Gene dynamics in populations of variable size

Final project for 8.592 - Statistical Physics in Biology

Bernardo Pando[†]

[†] *Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139*

In class, a model for the dynamics of a given gene mutation in a population of fixed size was described and studied. The purpose of this report is to review some efforts in extending that description to populations of variable size.

I. INTRODUCTION

The mathematical theory of evolution viewed as a stochastic process started its development in the early 30s with the seminal works of Fisher [1–3], Haldane [4], Wright [5] and others. From that moment people have postulated different models of evolution and tried to develop mathematical techniques for studying them. This was not only an intellectual challenge but it was, and is, a very important tool for probing evolution theories about the history of living organisms on data gathered in the present. During the last few years, due to an explosion in DNA sequencing and other genotyping technologies a huge amount of data have become available and the field has turned into a very active one. In [6] the state of the art in the field, as well as some of the current trends is reviewed.

Some of the ongoing efforts involve trying to deduce facts about the history of populations based on genetic data [7]. Questions as "When did the human population size start increasing?" [8], whether the human population went through some "bottleneck" at some point in history [9] and several related ones [10–13] are tried to be answered confronting genetic data to models of genes' evolution. It is evident that in order to be able to probe for the history of a given population is imperative to design and understand models that take population size changes into account.

Nowadays, most inferences are made through *backwards in time* simulations of a stochastic process referred to as "the coalescent" [14], initially introduced by Kingman [15] and others, the advantage of this approach being the relatively easy reconstruction of genealogical trees [16–19]. This was not the approach followed during the course; in the class emphasis was put on trying to predict the probability distribution of the frequency of finding some allele of a given gene that was expressed in two different forms in a population, or in trying to compute the fixation time of one of the two forms. Accordingly, in this report I will focus on the same lines of analysis but trying to extend the results that were discussed for populations of constant size to situations in which the number of individuals vary with time.

The rest of this report is structured as follows: in II a model of genetic evolution is introduced along with a numerical scheme for simulating it. In section III the approximations that allow to describe the behavior of that model by some partial differential equations are discussed. The constant size case studied in class is reviewed in section IV. And finally some exploratory results on the varying-size populations are presented in section V.

II. MODEL

Throughout the years several different stochatic models for population genetics have been introduced (many having been reviewed in some textbooks [20–22]). The simplest models try to describe the evolution of a gene that has only two possible forms (referred to as alleles) in a given population. Several evolutionary mechanisms can in principle affect the frequency of one of the two alleles in a given set of individuals; some of the most importants being:

1. *Mating.* In haploid populations the gene that an offspring will carry is selected randomly from the genes that its parents carry. If the individuals were diploids, each of the two copies of the gene carried by a newborn is selected randomly from one of the two copies of its parents.

2. *Mutation.* During the life of an individual, its genes can randomly mutate. In a model with two alleles the only possible mutations are those from one of the forms of the gene to the other and viceversa.

3. *Selective pressure.* It can be tought that individuals are selected based on the genes that they carry. The probability of survival of individuals that carry one of the forms of the gene can be different from the probability of survival of those that do not carry it.

4. *Migration.* Some individuals might decide to leave the population, whereas some foreigners (that is, individiuals that are originally from a population with a different genetic distribution) might decide to join it.

Every one of these effects can be modelled with a different level of detail and that is the main difference between most of the current models. For instance, the sex of individuals can be kept track of, the birth and death can be modelled more or less accurately and so on and so

Mating	Probability	Possible outcomes
AA + AA	x^4	AA
BB + BB	$(1-x)^4$	BB
AA + BB	$2x^2(1-x)^2$	AB
AA + AB	$4x^3(1-x)$	AA ($1/2$), AB ($1/2$)
BB + AB	$4x(1-x)^3$	BB ($1/2$), AB ($1/2$)
AB + AB	$2x^2(1-x)^2$	BB ($1/4$), BB ($1/4$), AB ($1/2$)

