Epistatic interactions between mutations

Sivaraman Ramaswamy
8.592J-HST.452J Statistical Physics in Biology

(Dated: May 15, 2013)

Drift-diffusion equation remains a popular approach to find the probability distribution of a particular gene in the population. This contribution deals with the case when a gene interacts with another, thereby affecting each other's fitnesses of the mutant genotypes. The proposed model modifies the drift velocity contribution from the selection of the genotypes and reformulates the drift-diffusion equation. A comparison with the case of no interaction between the mutants is also presented.

Introduction

Probabilistic ideas became an important part in the study of population genetics when it was first used by Fisher [1, 2] and Wright [3]. They used it to formulate the Fisher-Wright model to explain genetic drift in alleles. In their model, they calculate what the probability of having n of the $2N$ genes (diploid population of size N) in the population at any time t of a given type is. This formulation was further extended by Kimura [4, 5]. Crow and Kimura [6] then developed the diffusion theory, which played a prominent part in the theoretical studies of population genetics. In particular, it incorporates both mutation and selection of a particular genetic type, in addition to the genetic drift due to mating. This model helps one predict the probability with which any allele or its mutant gets fixed in the population and the rate at which polymorphism decays. In this model, the probability of the proportion of the individuals of a particular genetic type follows a drift-diffusion equation, also known as forward Kolmogorov equation. Other standard texts [7] have also used these results. An alternate mathematical theory, called coalescent theory, was also developed [8] and used [9]. However, drift-diffusion equation remains the most popular approach due its simplicity.

The drift-diffusion equation have been used in single-locus models [10, 11] as well as in multiple loci [12, 13]. However, no one has dealt with the case when the fitnesses of mutations in different loci depend on each other. In this work, a new model to deal with "interacting mutations" in two loci is presented. This was done by modifying the drift expression in the governing equation to incorporate the epistatic interactions. The advantage of this model is that it can be easily extended for more loci.

Drift-diffusion equation for single locus

Consider a single locus in a population of N diploid individuals. Let A be the dominant allele in that location. The ratio of the total number of copies of A in the population (n), to the total number of alleles in that locus ($2N$), is called the gene frequency and is denoted by x. Let $p(x,t)$ be the probability distribution function of x at time t. Let a be the recessive mutant in that locus, such the forward rate of mutation is μ_2 and the backward rate is μ_1, and s_1 be the quantity to define the fitnesses of its genotypes. Table 1 shows the fitness of the various genotypes occuring from these allele types.

TABLE. 1 Fitness of various genotypes

Genotype	AA	Aa	aa
	1	$1-hs_1$	$1-s_1$

The drift-diffusion equation governing this probability is

$$\frac{\partial}{\partial t}p(x,t) = -\frac{\partial}{\partial x}(v(x)p(x,t)) + \frac{\partial^2}{\partial x^2}(D(x)p(x,t)) \quad (1)$$

where $v(x)$ is the drift velocity and $D(x)$ is the diffusion coefficient.

Under only genetic drift, the drift and diffusion coefficients are [6] -

$$\begin{aligned} v(x) &= 0 \\ D(x) &= \frac{1}{4N}x(1-x) \end{aligned} \quad (2)$$

Under only mutation, these quantities can be easily calculated.

$$\begin{aligned} v(x) &= \mu_1(1-x) - \mu_2 x \\ D(x) &= \frac{1}{2N}[\mu_1(1-x) + \mu_2 x] \end{aligned} \quad (3)$$

Under only selection, assuming no dominance ($h = 0.5$), the quantities are [6] -

$$\begin{aligned} v(x) &= \frac{s_1}{2}x(1-x) \\ D(x) &= 0 \end{aligned} \quad (4)$$

Thus the effective rates become -

$$v(x) = \mu_1(1-x) - \mu_2 x + \frac{s_1}{2}x(1-x)$$
$$D(x) = \frac{1}{4N}x(1-x) \tag{5}$$

The diffusion coefficient contains contribution due to only the genetic drift, as the contribution from mutations is comparitively insignificant. This is due to the fact that the mutation rates are generally very small.

