E\left[x_1(\Delta t)|\mathcal{F}_t\right] + E\left[x_2(\Delta t)|\mathcal{F}_t\right] = x_1(sx_2 + x_3) - \\ - sx_1x_2 = sx_1(x_2 + x_3 - x_2) = sf_F(1-f).\end{aligned}$$

□

As above, we expand the solution in powers of $1/S = 1/2N_es$:

$$\xi(f|f_F(0),t) = \sum_{n=0}^{\infty}\left(\frac{1}{2N_es}\right)^n \xi_n(f|f_F(0),t).$$

Supposing that $1 \gg s \gg 1/N_e > 0$ we finally found the following equations:

$$\begin{aligned}\frac{\partial \xi_0(f|f_F(0),t)}{\partial t} &= \frac{\partial}{\partial f}\left[(1-f)f_F\xi_0(f|f_F(0),t)\right], \\ \frac{\partial \xi_{i+1}(f|f_F(0),t)}{\partial t} &= \frac{\partial^2}{\partial f^2}\left[f(1-f)\xi_i(f|f_F(0),t)\right] - \frac{\partial}{\partial f}\left[(1-f)f_F\xi_{i+1}(f|f_F(0),t)\right].\end{aligned} \tag{3.4.1}$$

Green function of the system is going to solve both partial differential equations.

Theorem 3.4.2. *In the containing case the Green function for $s = 0$ is*

$$G(f,t|\tilde{f},\tilde{t}) = \theta\left(t-\tilde{t}\right)\delta\left(\tilde{f} - \frac{f(1-f_F(\tilde{t})) - f_F(t) + f_F(\tilde{t})}{1-f_F(t)}\right)\frac{1-f_F(\tilde{t})}{1-f_F(t)}.$$

Proof. Consider $\tilde{f}$ the initial frequency of f and $\tilde{t}$ the initial time, then the Green function is the function G such that satisfies:

$$\begin{aligned}
&\frac{\partial G(f,t|\tilde{f},\tilde{t})}{\partial t} = \delta(f-\tilde{f})\delta(t-\tilde{t}) - \frac{\partial}{\partial f}\left[(1-f)f_F G(f,t|\tilde{f},\tilde{t})\right] \Rightarrow \\
&\frac{\partial G}{\partial t} + (1-f)f_F\frac{\partial G}{\partial f} = \delta(f-\tilde{f})\delta(t-\tilde{t}) + f_F G.
\end{aligned}$$

As in the previous case, the solution is 0 for $t < \tilde{t}$, while the term $\delta(f-\tilde{f})\delta(t-\tilde{t})$ generates the initial condition $\delta(f-\tilde{f})$ for the equation when $t > \tilde{t}$:

$$\frac{\partial G_2}{\partial t} + f(1-f_F)\frac{\partial G_2}{\partial f} = f_F G_2.$$

Hence, we apply the Characteristics method to the previous equation:

$$\begin{aligned}
&\frac{\partial t}{\partial s} = 1 \Rightarrow t = s + \tilde{t}, \\
&\frac{\partial f}{\partial s} = (1-f)f_F = (1-f)\frac{e^t}{e^t + \frac{1-f_F(\tilde{t})}{f_F(\tilde{t})}} \Rightarrow \\
&f(t) = \frac{f(\tilde{t}) + (e^t-1)f_F(\tilde{t})}{1+(e^t-1)f_F(\tilde{t})} = \frac{f(\tilde{t}) + \left(\frac{f_F(t)(1-f_F(\tilde{t}))}{f_F(\tilde{t})(1-f_F(t))} - 1\right)f_F(\tilde{t})}{1 + \left(\frac{f_F(t)(1-f_F(\tilde{t}))}{f_F(\tilde{t})(1-f_F(t))} - 1\right)f_F(\tilde{t})} \\
&\quad = \frac{f(\tilde{t})(1-f_F(t)) + f_F(t)(1-f_F(\tilde{t})) - (1-f_F(t))f_F(\tilde{t})}{(1-f_F(t)) + f_F(t)(1-f_F(\tilde{t})) - f_F(\tilde{t})(1-f_F(t))} \\
&\quad = \frac{f(\tilde{t})(1-f_F(t)) + f_F(t) - f_F(\tilde{t})}{1-f_F(\tilde{t})}, \\
&\frac{\partial G_2}{\partial s} = f_F G_2 \Rightarrow G_2(f(t),t) = \ldots = G_2(f(\tilde{t}),\tilde{t})\frac{1-f_F(\tilde{t})}{1-f_F(t)}.
\end{aligned}$$