TABLE I: Mating possibilities in a diploid population. x represents the fraction of copies of the gene A in the whole population. In the case of multiple outcomes the corresponding propabilities are inidicated in parenthesis.

forth. But following what was done in class I will take the simplest approach, paralleling the lines of what is known as the Wright-Fisher model [5, 20, 22]. The main assumptions of this model are: non-overlapping generations, "random mating", no migration and Poissonian mutation/selection. In this model, sex of the individuals is not taken into account and, for large enough diploid populations there's no need to keep track of the details in the distribution of heterozygotes. Therefore, the state of the population at any given time is described only by two numbers: the total number of genes (which we will denote by N [29]) and the amount of one of the two forms of the gene (we will denote this number by n and will refer to that form of the gene as 'A', which is going to be opposed to 'B').

In every generation, during the mating stage, genes are randomly drawn from the current population (non-overlapping generations assumption) according to the mating principles and the fate of the newborns is decided using some probabilities that characterize the selection pressure. So, for instance, if we are dealing with a population of diploid individuals, we can assign the following probabilities of birth based on the mating table I:

$$\begin{aligned} P_{\mathrm{b}}(\mathtt{AA}) &= x^2 \\ P_{\mathrm{b}}(\mathtt{AB}) &= 2x(1-x) \\ P_{\mathrm{b}}(\mathtt{BB}) &= (1-x)^2, \end{aligned} \tag{1}$$

where we have defined $x \equiv n/N$, the fraction of genes of type A in the whole population. If we furthermore assume a set $\{w_{\mathtt{AA}}, w_{\mathtt{AB}}, w_{\mathtt{BB}}\}$ of probabilities of survival given a birth, we can express the probabilities of obtaining a mature individidual with a given genotype as

$$P_{\mathrm{m}}(\cdot) = w_{(\cdot)} P_{\mathrm{b}}(\cdot), \tag{2}$$

being these normalized by the condition $1 = P_{\mathrm{m}}(\mathtt{AA}) + P_{\mathrm{m}}(\mathtt{AB}) + P_{\mathrm{m}}(\mathtt{BB})$.

In the present case, in which we are not interested in keeping track of the individual characteristics, but only on the gene count, we need to compute the probability of obtaining a 'mature' A. As we are assuming that the frequency of this gene is enough for describing the fate of the population it is valid to think that 'mature' genes are picked at random from the two genes of a mature individual. This reasoning leads us to

$$P_{\mathrm{m}}(\mathtt{A}) = P_{\mathrm{m}}(\mathtt{AA}) + \tfrac{1}{2} P_{\mathrm{m}}(\mathtt{AB}). \tag{3}$$

Given that the probabilities of obtaining a mature subject with a given genotype can be computed out of the current state of the population, the number n' of sexually active As in the next generation (assumed to be of size N') can then be simply drawn from a binomial distribution [30]:

$$n' = \mathrm{Bin}\left[N', P_{\mathrm{m}}(\mathtt{A})(n, N)\right]. \tag{4}$$

In the simplest description the effect of mutations can be described by two numbers, $\mu_{\mathtt{AB}}$ and $\mu_{\mathtt{BA}}$, that represent the probabilities of a gene to transmutate into the other form. So, after the mating step we can think that the number of copies of A is updated to n'' according to

$$n'' = \mathrm{Bin}\left[n', 1-\mu_{\mathtt{AB}}\right] + \mathrm{Bin}\left[N'-n', \mu_{\mathtt{BA}}\right]. \tag{5}$$

The meaning of this equation is clear: the number of As after mutation is equal to the number of copies of A that did not mutate plus the number of copies of B that got transformed.

