Rate of Fixation of Rare Variants in a Population

Bhavin S. Khatri[1, 2]
[1] *The Francis Crick Institute,*
1 Midland Road,
London, NW1 1AT, U.K.
[2] *Division of Infection & Immunity,*
University College London,
London, WC1E 6BT, U.K.

(Dated: 2 April 2017)

The process of molecular evolution has been dominated by the Kimura paradigm for nearly 60 years; mutations arise at a certain rate in the population and they go to fixation with a probability given by Kimura's classic formula, which assumes there are no further mutations that interfere with the fixation process. An alternative view is that rare variants exist in the population in a mutation-drift-selection balance and rise to fixation through a combination of chance (genetic drift), selection and mutation. When mutations increase in strength, but still in the weak regime, we would expect the Kimura rate approximation to be an overestimate, as a rare variant which grows in frequency will suffer a greater backward flux of mutations, slowing progress to fixation. However, to date calculating important quantities for a general model of selection and mutation, like the rate of fixation of these rare variants has not been tractable in the conventional diffusion approximation of population genetics. Here, we use Fisher's angular transformation to convert the frequency-dependent diffusion inherent in population genetics to simple diffusion in an effective potential, which describes the forces of selection, drift and mutation. Once this potential is defined it is simple to show that the mean first passage time is given by a double integral which relate to populations at the barrier. Exact numerical integration shows excellent agreement with discrete Wright-Fisher simulations, which do show a slowing down of the fixation of mutants at higher mutation rates and for strong positive selection, compared to the Kimura prediction. We then seek a closed-form analytical expression for the rate of fixation of mutants, by adapting Kramer's approximation for the mean first passage time. This overall gives an accurate approximation, but however, does not improve on the Kimura rate.

I. INTRODUCTION

The probability of fixation of a mutant in a wild-type population is a key quantity in population genetics; given some initial frequency x_0 in a population of finite size N, what is the probability that the frequency of this mutant variant reaches $x = 1$. In the simple case with no mutations, Kimura calculated his famous equation[1], which describes a probability of fixation which has a saturating form, in the diffusion limit of the Wright-Fisher model. This has formed the basis of understanding the substitution process in studies of molecular evolution in the weak mutation regime, where mutations arise do novo and fix with a probability given by Kimura's equation; the overall rate of substitution of different mutants is simply the per individual mutation rate multiplied by the population size and Kimura's fixation probability.

However, in reality when mutations are of sufficient strength, but still in the weak regime ($N\mu < 1$), we would expect even when one variant is nominally "fixed", there will be polymorphisms that co-exist at low frequency in the population. Further, as a variant goes towards fixation, back mutations should retard its progress, causing a slowing of the rate of fixation, compared to the Kimura rate. If mutations are included in the diffusion approximation of a Wright Fisher process, it is not possible to calculate the mean time to fixation in simple closed-form. A key difficultly of the diffusion approximation is that the effective diffusion constant for variant frequencies, depends on the frequency of the variant. When a variant is close to fixation or loss, diffusion slows down compared to intermediate frequencies since the variance in the change in variant frequency has a characteristic binomial form. Fisher found a transformation to an angular frequency, where diffusion is independent of angular frequency, which, however, comes at the cost of introducing an effective non-linear potential that describes the flux of diffusers to the boundaries[2]. This non-linearity makes any exact calculations of the dynamics difficult, but Khatri[2], developed a heuristic method to find accurate asymptotic Gaussian Greens functions in the short-time limit. Here we show that given this potential, a standard technique can be used to calculate the MFPT, where it is given as a double intergral over the barrier and well populations. When compared to discrete Wright-Fisher simulations, numerical evaluation of this integral gives very accurate estimates of the MFPT and confirms that variants at higher mutation rates have their fixation rate retarded, compared to the Kimura rate. We also calculate a Kramers-type approximation[3], which is classically used for the MFPT of chemical reactions, where the potential energy barrier impeding a reaction is large; comparing to the discrete Wright-Fisher simulations we find the calculation is accurate, but does not improve on the Kimura rate. In addition, the Kramers approximation, as for Kimura theory, fails to predict the reduction of fixation rate at higher mutation rates and strong positive selection.