Considering only the steady state probability distribution, the ordinary differential equation can be solved. The boundary conditions come from the fact that the probability current must be zero and that the probability should be normalized. The probability distribution then becomes proportional to

$$p^*(x) \propto \frac{x^{4N\mu_1}(1-x)^{4N\mu_2}e^{2Ns_1x}}{x(1-x)} \tag{6}$$

Drift-diffusion equation for two indepedent mutations

Now, consider a pair of loci where there will not be any influence of mutations of one gene over the other, i.e. the fitness of one kind of genotype will not affect the other genotype. Let B and b be the alleles in the second locus, with y being the gene frequency. Let η_2 and η_1 be the forward and backward mutation rates respectively and s_2 be the corresponding quantitiy to define the fitness of its genotypes. If $p(x, y, t)$ is the joint probability distribution function of x and y at time t, the drift-diffusion equation governing this joint probability is given by

$$\begin{aligned}\frac{\partial}{\partial t}p(x,y,t) = &-\frac{\partial}{\partial x}(v_x(x,y)p(x,y,t)) \\ &-\frac{\partial}{\partial y}(v_y(x,y)p(x,y,t)) \\ &+\frac{\partial^2}{\partial x^2}(D_{xx}(x,y)p(x,y,t)) \\ &+\frac{\partial^2}{\partial y^2}(D_{yy}(x,y)p(x,y,t))\end{aligned} \tag{7}$$

with

$$v_x(x,y) = \mu_1(1-x) - \mu_2 x + \frac{s_1}{2}x(1-x) \tag{8}$$
$$D_{xx}(x,y) = \frac{1}{4N}x(1-x) \tag{9}$$
$$v_y(x,y) = \eta_1(1-y) - \eta_2 y + \frac{s_2}{2}y(1-y) \tag{10}$$
$$D_{yy}(x,y) = \frac{1}{4N}y(1-y) \tag{11}$$

Since the genotypes are indepedent and assuming no dominance in the second type of alleles too, the expressions for the $v_x(x, y)$ and $v_y(x, y)$ (and similarly for $D_{xx}(x, y)$ and $D_{yy}(x, y)$) depends only only x and y respectively. The cross terms $D_{xy}(x, y)$ and $D_{yx}(x, y)$ are zeroes due to the independence of the genes and hence those terms were not considered. Therefore, the steady state probability distribution can be calculated easily.

$$\begin{aligned}p_I^*(x,y) &\propto \left[\frac{x^{4N\mu_1}(1-x)^{4N\mu_2}e^{2Ns_1x}}{x(1-x)}\right] \\ &\quad\times\left[\frac{y^{4N\eta_1}(1-y)^{4N\eta_2}e^{2Ns_2y}}{y(1-y)}\right] \\ &\propto \frac{x^{4N\mu_1}(1-x)^{4N\mu_2}y^{4N\eta_1}(1-y)^{4N\eta_2}e^{2N(s_1x+s_2y)}}{x(1-x)y(1-y)}\end{aligned} \tag{12}$$

Drift-diffusion equation for two dependent mutations

Suppose the two loci were such that the fitnesses of the genotypes of the first gene affects the fitnesses of the genotypes of the second gene. Let the fitness of the combination of the genotypes be the product of the individual fitnesses, as shown in Table 2.

TABLE. 2 Fitness of various genotypes

Genotype	AA	Aa	aa
BB	1	$1-hs_1$	$1-s_1$
Bb	$(1-hs_2)$	$(1-hs_2)(1-hs_1)$	$(1-hs_2)(1-s_1)$
bb	$(1-s_2)$	$(1-s_2)(1-hs_1)$	$(1-s_2)(1-s_1)$

Considering the first row of Table 2, the equilibrium frequency of the BB genotype, as found by the Hardy-Weinberg equilibrium, is y^2. Since only the drift velocity depends on the fitness of the genotypes, it can be assumed that the contribution of BB to the drift velocity of the two-gene genotype in the first column would be -

$$v_x(x,y) = \left[\mu_1(1-x) - \mu_2 x + \frac{s_1}{2}x(1-x)\right]y^2 \tag{13}$$

Similarly, considering the frequency of Bb, which is $2y(1-y)$, and bb, which is $(1-y)^2$, and the multiplying the fitness expression with the corresponding factor, the drift velocity for the first gene is obtained

by adding the individual contributions.

$$\begin{aligned} v_x(x,y) = &\ \mu_1(1-x) - \mu_2 x + \left[\frac{s_1}{2}x(1-x)\right] y^2 \\ &+ \left[(1-hs_2)\frac{s_1}{2}x(1-x)\right] 2y(1-y) \\ &+ \left[(1-s_2)\frac{s_1}{2}x(1-x)\right] (1-y)^2 \\ = &\ \mu_1(1-x) - \mu_2 x + [1 - s_2(1-y)]\,\frac{s_1}{2}x(1-x) \end{aligned} \tag{14}$$

This expression is consistent with the fact that drift velocity due to mutations remains unchanged and only the selection term is affected. In the same way, the drift velocity for the other gene can also be calculated.