So, gathering everything together this model of gene evolution is defined as follows: the state of the system is described by n_t, the number of copies of A in the tth generation and its evolution is defined by the following equations:

$$\begin{aligned} x &= \frac{n_t}{N_t} \\ W &= w_{\mathtt{AA}} x^2 + w_{\mathtt{BB}}(1-x)^2 + 2w_{\mathtt{AB}} x(1-x) \\ p &= x + \frac{x(1-x)}{2} \frac{\mathrm{d}}{\mathrm{d}x} \left[\ln W(x)\right] \\ n' &= \mathrm{Bin}\left[N_{t+1}, p\right] \\ n_{t+1} &= \mathrm{Bin}\left[n', 1-\mu_{\mathtt{AB}}\right] + \mathrm{Bin}\left[N_{t+1}-n', \mu_{\mathtt{BA}}\right], \end{aligned} \tag{6}$$

where p is what was previously referred to as $P_{\mathrm{M}}(\mathtt{A})(n, N)$, and the reported expression is just an useful way of writing what is obtained by following the replacements indicated in equations (1–3) along with the normalization condition for the P_{m} probabilities.

This way of describing the model provides us with a full systematic specification of the system at a "microscopic" scale (rules of evolution per generation and bookeeping of all the individuals), useful for testing and trying to understand the validity of simplifications that might be introduced later on. But, furthermore, the process so described turns out to be very easy to simulate numerically as the only nontrivial part involves sampling numbers from binomial distributions, feature that is included in most commercial packages, or could be implemented in a fairly simple manner (see for example [24]).

It can also be seen that the fact that the population might change size was included in the model without any effort. But, it is fair to say that this approach is only valid

for situations in which the dynamics of the population size in decoupled from the dynamics of the frequency of the gene under observation. Only a before-hand known N_t allows the model specification to be complete. Therefore, this kind of model might be useful for describing dynamics of neutral genes, *i.e.* genes that are not detrimental for the population size. This doesn't mean that selection can not be taken into account (as a matter of fact, it is included in the model!) but that the selection pressure is supposed to be small: the model works under the assumption that the tth generation has the means for procreating all N_{t+1} mature individuals of the next wave, but this can only be true if a significant portion of the births reach maturity (as the amount of time and resources that a generation has for procreating the next one is limited).

III. DIFFUSION APPROXIMATIONS

The stochastic model (6) is difficult to analyze as it is and some approximations are called for in order to try to gain insight into its inner workings. In 1964 Kimura [25] introduced the idea of using diffusion approximations for describing the behavior of models of this sort by means of partial differential equations. The approximations that underlie this approach are to consider that the popolation size is large enough so that the frequency of the form `A` can be regarded as a continuous variable and that we are only interested in describing the behaviour in a timescale much longer than a generation life cycle, so that the generation number can be also considered continous. It is worth saying that under these coarse-graining assumptions many a priori different models exhibit similar behaviors the difference lying only in a different interpretation of the parameters. So, for instance, some models that keep track of the sex of the individuals might be reduced to a model that doesn't take sex into account by a simple rescaling of the population size [21].

Under these conditions the model is a stochastic system with a continuum variable in the range $[0, 1]$ and its behaviour can be described in terms of Chapman-Kolmogorov or Fokker-Planck equations [26]. If we call $p(x, t|y, t')$ the probability density of seeing the gene with frequency x in the population at time t given that it was y at time t', then the following two equations rule the evolution of this quantity [26]:

$$\begin{cases} \dfrac{\partial p}{\partial t} = -\dfrac{\partial}{\partial x}\left[M(x,t)p\right] + \dfrac{\partial^2}{\partial x^2}\left[\dfrac{V(x,t)}{2}p\right] \\ \dfrac{\partial p}{\partial t'} = M(y,t')\dfrac{\partial p}{\partial y} + \dfrac{V(y,t')}{2}\dfrac{\partial^2 p}{\partial y^2}. \end{cases} \tag{7}$$

In this equations, $M(x, t)$ is the mean of the change in x during the infinitesimal interval $(t, t + \delta t)$, and $V(x, t)$ is the variance of such a change.