II. FISHER'S ANGULAR TRANSFORMATION FOR 2-VARIANTS

In the diffusion approximation the stochastic dynamics of gene frequency x ($= n/N$, where n is the number of copies of the mutant variant a_1 and N the total population) is given by

$$\frac{\partial p(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left(A(x)p(x,t)\right) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left(B(x)p(x,t)\right), \quad (1)$$

where $p(x,t)$ is the probability density of gene frequency and $A(x)$ is the mean change in gene frequency per generation and $B(x)$ is the variance of gene frequency change per generation - we assume these only have a time-dependence though $x(t)$. This is the forward Fokker-Planck equation whose solutions represent a progression forward though time given an initial condition $p(x,0) = \delta(x - x_0)$ - i.e. we know the initial gene frequency at time zero. For selection and arbitrary mutation:

$$A(x) = \Delta f x(1-x) + \mu_{21}(1-x) - \mu_{12}x \quad (2)$$

where $\Delta f = f(a_2) - f(a_1)$, so $\Delta F > 0$ means selection favours variant a_2, μ_{12} is the mutation rate from variant $1 \rightarrow 2$ and μ_{21} the mutation rate from variant $2 \rightarrow 1$ and

$$B(x) = \frac{1}{N}x(1-x). \quad (3)$$

They key difficulty with Eqn.1 is that the variance of gene frequency change per generation depends on the frequency x. Fisher proposed a transformation to a different co-ordinate θ:

$$\theta = \cos^{-1}(1-2x). \quad (4)$$

It is simple to show that the resultant Fokker-Planck equation is a diffusion equation in an effective potential

$$\frac{\partial q}{\partial t} = \frac{1}{2N}\frac{\partial^2 q}{\partial \theta^2} + \frac{\partial}{\partial \theta}\left(\frac{\partial U(\theta)}{\partial \theta} q\right) \quad (5)$$

where the derivative of the effective potential in the above equation is given by,

$$2N\frac{\partial U}{\partial \theta} = -\Bigg(N\Delta f \sin(\theta) + (2N(\mu_{12}+\mu_{21}) - 1)\cot(\theta) + \frac{2N(\mu_{21}-\mu_{12})}{\sin(\theta)}\Bigg). \quad (6)$$

so that the potential is found on integration to be

$$2NU(\theta) = \big(N\Delta f \cos(\theta) - (2N(\mu_{12}+\mu_{21}) - 1)\ln(\sin(\theta)) - 2N(\mu_{21}-\mu_{12})\ln(\tan(\theta/2))\big). \quad (7)$$

From Eqn.5 it is then simple to show that the equilibrium pdf is given by

$$p^*(\theta) = \frac{1}{Z}e^{-2NU(\theta)}. \quad (8)$$

In Fig.1, we have a plot of the potential for $2N\mu_{12} = 0.1$ and $\mu_{21} = 2\mu_{12}$ for various values of $2N\Delta F$.

III. MEAN FIRST PASSAGE TIME FOR 2-VARIANTS

The first passage time is the time it takes for a diffuser to *first* reach a boundary – this will have a distribution and we want to calculate it's mean, which is the mean first passage time (MFPT). There are a number of ways to do this, one of which is to solve the corresponding backward Fokker-Planck equation[4,5] for the mean first passage time for an initial frequency; however, under selection and mutation, no-known closed form solution is known. Here, we will assume that the initial frequency of the rare variant is not known, but that it is in quasi-equilibrium due to a mutation-selection-drift balance near $x = 0$ or $\theta = 0$. After some period of time the rare variant will drift in frequency to a critical value, after which a combination of selection and mutation take the variant to fixation. The critical frequency θ^* will be given by the maximum in the effective potential $U(\theta)$. A method to solve such problems was developed by Kramers[3]. Here, the simplest exposition is to consider that we inject rare variants into the system at $\theta = 0$, at a rate J, and then remove them when they reach fixation. From Eqn.5 the net flux in the system must equal to a constant, the number we inject per generation J:

$$-\frac{1}{2N}\frac{\partial \tilde{q}}{\partial \theta} - \left(\frac{\partial U(\theta)}{\partial \theta}\tilde{q}\right) = J \quad (9)$$

where $\tilde{q}(\theta)$ is the non-equilibrium *steady-state* distribution when there is a net flux through the system. Its integral is the number of diffusers in the system and so the MFPT τ simply obeys

$$\tau = \frac{\int_0^\pi \mathrm{d}\theta \tilde{q}(\theta)}{J}, \quad (10)$$

An expression for the steady-state distribution can be obtained by integrating Eqn.9:

$$\tilde{q}(\theta) = 2NJe^{-2NU(\theta)}\int_\theta^{\theta^\dagger} \mathrm{d}\theta' e^{2NU(\theta')} \quad (11)$$

