MIT-CTP-3525

On the Population Genetics

Chris Kouvaris[1]

[1] *Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

(Dated: May 10, 2005)

In this paper we review the basic mechanisms that drive the change in frequencies of populations. In particular we are going to focus on mutations. In a system of diploids with back and forward mutations, the diffusion equation (Kolmogorov) that governs the dynamics of the population should be modified in order to include the effect of the mutations. Mutations not only change the mean value of genes going from one species to the other, but they also change the variance. We include the effect of the non-zero variance of the mutations in the Kolmogorov equation and we solve the equation. We show that in order to disregard the variance of the mutations from the Kolmogorov equation, the probability of the mutation should scale as $1/N$ or at a higher power. We also show that in cases where the mutation rates are high, the probability distribution of the population number follows a Gaussian with mean value and standard deviation of 1/2.

PACS numbers: 12.38.-t, 25.75.Nq

I. INTRODUCTION

Theoretical population genetics is arguably the area of biology in which mathematics has been most succesfully applied. Other areas such theoretical ecology model phenomena which are intrinsically more important to human welfare, and which have a much larger base of observations to work with, but are nevertheless not as succesfully modeled. The major reason why theory is more readily applied to population genetics is that there is a framework - Mendelian segregation - on which to hang it. The Mendelian mechanism is a highly regular process with strong geometric and algebraic overtones.

The other reason why Mendelian segregation is particularly important to population genetics is that it occurs whether or not natural selection is present, whether or not mutation is present, and whether or not migration is present.

In a large random-mating population with no selection, mutation or migration, the gene frequencies and the genotypa frequencies are constant from generation to generation. Furthermore, there is a simple relationship between the gene frequencies and the genotype frequencies. These properties are derived from a theorem, or principle known as the Hardy-weinberg law after Hardy and Weinberg, who independently demostrated it in 1908. A population with constant gene and genotype grequenciesis said to be in Hardy-Weinberg equilibrium. The relationship between gene frequencies and genotype frequencies is of the greatest importance because many of the deductions about population genetics and quantitative genetics rest on it. The relationship is this: if the gene frequencies of two alleles among the parents are p and q, then the genotype grequencies among the progeny are p^2, 2pq and q^2. This relationship refers to autosomal genes; sex-linked genes are not quite so simple. The conditions of random mating and no selection, required for the Hardy-Weinberg law to hold, refer only to the genotypes under consideration. There may be preferential mating with respect to other attributes, and genotypes of other loci may be subject to selection, without affecting the issue. Two additional conditions are that the genes segregate normally in gametogenesis and that the gene frequencies are the same in males and females. The proof of the theorem can be seen in the original papers of Hardy and Weinberg [1, 2].

Although the large random-mating population is stable with respect to gene frequencies and genotype frequencies, if we include agencies that tend to change the genetic properties, then the frequencies can also change. There are two sorts of process: systematic process, which tend to change the frequency in a manner predictable both in amount and in direction; and the dispersive process, which arises in small populations from effects of sampling, and is predictable in amount but not in direction. We are going to concentrate on the systematic processes for the moment. There are three basic processes: migration, selection, and mutation.

The effect of migration is very simply dealt with. In a large population that consists of a proportion x of new immigrants in each generation and $1-x$ natives, and the frequency of a certain gene is p_m for the immigrants and p_0 for the natives, the frequency of the gene in the mixed population p_1 is

$$p_1 = mp_m + (1-m)p_0 \tag{1}$$

The change of gene frequency, Δp, brought about by one generation of immigration is the difference between the frequency before immigration and the frequency after immigration. Therefore

$$\Delta p = p_1 - p_0 \tag{2}$$

In the case of the selection, we must take account of the fact that individuals differ in viability and fertility, and that therefore contribute different numbers of offspring to the next generation. The contribution of offspring to the next generation is called the fitness of the individual, or sometimes the adaptive value, or selective value. If the differences of fitness are in any way associated with the presence or absence of a particular gene in the individual's genotype, the selection operates on that gene. When a gene is subject to selection its frequency in the offspring is not the same as in the parents, since parents of different genotypes pass on their genes unequally to the next generation. In this way selection causes a change of gene frequency and consequently also a genotype frequency. The change of gene frequency resulting from selection is more complicated to describe than that resulting from mutation, because the differences of fitness that give rise to the selection are an aspect of the phenotype. Reviews on this subject can be found in [3–8].