Given the coarse-graining assumption we can compute such quantities taking as the "infinitesimal" interval one generation. From (6) the mean value of the change in x (the fraction of genes of the form `A`) throughout one generation assuming a large enough population can be approximately computed as

$$M(x,t) \equiv \langle \Delta x \rangle \approx \langle \Delta x_{\mathrm{ms}} \rangle + \langle \Delta x_{\mathrm{m}} \rangle \tag{8}$$

where $\langle \Delta x_{\mathrm{ms}} \rangle$ is the average change due to the mating/selection process and $\langle \Delta x_{\mathrm{m}} \rangle$ is the mean change due to mutations. Using the model specification it is straightforward to compute these two quantities:

$$\begin{aligned} \langle \Delta x_{\mathrm{ms}} \rangle &= \langle x_{\mathrm{ms},t+1} \rangle - x_t = \frac{\langle n' \rangle}{N_{t+1}} - x \\ &= \frac{x(x-1)}{2}\frac{\mathrm{d}}{\mathrm{d}x}\left[\ln W(x)\right] \\ \langle \Delta x_{\mathrm{m}} \rangle &= x(1-\bar{\mu}_{\mathtt{AB}}) + (1-x)\bar{\mu}_{\mathtt{BA}} - x \\ &= \bar{\mu}_{\mathtt{BA}} - (\bar{\mu}_{\mathtt{AB}} + \bar{\mu}_{\mathtt{BA}})x. \end{aligned} \tag{9}$$

As for the variances, again under the assumption of a large enough population we can approximate

$$V(x,t) \equiv \langle (\Delta x)^2 \rangle_c \approx \langle (\Delta x_{\mathrm{ms}})^2 \rangle_c + \langle (\Delta x_{\mathrm{m}})^2 \rangle_c. \tag{10}$$

But before computing this quantity let's note that the model makes sense only in the case of high survival probabilities (as one generation has only a limited number of time and resources to breed the mature individuals of the next wave) and that usually the probabilities of mutation are fairly low. Therefore we can assume that the probabilities of survival take the form $w_{(\cdot)} = 1 - \epsilon\bar{w}_{(\cdot)}$ whereas the mutation probabilities are written as $\mu_{(\cdot)} = \epsilon\bar{\mu}_{(\cdot)}$ with ϵ a small number. Computing the variances using this approach we get

$$V(x,t) \approx \frac{x(1-x) + O(\epsilon)}{N(t)} \tag{11}$$

where the fact that the population size might be changing throughout the generations first enter the picture.

So, disregarding terms of order ϵ in the variance (which might turn out to be important for the description to be accurate near the boundaries) the forward and backwards evolution equations for $p(x, t|y, t')$ are

$$\begin{cases} \dfrac{\partial p}{\partial t} = -\dfrac{\partial}{\partial x}\left[M(x)p\right] + \dfrac{\partial^2}{\partial x^2}\left[\dfrac{x(1-x)}{2N(t)}p\right] \\ \dfrac{\partial p}{\partial t'} = M(y)\dfrac{\partial p}{\partial y} + \dfrac{x(1-x)}{2N(t')}\dfrac{\partial^2 p}{\partial y^2}. \end{cases} \tag{12}$$

with $M(x)$ given by

$$M(x) = \frac{x(1-x)}{2}\frac{\mathrm{d}}{\mathrm{d}x}\left[\ln W(x)\right] + \mu_{\mathtt{BA}}(1-x) - \mu_{\mathtt{AB}}x. \tag{13}$$

These equations provide an accurate description of the local evolution of the distribution of frequencies away from the boundaries located at $x = 0$ and $x = 1$. At those points, some boundary conditions need to be specified in