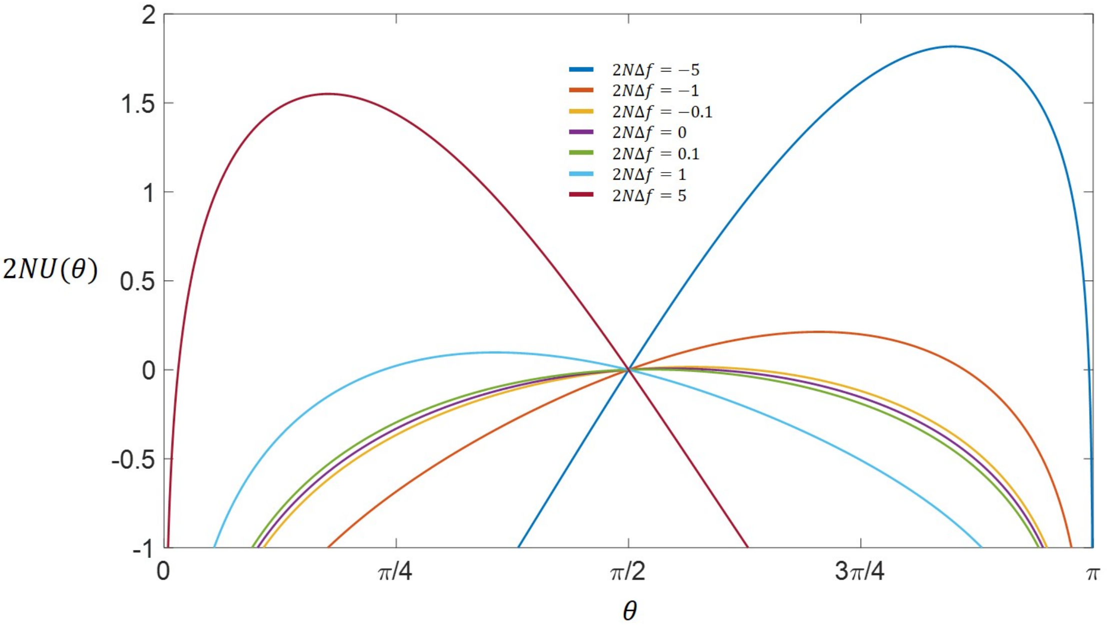

FIG. 1. The effective potential $U(\theta)$ that arises after Fisher's angular transformation Eqn.4, for $2N\mu_{12} = 0.1$ and $\mu_{21} = 2\mu_{12}$ for various values of $2N\Delta F$.

where $\theta^\dagger$ is the angular frequency, which corresponds to fixation - as we shall see our answer will not be very critical on the exact value of $\theta^\dagger$. Using Eqn.10 and 11, and swapping the order of integration, we find the MFPT is given by the following double integral:

$$\tau = 2N \int_0^{\theta^\dagger} \mathrm{d}\theta' e^{2NU(\theta')} \int_0^{\theta'} \mathrm{d}\theta e^{-2NU(\theta)}. \qquad (12)$$

This expression is exact and can be numerically integrated for given values of Δf, N, μ_{12}, and μ_{21}. However, the potential is singular at $\theta = 0$ and $\theta = \pi$, which requires a careful numerical integration scheme that essentially integrates a modified integrand, where the singularity is removed by choice of a function that has the same limiting form at the singular points and is itself is integrable over the region of interest. Using the limiting form $\lim_{\theta\to 0}\{e^{-2NU}\} \to e^{-N\Delta f}2^{-2N\Delta\mu}\theta^{4N\mu_{21}-1}$ as a function that is simple to integrate, we perform this numerical integration using a standard numerical routine in Matlab. The results are shown in Fig.2 and show excellent agreement with the MFPTs calculated from simulations of the discrete Wright-Fisher process.

We can also develop an approximation of the integral as follows, which is a modification of Kramer's approximation[3]: the outer integral is of the form $\int_0^{\theta^\dagger} \mathrm{d}\theta' P(\theta < \theta')e^{2NU}$ and so will be dominated by values of theta where U is maximum, which when mutation is weak, will correspond to the top of the barrier of U – at the barrier the function $P(\theta < \theta')$ (which comes from the inner integral above) will vary slowly as most of the density will be to the "left" of the barrier, so we can take this factor out of the integral, evaluated at the barrier θ^*, to give $\tau \approx 2NP(\theta < \theta^*)\int \mathrm{d}\theta e^{2NU}$. To evaluate $P(\theta < \theta^*)$, we make the approximation that $\theta \ll 1$, $2NU(\theta) \approx Ns(1-\theta^2/2) - (2N\mu_\Sigma - 1)\ln(\theta) - 2N\Delta\mu\ln(\theta/2)$ (where $\mu_\Sigma = \mu_{12} + \mu_{21}$ and $\Delta\mu = \mu_{21} - \mu_{12}$). The resultant integral that defines $P(\theta < \theta^*)$ can then be rearranged by change of variable to give the definition of the lower incomplete gamma function $\gamma(z,a) = \int_0^a \mathrm{d}uu^{z-1}e^{-u}$, so that the inner integral is given by