II. MUTATIONS

The third case where we can have a change in the frequencies is mutation. In this section we are going to investigate thouroughly how forward and back mutations change the population of the species. Let's consider a diploid, namely a system that has two species A_1 and A_2. The total number of alleles is $2N$. We can define x to be the fraction of A_1 alleles and $1-x$ the fraction of A_2 alleles. This means that if there are M A_1, $x = M/(2N)$. Furthermore, we assume that there is a probability μ_1 of having a A_2 mutating to an A_1 and a probability μ_2 of having a A_1 mutating to an A_2. We want to calculate the probability to have a change in the frequencies by Δx after one time step. For simplicity we assume that we have x_1 of the x A_1's that convert to A_2's and x_2 from the $y = 1-x$ A_2's converting to A_1's. The probability to have a change in x, $\Delta x = x_2 - x_1$ is

$$P(\Delta x) = P(x_2 - x_1) = \sum_{x_1, x_2} \mu_2^{x_1}(1-\mu_2)^{x-x_1} \begin{pmatrix} x \\ x_1 \end{pmatrix} \mu_1^{x_2}(1-\mu_1)^{y-x_2} \begin{pmatrix} y \\ x_2 \end{pmatrix}$$

where the sum is over all possible values of x_1 and x_2, under the constraint $x_2 - x_1 = \Delta x$. It is very easy to justify this. The probability is the product of two binomials. It is the probability of picking x_1 A_1 from x times the probability of not picking the rest, times the same contribution for x_2 A_2 picked from y A_2. The constraint appears because we are not interested in the probability of having specifically x_1 A_1 converting to A_2 and x_2 A_2 converting to A_1, but we are interested in the probability with the difference $x_2 - x_1$ constant. So we have to sum over all values of x_1 and x_2, under the above constraint. Now, we can calculate the mean value and the variance of Δx, $\langle \Delta x \rangle$ and $\langle \Delta x^2 \rangle_c$ respectively. For $\langle \Delta x \rangle$ we have:

$$\langle \Delta x \rangle = \sum_{x_1, x_2} (x_2 - x_1)\mu_2^{x_1}(1-\mu_2)^{x-x_1} \begin{pmatrix} x \\ x_1 \end{pmatrix} \mu_1^{x_2}(1-\mu_1)^{y-x_2} \begin{pmatrix} y \\ x_2 \end{pmatrix}$$

where the sum is over all values of x_1 and x_2 without the constraint. The final result for the mean value is

$$\langle \Delta x \rangle = (1-x)\mu_1 - x\mu_2. \qquad (3)$$

Similarly we calculate the variance:

$$\langle \Delta x^2 \rangle_c = (1-x)\mu_1(1-\mu_1) + x\mu_2(1-\mu_2). \qquad (4)$$

Now we can present the mathematical formalism that will enable us to calculate the frequency changes.

III. THE KOLMOGOROV EQUATION

In population genetics the fundamental quantity used for describing the genetic composition of a Mendelian population is gene frequency rather than genotype frequency. The main reason for this is that each gene is a self-reproducing entity and its frequency changes almost continuously with time as long as the population is reasonably large. On the other hand, genotypes are produced anew in each generation by recombination of genes and therefore do not have the continuity that genes have.

The way we can formulate mathematically the process of change of gene frequency is to treat the process as stochastic. Roughly speaking, this means that as the time interval becomes smaller, the amount of change in gene frequency during that interval is expected to be smaller. We assume that the process of change in gene frequency is Markovian, that is, the probability distribution of gene frequency at a given time t_1 depends on the gene frequencies at a preceding time t_0 but not on the previous history which led to the gene frequencies at t_0, where $t_0 < t_1$.

The fundamental equations used to study this continuous Markov process are the Kolmogorov forward and backward equations [9]. The forward Kolmogorov equation was first introduced into the field of population genetics by Wright (1945), while the backward equation was first used in this field to study the problem of gene fixation by Kimura [10, 11]. The derivation of the equations can be found in [8]. The forward Kolmogorov equation is

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial (M(x)P(x,t))}{\partial x} + \frac{1}{2}\frac{\partial^2 (V(x)P(x,t))}{\partial x^2}, \quad (5)$$

where x was defined in the previous section and t is the time. The forward Kolmogorov equation is a diffusion equation. It describes the change over time in the distribution of allele frequencies in terms of some measures of the shape of that distribution. The function $M(x)$ represents the mean value in the change of x due to directional forces such as selection, mutation etc, whereas the function $V(x)$ represents the variance due to these processes and due to sampling.