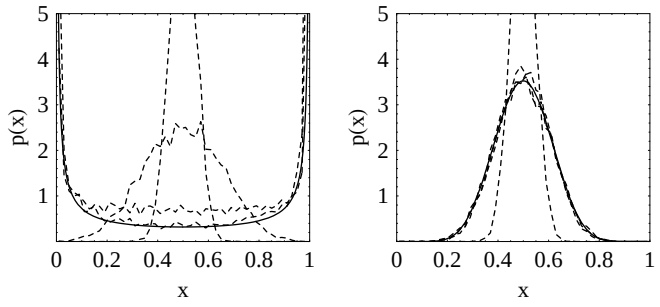

FIG. 1: Distribution of frequencies of a neutral allele in populations of fixed size. In each figure the stationary distribution (15) is plotted (solid line) along with the results of Monte Carlo simulations (dashed curves). In both cases the size of the population was taken to be $N = 1000$; in the left panel the mutation probabilities were set to $\mu_{\mathtt{AB}} = \mu_{\mathtt{BA}} = 0.0001$ whereas in the right panel the corresponding value is 0.005. Simulations were started from a population in which the frequency of the gene was 1/2 and 10000 realizations of the process were drawn. The different dashed lines in each panel correspond to histograms (bin size = 0.02) constructed from the simulation results obtained after 10, 100, 1000 and 10000 generations (from the curves most concentrated around 1/2 to the broadest ones).

order for the system to be well posed. Nevertheless we know that the magnitude under study represents a probability distribution and as such it should be normalized, so that at all times we must have

$$\int_0^1 p(x,t|y,t')\,\mathrm{d}x = 1. \tag{14}$$

For some computations this condition turns out to be enough and boundary specifications need not be taken into account. Specific examples are the cases of the stationary distribution in the case of a constant size population discussed in the next section and the explicit solution found by Kimura of the process with neither mutation nor selection [27].

IV. CONSTANT SIZE

The case of constant size was studied in class. It was seen that the probability distribution reached a steady state (obtained by setting $\frac{\partial p}{\partial t} = 0$ in (12) and considering zero probability flux in equilibrium) that could be expressed in the form

$$p_{\mathrm{ss}}(x) \propto [W(x)]^N x^{2N\mu_{\mathtt{BA}}-1}(1-x)^{2N\mu_{\mathtt{AB}}-1}. \tag{15}$$

It was also shown that for the case of equal mutation probabilities ($\mu_{\mathtt{AB}} = \mu_{\mathtt{BA}} \equiv \mu$)and no selection ($W(x) = 1$) the steady state distribution undergoes a transition when $2N\mu = 1$. For $2N\mu > 1$ the mode of distribution is located at $x = 1/2$, corresponding to a situation in which the gene is guaranteed to be present in future generations in its two possible forms; on the other hand, for $2N\mu < 1$ the distribution becomes singular (though yet normalizable) at $x = 0$ and $x = 1$ meaning that in this case there is a finite probability for one of the forms of the gene to take over the whole population. This analytical result was used to test the behavior of the implemented Monte Carlo simulations to make sure that everything was working properly. Some results can be seen in figure 1. The incorporation of selective pressure will just deform these general results: for large enough populations the distribution will be monomodal (with the mode very close to some boundary if the selective pressure is high and the mutation probabilites are low) and for sizes below a critical value there will be finite probabilities for the alleles to spread over all the individuals.

V. MONOTONICALLY VARYING POPULATIONS

If the size of the population is changing with time there is not guarantee that a stationary distribution will be ever achieved and if it some steady state might be reached there's no guarantee that it will be unique (except perhaps in those cases in which the population size reaches some limiting value). The critical size of the population that separates the two behaviors discussed in the previous section will play a dominant role. In a increasing population the characteristics of the frequency distribution that reaches the critical value will be emphazized from that point on as the variance will be dimishing forever. Forever decreasing populations can not be fully analyzed using the current scope as it is only valid when the number of individuals is fairly large, but intuitevily, if the critical point is reached when the size of the population is large enough, then we can say that one of the two forms of the gene will take over the population. In situations in which the size is oscillating (which might not be an unrealistic situation for some species that compete with others, as proposed in models of the Lotka-Volterra type [28]) the behaviour is not easy to predict based on these grounds, but the two main alternatives are either reaching a more or less static distribution (for example a monomodal distribution with small periodic fluctations in mean and or variance) or behaving in an oscillatory manner, changing periodically from a situation in which the genes tend to fixate to another scheme in which they tend to coexist.