$$\begin{aligned} P(\theta < \theta^*) &\approx e^{-N\Delta f}2^{-2N\Delta\mu}\int_0^{\theta^*} \mathrm{d}\theta e^{\frac{1}{2}N\Delta f\theta^2}\theta^{4N\mu_{21}-1} \\ &= 2^{2N\mu_{12}-1}e^{-N\Delta f}(-N\Delta f)^{-2N\mu_{21}} \\ &\quad \gamma(2N\mu_{21}, -N\Delta f(\theta^*)^2/2). \qquad (13) \end{aligned}$$

Next to evaluate the outer integral, we approximate the potential around the maximum of the barrier by a quadratic, so $U(\theta) \approx U(\theta^*) - \frac{1}{2}\kappa(\theta - \theta^*)^2$, where $\kappa = |\mathrm{d}^2U/\mathrm{d}\theta^2|_{\theta=\theta^*}$ so that,

$$\begin{aligned} \int_0^{\theta^\dagger} \mathrm{d}\theta' e^{2NU(\theta')} &\approx e^{2NU(\theta^*)}\int_0^{\theta^\dagger} \mathrm{d}\theta' e^{-N\kappa(\theta'-\theta^*)^2} \\ &\approx e^{2NU(\theta^*)}\sqrt{\frac{\pi}{N\kappa}}, \qquad (14) \end{aligned}$$

where we have extended the upper and lower limits to $\pm\infty$, so we can use the standard integral $\int_{-\infty}^{\infty} \mathrm{d}ze^{-\alpha z^2} = \sqrt{\pi/\alpha}$.

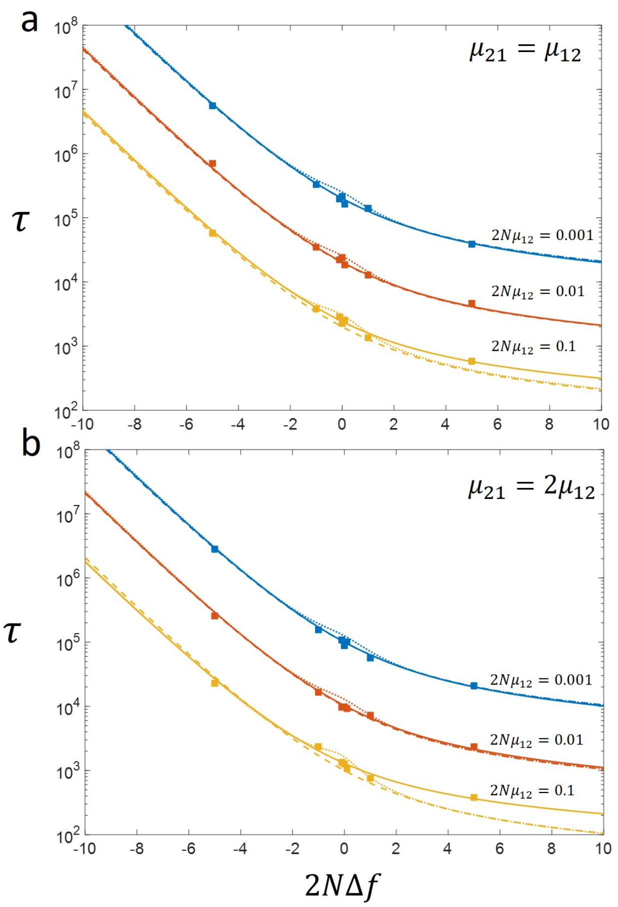

FIG. 2. Comparison of discrete Wright Fisher simulations of MFPT (squares) to numerical integration of Eqn.14 (solid line) and Kimura's approximation (dashed line) and Kramer's method (dotted line). Top graph is for $\mu_{12} = \mu_{21}$ and bottom graph is for $\mu_{21} = 2\mu_{12}$.