At equilibrium we want to have a steady state, which means that $\partial P(x,t)/\partial t = 0$. In this case the equation can be written as

$$0 = -\frac{\partial (M(x)P(x,t))}{\partial x} + \frac{1}{2}\frac{\partial^2 (V(x)P(x,t))}{\partial x^2}. \quad (6)$$

If we integrate over x once and set the integration constant to zero we have

$$\frac{1}{2}\frac{\partial (V(x)P(x,t))}{\partial x} - M(x)P(x,t) = 0. \quad (7)$$

If we solve the differential equation we get the solution:

$$P(x) = \frac{C}{V(x)} e^{\int \frac{2M(x)}{V(x)} dx}, \quad (8)$$

where C is a constant fixed by normalizing the probability density. It is instructive to see what results we get if we have a system with random mating and processes such as natural selection and mutation. In this particular case the Kolmogorov equation is written as

$$\frac{1}{2}\frac{\partial}{\partial x}(\frac{x(1-x)}{2N}P(x,t)) - (\frac{x(1-x)}{2w}\frac{dw}{dx} + (1-x)\mu_1 - x\mu_2)P(x,t) = 0. \quad (9)$$

The variance $x(1-x)/(2N)$ is the variance due to sampling. The term with the w is the mean value for Δx due to selection and the second term is the mean value for mutations we derived in the previous section. The steady state solution of this equation is

$$P(x) \propto x^{4N\mu_1 - 1}(1-x)^{4N\mu_2 - 1} w^{2N}. \quad (10)$$

The above solution has been analyzed in many papers and books in literature, for example in [3, 8, 10, 11]. In all these treatments the effect of the variance of the mutation has been ignored, meaning that the variance of Δx due to mutations from Eq. 4 has not been included as part of the function $V(x)$ in Eq. 9. The main reason for neglecting the variance of the mutations is that usually it is very small compared the variance of the sampling. However we shall show that there are interesting limits where the above statement is not correct. Another possible reason for neglecting the variance of the mutation is that the time scales for random mating and mutations might be unequal. In this paper we solve the forward Kolmogorov equation including the contribution of the variance of the mutations and we study the solutions. For the purpose of generality we also include an arbitrary constant τ in front of the variance of the mutation in order to account for different time scales between random mating and mutation. We are going to consider a system without natural selection, which means that we set $w = 1$. The forward Kolmogorov equation can be written as

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial}{\partial x}(((1-x)\mu_1 - x\mu_2)P(x,t)) + \frac{1}{2}\frac{\partial^2}{\partial x^2}((\frac{x(1-x)}{2N} + \tau((1-x)\mu_1(1-\mu_1) + x\mu_2(1-\mu_2)))P(x,t)). \quad (11)$$

We are interested again in steady state solutions independent of time. The solution is given by Eq. 8, using $M(x)$ and $V(x)$ specified in the equation above. After

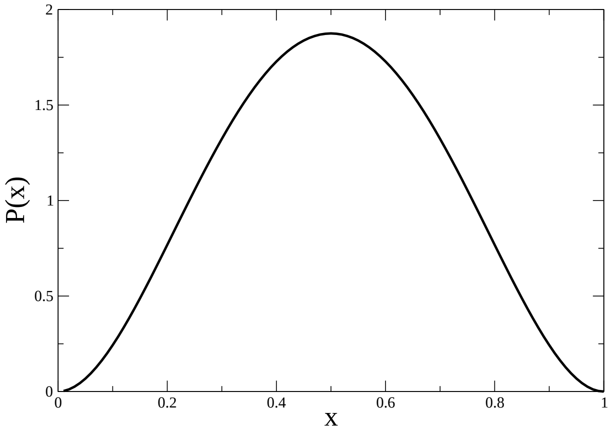

FIG. 1: Probability distribution for the case A with $N\mu = 1/2$

some algebra we can write the solution

$$P(x) \propto (x - r_1)^{-\Lambda(\mu_1 - (\mu_1+\mu_2)r_1)-1}(x-r_2)^{\Lambda(\mu_1-(\mu_1+\mu_2)r_2)-1}. \quad (12)$$

We have introduced the constants Λ, r_1 and r_2:

$$\Lambda = \frac{4N}{\sqrt{(n_1+n_2+1)^2 - 4n_1n_2}}, \quad (13)$$

where $n_1 = 2N\mu_1(1-\mu_1)\tau$ and $n_2 = 2N\mu_2(1-\mu_2)\tau$. The constants r_1 and r_2 are

$$r_{1,2} = \frac{1}{2}(n_2 - n_1 + 1 \pm \sqrt{(n_1+n_2+1)^2 - 4n_1n_2}), \quad (14)$$

with r_1 being the one with the plus sign in front of the square root. We can take different limits and see how important is the variance of the mutations.