In the previous equalities we have used the same tricks than in the proof of the Theorem 3.2.2. In addition for this case, we recommend to use a derivative software in order to check the correctness of the formula for $f(x)$ because there are lots of calculus.

Now, using the Lemma 3.2.3 with:

$$\varphi(t) = \frac{1-f_F(t)}{1-f_F(\tilde{t})}, \qquad \phi(t) = \frac{f_F(t)-f_F(\tilde{t})}{1-f_F(\tilde{t})}, \qquad \psi(t) = \frac{1-f_F(\tilde{t})}{1-f_F(t)}.$$

we found the Green function for the containing case:

$$G(f,t|\tilde{f},\tilde{t}) = \theta\left(t-\tilde{t}\right)\delta\left(\tilde{f} - \frac{f(1-f_F(\tilde{t})) - f_F(t) + f_F(\tilde{t})}{1-f_F(t)}\right)\frac{1-f_F(\tilde{t})}{1-f_F(t)}.$$

□

Corollary 3.4.3. *Via the Green function, the 0th order is the following:*

$$\xi_0(f|f_F(0),t) = \xi_N\left(\frac{f(1-f_F(0)) - f_F(t) + f_F(0)}{1-f_F(t)}\middle| f_F(0)\right)\frac{1-f_F(0)}{1-f_F(t)}.$$

The higher orders can be solved recursively as follows:

$$\xi_{i+1}(f|f_F(0),t) = \\ = \int_0^t \left[\frac{\partial^2}{\partial x^2}\left[x(1-x)\xi_i(x|f_F(0),t)\right]\right]_{x=\frac{f(1-f_F(u))-f_F(t)+f_F(u)}{1-f_F(t)}} \frac{1-f_F(u)}{1-f_F(t)}du.$$

Proof. The proof of that Corollary is the same than the one for Corollary 3.2.5 despite the fact that in that case we use the Theorem 3.2.2 and Equation 3.4.1.

□

Conclusions

In this work we have studied the stationary solutions of a modified Wright-Fisher model.

In the case of two mutations, we proved that pairs of nested and exclusive neutral mutations have the following frequency spectrum, i.e. stationary distribution of frequencies:

$$\xi^{(n)}(f|f_0) = \frac{f_0}{(1-f)^2}\left(1+\frac{1}{f}+\frac{2\ln(f)}{1-f}\right)$$
$$\xi^{(e)}(f|f_0) = \left[\frac{1}{f}-\frac{f_0}{(1-f)^2}\left(1+\frac{1}{f}+\frac{2\ln(f)}{1-f}\right)-\frac{f_0}{(1-f_0)^2}\left(1+\frac{1}{f_0}+\frac{2\ln(f_0)}{1-f_0}\right)\right]$$

for the nested and exclusive case respectively, and we characterized the equations for a general solution with arbitrary selection.

These result are the first steps towards a full description of the effects of the evolution of a mutation on a nearby mutation. They can be used to elucidate the structure of linkage disequilibrium (non-independence between close mutations), to model important biological phenomena like background selection (selection on nearby deleterious mutations) and pervasive hitchhiking and genetic draft (selection on nearby beneficial mutations) and improve statistical inference from DNA sequence data. An important but difficult extension of this work would be the inclusion of recombination, i.e. the exchange of genetic material between different sequences during sexual reproduction.