$$\begin{aligned} v_y(x,y) = &\ \eta_1(1-x) - \eta_2 x + \left[\frac{s_2}{2}y(1-y)\right] x^2 \\ &+ \left[(1-hs_1)\frac{s_2}{2}y(1-y)\right] 2x(1-x) \\ &+ \left[(1-s_1)\frac{s_2}{2}y(1-y)\right] (1-x)^2 \\ = &\ \eta_1(1-y) - \eta_2 y + [1 - s_1(1-x)]\,\frac{s_2}{2}y(1-y) \end{aligned} \tag{15}$$

The diffusion coefficients $D_{xx}(x,y)$ and $D_{yy}(x,y)$ remains unchanged as the mutation and selection contribution is much smaller than the genetic drift contribution and for the same reason, the cross terms $D_{xy}(x,y)$ and $D_{yx}(x,y)$ remain zeroes.

Using Eq. (7) and using similar boundary conditions, the steady state probability for this case can be obtained.

$$\begin{aligned} p_D^*(x,y) \propto &\frac{x^{4N\mu_1}(1-x)^{4N\mu_2}y^{4N\eta_1}(1-y)^{4N\eta_2}}{x(1-x)y(1-y)} \\ &\times e^{2N[s_1x + s_2y - s_1s_2(x+y-xy)]} \end{aligned} \tag{16}$$

Comparison of the probability distributions

These probabilities, $p_I^*(x,y)$ and $p_D^*(x,y)$, can be compared for different values of N, s_1 and s_2. For the sake of comparison, it can be assumed that μ_1, μ_2, η_1 and η_2 have the same numerical value, as they will affect both probabilities in the same manner. This value was arbitrarily chosen as 0.075.

In figure 1, a low value of 10 was used for N and the values of s_1 and s_2 were both 0.1. The figure shows the contour plots of the probability distribution, with x and y being the x and y-axis respectively. From the value of the contours and the position of the peak, it can be seen that there is not much difference in the probability distribution between the independent and the dependent mutations case. Similar plots were generated for higher values of N and for lower values of s_1 and s_2, none of which showed much difference in the probability distribution of the two cases.

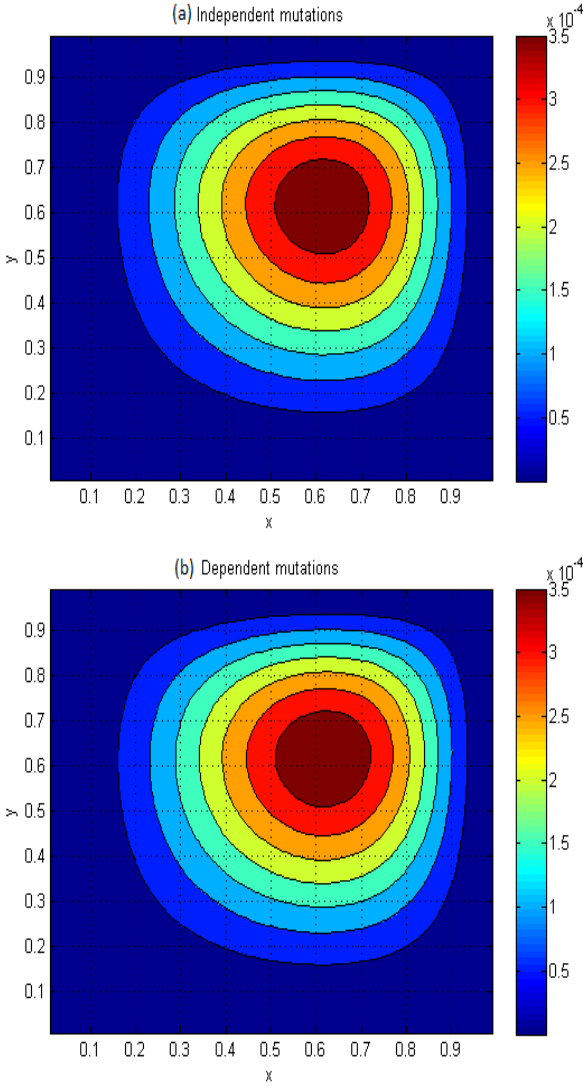

FIG. 1 Contour plots comparing the probability distributions of the (a) independent and (b) dependent mutations case. The parameter values used are $N = 10$ and $s_1 = s_2 = 0.1$. Not much difference is observed for these parameter values.