A. $N\mu_1(1-\mu_1)\tau << 1$ and $N\mu_2(1-\mu_2)\tau << 1$

First we take the limit where $N\mu_1(1-\mu_1)\tau << 1$ and $N\mu_2(1-\mu_2)\tau << 1$. This practically means that the population is small, or the rates of the back and forward mutations goes to zero as N grows. As we shall show, in this case the variance of the mutations can be ignored in the Kolmogorov equation and the solution for $P(x)$ is very well approximated by Eq. 10. In this case $n_1 << 1$ and $n_2 << 1$ and it is easy to see that $\Lambda \simeq 4N$, $r_1 \simeq 1$ and $r_2 \simeq 0$. With these values for the constants, it is trivial to show that the solution of the Kolmogorov equation (12) is approximated by (10). Therefore the limit where it is valid to drop the variance of the mutation from the Kolmogorov equation is $N\mu(1-\mu)\tau << 1$, where μ is either μ_1 or μ_2. Practically for very small μ, $1-\mu \simeq 1$. So the limit is $\mu\tau << 1/N$. If the time scales for random mating and mutations are comparable (meaning $\tau \simeq 1$) this limit becomes $\mu << 1/N$ and states the fact that the variance of the mutation is negligible if the rates for back and forward mutations are much smaller than the inverse of the population number.

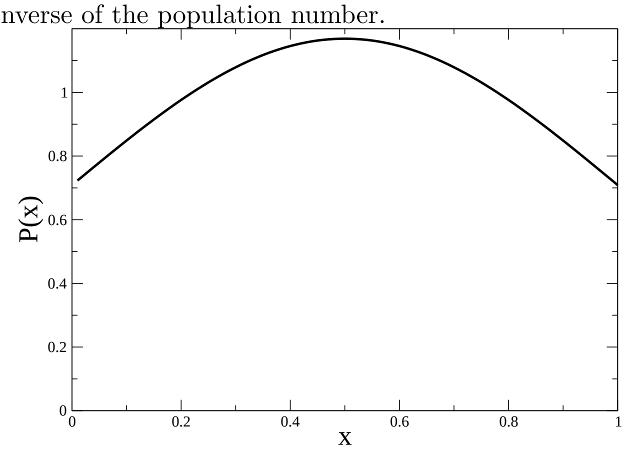

FIG. 2: Probability distribution for the case B with$N\mu \rightarrow \infty$

B. $N\mu_1(1-\mu_1)\tau >> 1$ and $N\mu_2(1-\mu_2)\tau >> 1$

This is the opposite limit from the previous case. In this case there are relatively significant rates for back and forward mutation. For simplicity we shall assume that $\mu_1 = \mu_2 = \mu$. The constant Λ is

$$\Lambda = \sqrt{\frac{2N}{\mu(1-\mu)\tau}}. \quad (15)$$

The constants r_1 and r_2 are

$$r_{1,2} = \frac{1}{2} \pm \sqrt{2N\mu(1-\mu)\tau}. \quad (16)$$

We can write now the solution for $P(x)$:

$$P(x) \propto (\frac{1}{2} + \sqrt{2N\mu(1-\mu)\tau} - x)^{4N\mu-1}(x + \sqrt{2N\mu(1-\mu)\tau} - \frac{1}{2})^{4N\mu-1}. \quad (17)$$

As we can see from the solution, although the power is exactly the same as in the previous case, the zeros where the probability vanishes are not at $x = 0$ and $x = 1$ anymore. They have moved to $1/2 \pm \sqrt{2N\mu(1-\mu)\tau}$. However the allowed values of x are between 0 and 1. Naively we might expect that as the $N\mu(1-\mu)\tau$ becomes larger and larger, and the roots r_1 and r_2 move away from the zero axis, the probability distribution tends to become uniform. However, this not what we found analytically and numerically. We derive now a formula for $P(x)$ valid at the $N\mu(1-\mu)\tau >> 1$ limit. Let's call $k = \sqrt{2N\mu(1-\mu)\tau} \simeq \sqrt{2N\mu}$, assuming $\tau \simeq 1$ and that $1-\mu \simeq 1$. In this case Eq. 17 can be written as

$$(1/2+k-x)^{2k^2-1}(x-1/2+k)^{2k^2-1} = k^{2k^2-1}(1-\frac{x-1/2}{k})^{2k^2-1}(1+\frac{x-1/2}{k})^{2k^2-1} \propto (1-\frac{(x-1/2)^2}{k^2})^{2k^2-1}. \quad (18)$$