We also characterized the distribution of frequencies following an environmental change and the related change in selective pressure. Assuming that after this change, the mutation under selection would evolve according to the deterministic equation

$$f_F(t) = \frac{f_F(0)}{f_F(0)+(1-f_F(0))e^{-t}},$$

the evolution of the second mutation follows the power series

$$\xi(f|f_F(0),t) = \sum_{i=0}^{\infty}\left(\frac{1}{2N_e s}\right)^i \xi_i(f|f_F(0),t),$$

We found the solution of the Kolmogorov equation for the coefficients of this expan-

sion:

$$\xi_0^{(n)}(f|f_F(0),t) = \xi_N\left(\frac{ff_F(0)}{f_F(t)}\middle| f_F(0)\right)\frac{f_F(0)}{f_F(t)},$$

$$\xi_{i+1}^{(n)}(f|f_F(0),t) = \int_0^t \left[\frac{\partial^2}{\partial x^2}\left[x(1-x)\xi_i(x|f_F(0),t)\right]\right]_{x=\frac{ff_F(u)}{f_F(t)}} \frac{f_F(u)}{f_F(t)}du;$$

$$\xi_0^{(e)}(f|f_F(0),t) = \xi_N\left(\frac{f(1-f_F(0))}{1-f_F(t)}\middle| f_F(0)\right)\frac{1-f_F(0)}{1-f_F(t)},$$

$$\xi_{i+1}^{(e)}(f|f_F(0),t) = \int_0^t \left[\frac{\partial^2}{\partial x^2}\left[x(1-x)\xi_i(x|f_F(0),t)\right]\right]_{x=\frac{f(1-f_F(u))}{1-f_F(t)}} \frac{1-f_F(u)}{1-f_F(t)}du;$$

$$\xi_0^{(c)}(f|f_F(0),t) = \xi_N\left(\frac{f(1-f_F(0))-f_F(t)+f_F(0)}{1-f_F(t)}\middle| f_F(0)\right)\frac{1-f_F(0)}{1-f_F(t)},$$

$$\xi_{i+1}^{(c)}(f|f_F(0),t) =$$

$$= \int_0^t \left[\frac{\partial^2}{\partial x^2}\left[x(1-x)\xi_i(x|f_F(0),t)\right]\right]_{x=\frac{f(1-f_F(u))-f_F(t)+f_F(u)}{1-f_F(t)}} \frac{1-f_F(u)}{1-f_F(t)}du$$

respectively for the nested, exclusive and containing case. This is the first characterization of the frequency spectrum after an environmental change, and is specially interesting because it represent an observable DNA footprint of recent adaptation to changing environments. These results can be used to detect regions of the genome under recent selection based on DNA sequence data. They can also be used to detect the initial and final frequencies of the mutation under selection and therefore distinguish between different evolutionary scenarios (hard selective sweeps, incomplete selective sweeps, soft selective sweeps from standing variation) that differ in the initial and final frequency of the selected mutation in the population. It would be interesting to extend these results to the case with recombination, but similarly to the neutral case, there are difficult problems to be faced.

A general, interesting extension of this work would be the analysis of three or more mutations. This would be useful to characterize the noise of the frequency spectrum and other statistics.

In addition, another option to extend the project is to bound the error of the series expansion of chapter two or compare the results with numeric simulations.

Bibliography

[1] Richard Durrett. *Probability models for DNA sequence evolution*. Probability and its Applications (New York). Springer, New York, second edition, 2008.

[2] Warren J. Ewens. *Mathematical population genetics. I*, volume 27 of *Interdisciplinary Applied Mathematics*. Springer-Verlag, New York, second edition, 2004. Theoretical introduction.

[3] Luca Ferretti, Emanuele Raineri, and Sebastian Ramos-Onsins. Neutrality tests for sequences with missing data. *Genetics*, 191(4):1397–1401, 2012.

[4] R. C Griffiths. A transition density expansion for a multi-allele diffusion model. *Adv. in Appl. Probab.*, 11(2):310–325, 1979.

[5] R. A. Littler and E. D. Fackerell. Transition densities for neutral multi-allele diffusion models. *Biometrics*, 31:117–123, 1975.

[6] Xiaohui Xie. The site-frequency spectrum of linked sites. *Bull. Math. Biol.*, 73(3):459–494, 2011.