In figure 2, a higher value of 0.5 was selected for s_1 and s_2. N was kept same at 10. In this figure, a more significant difference is observed between the two probability distributions. Since the fitness factor, s_1 and s_2, is high for both genes, the dominant genotype is highly favored. Hence, at equilibrium, the peak is shifted towards the higher values of x and y. In figure

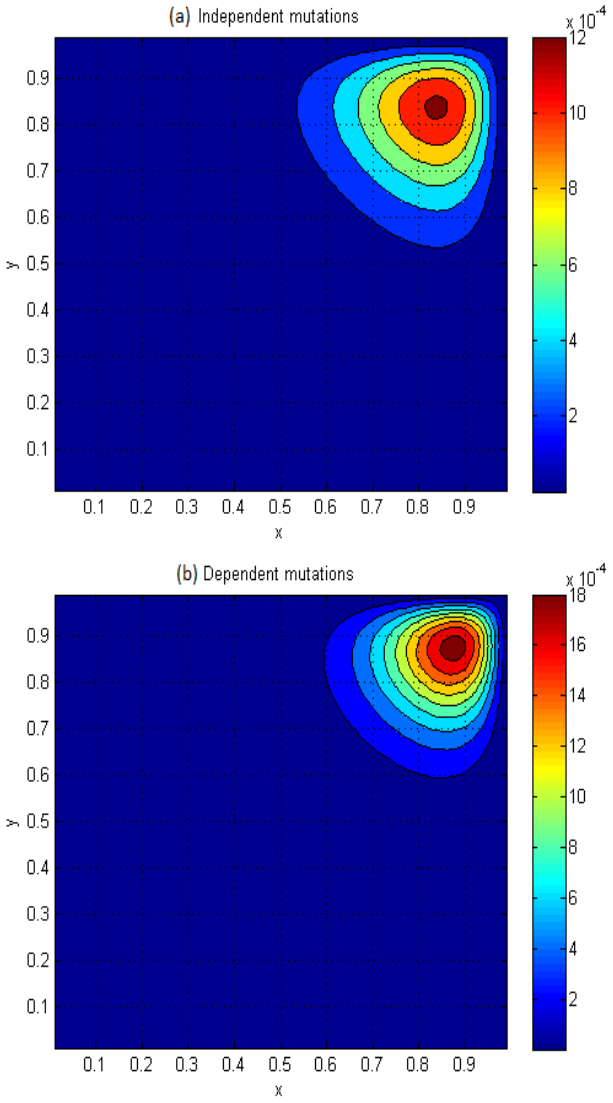

FIG. 2 Contour plots comparing the probability distributions of the (a) independent and (b) dependent mutations case. The parameter values used are $N = 10$ and $s_1 = s_2 = 0.5$. Significant difference is observed for these parameter values.

2, the peak of the dependent mutations case is more shifted towards the higher values of x and y than for the independent mutations case. Also, the probability distribution for the independent mutations case is more spread out, with the peak value lower than that for the dependent mutations case. These features indicate that dependent mutations favour the two-gene dominant genotype more than the independent mutations. Similar plots where generated for higher values of N, but in those cases, the absolute difference in the two plots were really small. Those plots have been omitted due to space constraints.

Conclusion

A new model is presented to explain the effects of mutations in different loci that affect each other's fitness. This model is based on the original drift-diffusion equation proposed in literature, with modifications to the drift velocity term. The obtained steady state probability distribution for the gene frequencies was also compared with the case when the mutations' fitnesses are independent. The results indicate that the differences in the steady state probability distribution are significant only for the case of small population and significant dominance of one genotype. This model can be easily extended for more than two loci. However, experimental data is required to validate the proposed model.

References

[1] R.A. Fisher. *Proc. Roy. Soc. Edin.*, 42:321, 1922.

[2] R.A. Fisher. *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, 1930.

[3] S. Wright. *Genetics*, 16(2):97, 1931.

[4] M. Kimura. *Cold Spring Harbor Symposia on Quantitative Biology*, 20(1):33, 1955.

[5] G.A. Watterson. *Theor Popul Biol.*, 49:154, 1996.

[6] J.F. Crow, M. Kimura, et al. *An introduction to population genetics theory*. Harper & Row, New York, 1970.

[7] W.J. Ewens. *Mathematical population genetics*. Springer Verlag, Berlin, 1979.

[8] J. F. C. Kingman. *Journal of Applied Probability*, 19:27, 1982.

[9] M. Nordborg. *Coalescent Theory: Handbook of Statistical Genetics*. John Wiley & Sons, Ltd, 2004.

[10] M. Kimura. *Genetics*, 47(6):713, 1962.

[11] M. Kimura and J.F. Crow. *Genetics*, 49(4):725, 1964.

[12] S. Karlin. *Theoretical Population Biology*, 7(3):364, 1975.

[13] G.J. Baxter, R.A. Blythe, and A.J. McKane. *Mathematical Biosciences*, 209(1):124, 2007.