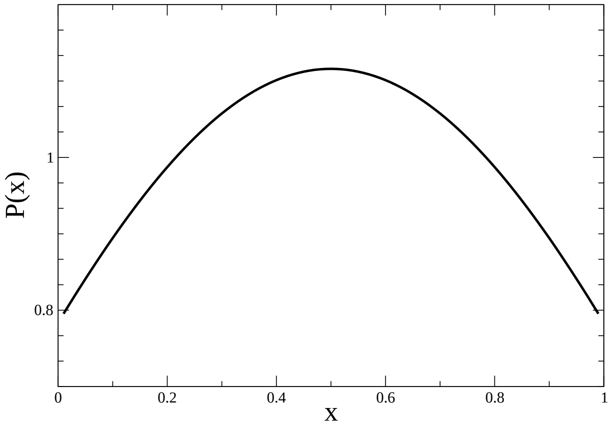

FIG. 3: Probability distribution for the case C

If we take the limit $k \to \infty$ this expression is $e^{-2(x-1/2)^2}$. So instead of a uniform distribution we get a Gaussian distribution centered at $x = 1/2$ and with a deviation $\sigma = 1/2$ again. This is the main result of this paper. In case with strong mutations independent of the population number N, the variance of the mutation makes the probability distribution $P(x)$ a Gaussian. We checked the above analytical result numerically. Even for a k as low as 10 Eq. 17 is in perfect agreement with the formula

$$P(x) = 0.855e^{-2(x-1/2)^2}, \quad (19)$$

where the prefactor is fixed by normalization.

C. $\mu = \gamma/N$

So far we investigated cases where the back and forward mutation rates were independent of N. We explored the two different limits where the variance of the mutation is negligible (first case) and significant (second case). It is very interesting to see what happens however if the mutation rates depend on the population number. In this subsection we examine the case where the rates μ_1 and μ_2 scale as $1/N$. More precisely we assume that $\mu_1 = \mu_2 = \mu = \gamma/N$, where γ is an arbitrary constant. Again we can rewrite the constant Λ as

$$\Lambda = \frac{4N}{\sqrt{1+8\gamma\tau}}. \quad (20)$$

The constants r_1 and r_2 are

$$r_{1,2} = \frac{1}{2}(1 \pm \sqrt{1+8\gamma\tau}. \quad (21)$$

The solution of the Kolmogorov equation is

$$P(x) \propto (\frac{1}{2}(1+\sqrt{1+8\gamma\tau})-x)^{4\gamma-1}(x+\frac{1}{2}(\sqrt{1+8\gamma\tau}-1))^{4\gamma-1}. \quad (22)$$

As we can see this is an intermediate situation between the cases A and B. If we take a characterisitc value of $\gamma = 1$ and again we consider $\tau \simeq 1$, then the zeros of the probability density are in -1 and 2, not very far away from the interval [0,1]. This means that we don't have a Gaussian distribution but something between the cases A and B.

IV. CONCLUSIONS

In this paper we studied the effect of mutations in the Kolmogorov equation. Roughly speaking, we showed that in systems with high mutation rates $N\mu > 1$ the probability distribution for the number population is a Gaussian with mean value and standard deviation independent of both the mutation rates and the population number. We also showed that in the case of $N\mu << 1$ the variance of the mutation can be safely ignored from the Kolmogorov equation.

Acknowledgments

We acknowledge helpful conversation with Profs. M. Kardar, L. Mirny and with M. Desai.

[1] G.H. Hardy (1908) "Mendelian proportions in a mixed population." Science, **28**, 49-50.

[2] W Weinberg (1908) "Über den Nachweis der Vererbung

beim Menchen," Naturk.Württemb., **64**, 369-82.

[3] S.H. Rice "Evolution Theory: Mathematical and Conceptual Foundations" Sinauer Associates, Inc. Publishers.

[4] D.L. Hartl "A Primer of Population Genetics" Sinauer Associates, Inc. Publishers.

[5] J.H. Gillespie " Population Genetics" The John Hopkins University Press.

[6] D.S. Falconer ,T.F.C. Mackay "Introduction to Quantitative Genetics" Longman Group Limited.

[7] J.F. Crow "Genetics Notes: An Introduction to Genetics" Macmillan Publishing Company.

[8] J.F. Crow ,M. Kimura " An Introduction to Population Genetics Theory" Harper and Row, Publishers.

[9] A Kolmogorov (1931) "Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung," Math. Ann., 104, 415-458.

[10] M. Kimura (1954) "Process leading to quasi-fixation of genes in natural populations due to random fluctuations of selection intensities," Genetics 39:280-295.

[11] M. Kimura (1962) "On the probability of fixation of mutant genes in a population," Genetics 47:713-719.