This last approximation assumes that $1/\sqrt{2N\kappa} \ll \pi$, which as we can see from Fig.1 will not be very good for $2N\Delta f \ll 1$; we will see in Fig.2 that if we replace this approximation for the barrier integral with a numerical integration, this regime is where this calculation is less accurate. The final expression for the MFPT is then

$$\tau \approx e^{-N\Delta f} 2^{-2N\mu_{12}} \sqrt{\frac{4N\pi}{\kappa}} e^{2NU(\theta^*)} (-N\Delta f)^{-2N\mu_{21}} \gamma(2N\mu_{21}, -\frac{1}{2}N\Delta f(\theta^*)^2). \tag{15}$$

The rate of fixation of the rare variants is then $k = 1/\tau$. The only task left is to calculate the position of the barrier, θ^* and the curvature at the barrier, $\kappa = |\mathrm{d}^2U/\mathrm{d}\theta^2|_{\theta=\theta^*}$. The first is done by solving for $\mathrm{d}U/\mathrm{d}\theta = 0$, which gives a quadratic equation for $\cos\theta^*$, which has solution:

$$\cos(\theta^*) = \frac{2N\mu_\Sigma - 1 + 2N\kappa}{2N\Delta f}. \tag{16}$$

where κ is calculated directly from the second derivative of the potential as:

$$2N\kappa = \sqrt{(2N\mu_\Sigma - 1)^2 + 4N\Delta f(N\Delta f + 2N\Delta\mu)}. \tag{17}$$

We can see that comparing to simulations in Fig.2, Eqn.15 is generally quite accurate, where this accuracy diminishes somewhat for $2N\mu = 0.1$ and $2N\Delta f = 5$, where the calculation underestimates the MFPT. Finally, we compare this calculation to the Kimura rate:

$$k = \mu_{21} \frac{1 - e^{-2\Delta f}}{1 - e^{-2N\Delta f}}, \tag{18}$$

where $\tau = 1/k$ and we assume the net mutational input only depends on μ_{21}. We see that the Kimura rate (dashed lines in Fig.2 very accurately predicts the mean first passage time, including the regime $2N\Delta f \ll 1$, where Eqn.15, in contrast, is less accurate. However, for large and positive $2N\Delta f$ the Kimura formula underestimates the MFPT, as also found with Eqn.15.

These results support the original hypothesis that when mutations begin to increase in strength (but still in the weak mutation regime $N\mu \ll 1$), we should expect to see that the Kimura approximation will overestimate the rate of fixation due to the retardation effect of back mutations as a variant approaches fixation; this is consistent with the observation that the effect is stronger when the difference in mutation rates $\Delta\mu > 0$, such that back mutations are stronger. However, we find that this retardation effect is only significant when $2N\Delta f > 1$.

IV. CONCLUSIONS

Calculating the rate of fixation of rare variants or polymorphisms in population is an important quantity to understand the rate of molecular evolution. Although, the Kimura approximation for the rate of fixation of mutants is widely popular, it ignores the effect of mutations, as a mutant rises to fixation. We should expect it to overestimate the rate of fixation for large mutation rates, due to an increasing flux of mutations back to the wildtype, as the mutant allele approaches fixation. To investigate this effect, we show that techniques from chemical reaction kinetics for calculating the rate of crossing a potential energy barrier, can be adapted to this canonical population genetics problem; first the frequency-dependent diffusion of the Wright-Fisher process is removed by Fisher's angular transformation[6], which results in simple diffusion or Brownian motion in an effective potential, which is analogous to the potential energy in a chemical reaction. The mean first passage time is then expressed as a double integral over the potential surface; we show that careful numerical integration of this integral gives excellent agreement with the MFPT calculated from discrete Wright-Fisher simulations. In comparison, we see the Kimura approximation does underestimate the MFPT, but only for larger mutation rates and for strong positive selection. The failure of the Kimura theory is likely due to the effect of back mutations retarding increase in frequency of the allele, as it approaches fixation; in support of this we find that the magnitude of the discrepancy between Kimura and Wright-Fisher simulations, increases as the ratio of the backward to forward mutation rate increases. We also evaluated this integral approximately using a modification of Kramers' theory[3], although overall this does not improve on Kimura's theory.

In general, these results point to a new calculational technique, where Fisher's transformation converts classic problems in population genetics to one of simple Brownian motion in a potential, which from the literature in physics and chemistry[4,7] exist many approximate methods to find solutions.

REFERENCES

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

[2] B. S. Khatri, "Quantifying evolutionary dynamics from variant-frequency time series," Scientific Reports **6** (2016).

[3] H. A. Kramers, "Brownian motion in a field of force and the diffusion model of chemical reactions," Physica **7**, 284–304 (1940).

[4] C. Gardiner, *Stochastic Methods: A Handbook for the Natural and Social Sciences* (Springer, 2009).

[5] M. Kimura and T. Ohta, "The average number of generations until fixation of a mutant gene in a finite population." Genetics **61**, 763–771 (1969).

[6] R. A. Fisher, *The Genetical Theory of Natural Selection* (Oxford Univ. Press, Oxford, 1930).

[7] N. van Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, 1981).