Using the simulation tool developed, here are some exploratory results of some of these possibilities.

- NOT COMPLETE YET!-

[1] R.A. Fisher, *Proc. Roy. Soc. Edin.* **42**, 321 (1922) (cited in [25]).
[2] R.A. Fisher, *Proc. Roy. Soc. Edin.* **50**, 205 (1930) (cited in [25]).
[3] R.A. Fisher, *The Genetical Theory of Natural Selection*, 2nd ed. (Dover, New York, 1958) (original edition dated in 1930).
[4] J.B.S. Haldane, *Trans. Camb. Phil. Soc.* **23**, 19 (1924) (cited in [25]).
[5] S. Wright, *Genetics* **16**, 97 (1931).
[6] J. Wakeley, *Journal of Heredity* **95**, 397 (2004).
[7] G. Weiss and A. von Haeseler, *Genetics* **149**, 1539 (1998).
[8] J.D. Wall and M. Przeworski, *Genetics* **155**, 1865 (2000).
[9] J. Hawks, K. Hunley, S-H. Lee and M. Wolpoff, *Mol. Biol. Evol.* **17**, 2 (2000).
[10] A.R. Rogers and H. Harpending, *Mol. Biol. Evol.* **9**, 552 (1992).
[11] R. Foley, *Genome Research* **8**, 339 (1998).
[12] A.R. Rogers, *Proc. Natl. Acad. Sci. USA* **98**, 779 (2001).
[13] J.F. Storz, M.A. Beaumont and S.C. Alberts, *Mol. Biol. Evol.* **19**, 1981 (2002).
[14] D.J. Balding, M.J. Bishop, C. Cannings (editors), *Handbook of statistical genetics* (John Wiley & Sons, Chichester, 2003).
[15] J.F.C. Kingman, *Stochastic Process Appl.* **13**, 235 (1982) (cited in [6]).
[16] S. Tavaré, *Theor. Pop. Biol.* **26**, 119 (1984).
[17] S.M. Krone and C. Neuhauser, *Theor. Pop. Biol.* **51**, 210 (1997).
[18] B. Rannala, *Heredity* **78**, 417 (1997).
[19] A. Sano and H. Tachida, *Genetics* **169**, 1687 (2005).
[20] D.L. Hartl, *A Primer of Population Genetics*, 2nd ed. (Sinauer Associates, Sunderland, 1987).
[21] M. Kimura and T. Otha, *Theoretical Aspects of Population Genetics* (Princeton University Press, Princeton, 1971).
[22] D.L. Hartl and A.G. Clark, *Principles of Population Genetics*, 3rd ed. (Sinauer Associates, Sunderland, 1997).
[23] S. Lessard (editor), *Mathematical and Statistical Developments of Evolutionary Theory* (Kluver Academic Publishers, Dordrecht, 1987).
[24] W.H. Press, S.A. Teukolsky, W.T. Vetterling and B.P. Flannery, *Numerical Recipes in C* (Cambridge University Press, Cambridge, 1992).
[25] M. Kimura, *J. Appl. Prob.* **1**, 177 (1964).
[26] C.W. Gardiner, *Handbook of Stochastic Methods for physics, chemistry, and the natural sciences* (Springer-Verlag, Berlin, 1985).
[27] M. Kimura, *Proc. Natl. Acad. Sci. USA* **41**, 144 (1955).
[28] S.H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering* (Perseus Books Group, 2001).

[29] For haploid populations N is equal to the number of individuals whereas for diploids it is twice that number.
[30] The notation $n = \mathrm{Bin}\,[N,p]$ denotes that n is a random variable drawn from a binomial distribution with parameters N, p; *i.e.* the probability of obtaining n after randomly drawing it is $p(n) = \binom{N}{n} p^n (1-p)^{N-n}$.