DOI: 10.1534/genetics.104.027797

Note

A Novel Solution for the Time-Dependent Probability of Gene Fixation or Loss Under Natural Selection

Ying Wang and Bruce Rannala[1]

Department of Medical Genetics, University of Alberta, Edmonton, Alberta T6G 2H7, Canada

Manuscript received February 17, 2004
Accepted for publication July 7, 2004

ABSTRACT

Kimura (1955b) proposed a solution for the time-dependent probability of fixation, or loss, of a gene under selection. Example calculations suggest the formulas are prone to numerical inaccuracies. An alternative solution is proposed; the correctness of the solution is confirmed by numerically solving the Kolmogorov backward equation and by simulation.

THE process of change of the frequency of a gene over time in a large random-mating population can be treated as a stochastic process and approximated by a diffusion process. Kimura modeled the dynamic process of gene frequency change over time under different models for mutation and selection by making use of diffusion theory (Kimura 1954, 1955a,b, 1957, 1964, 1978). Empirical studies have suggested that many mutant disease-related alleles are under natural selection, for example, the MHC-HLA gene family (Dean *et al.* 2002). Studying the dynamics of a disease-related allele under natural selection is important for understanding the evolutionary history of a disease (Slatkin and Rannala 2000). In this article, we review the existing theory of gene frequency evolution and suggest that for a selected allele in the case of no dominance it is not possible to accurately calculate the time-dependent fixation and loss probabilities using Kimura's solution; there are no numerical examples of this problem in Kimura's article (Kimura 1955b, 1964). By further investigating properties of the boundary conditions of the diffusion process and the solution of the oblate spheroidal equation, we propose an alternative solution for the time-dependent probability of fixation, or loss, of an allele that can be accurately calculated by using other existing solutions (*e.g.*, the ultimate probability of fixation, or loss, of an allele). The correctness of our result was confirmed by simulation studies and by numerically solving the Kolmogorov backward equation.

Consider a large random-mating population; two alleles exist at a locus with a selectively advantageous allele with frequency p_0 at generation 0. The probability density that the frequency of the favored allele is x ($x \in (0, 1)$) at generation t, denoted by $\phi(x, t|p_0)$, satisfies the Kolmogorov forward equation (or Fokker-Planck equation). The average change of allele frequency per generation under selection (with selection coefficient $s > 0$) is approximately

$$\delta x = sx(1 - x), \tag{1}$$

according to a diffusion approximation if there is no dominance. The variance of the change of gene frequency due to random drift is $x(1 - x)/(2N)$ and $\phi(x, t|p_0)$ can be obtained by solving the following partial differential equation (PDE),

$$\frac{\partial\phi(x, t|p_0)}{\partial t} = \frac{1}{4N}\frac{\partial^2\{x(1 - x)\phi(x, t|p_0)\}}{\partial x^2} - \frac{\partial\{sx(1 - x)\phi(x, t|p_0)\}}{\partial x}, \tag{2}$$

with boundaries $x = 0$ and $x = 1$, where N is the Wright-Fisher population size. Kimura solved this PDE by using the separation-of-variables method (Kimura 1955b, 1957, 1964). Applying this method, $\phi(x, t|p_0)$ can be expressed as a product of a function of x and t alone. It has been found that the function of x alone has a similar form to an intermediate solution of the oblate spheroidal equation (Stratton *et al.* 1941, 1956). Kimura proposed the solution for Equation 2 on the basis of the theory of the oblate spheroidal equation,

$$\phi(x, t|p_0) = \sum_{k=1}^{\infty} C_k e^{-((\lambda_k + c^2)/4N)t + 2\alpha x} \sum_{n=0,1}^{\prime} f_n^k T_n^1(z), \tag{3}$$

where

$$C_k = \frac{(1 - r^2)e^{-c(1-r)}\Sigma'_{n=0,1} f_n^k T_n^1(r)}{\Sigma'_{n=0,1}((n + 1)(n + 2))/(2n + 3)(f_n^k)^2}, \tag{4}$$

[1]*Corresponding author:* Department of Medical Genetics, 8-39 Medical Sciences Bldg., University of Alberta, Edmonton, AB T6G2H7, Canada. E-mail: brannala@ualberta.ca

and $r = 1 - 2p_0$, $z = 1 - 2x$, $c = Ns$, $T_n^1(\cdot)$ is the Gegenbauer polynomial, λ_k is the eigenvalue of the oblate spheroidal angular function, and f_n^k is the intermediate coefficient of the spheroidal harmonics. Feller (1952) has classified the boundaries associated with the one-dimensional diffusion process into four general classes. According to Feller's classification criteria, the boundaries both belong to the exit boundaries. From the nature of the process, the boundaries $x = 0$ and $x = 1$ are absorbing barriers. It was shown that the probability mass functions of loss and fixation before time t, denoted by $f_0(t|p_0)$ and $f_1(t|p_0)$, respectively, satisfy

$$\frac{df_0(t|p_0)}{dt} = \lim_{x\to 0}\left\{\frac{1}{4N}\frac{\partial[x(1-x)\phi(x,t|p_0)]}{\partial x} - sx(1-x)\phi(x,t|p_0)\right\}, \tag{5}$$

$$\frac{df_1(t|p_0)}{dt} = -\lim_{x\to 1}\left\{\frac{1}{4N}\frac{\partial[x(1-x)\phi(x,t|p_0)]}{\partial x} - sx(1-x)\phi(x,t|p_0)\right\}, \tag{6}$$

for an absorbing barrier process (Bharucha-Reid 1960). After simplification, Equations 5 and 6 become

$$\frac{df_0(t|p_0)}{dt} = \frac{1}{4N}\phi(0,t|p_0), \quad \frac{df_1(t|p_0)}{dt} = \frac{1}{4N}\phi(1,t|p_0), \tag{7}$$

which were used by Wright (1931), Wright and Kerr (1954), and Kimura (1955b, 1964). The explicit solution of the probability of fixation, or loss, of an allele by generation t, given initial frequency p_0, can be derived by means of the above relations and Equation 3. The procedure has been used by Kimura (1955b, 1964). For example, the loss probability of an allele within t generations was derived as

$$f_0(t|p_0) = \sum_{k=1}^{\infty}\frac{C_k}{\lambda_k + c^2}(1 - e^{-((\lambda_k+c^2)/4N)t})\sum_{n=0,1}^{'}\frac{(n+1)(n+2)}{2}f_n^k. \tag{8}$$

If we instead write the above equation in the form $f_0(t|p_0) = \Sigma_{k=1}^{\infty}A_k(1 - e^{-((\lambda_k+c^2)/4N)t})$, we can see that if t/N is not too small (*e.g.*, $t/N \geq 0.5$), $\Sigma_{k=1}^{\infty}A_k e^{-((\lambda_k+c^2)/4N)t}$ rapidly converges as the eigenvalue of the oblate spheroidal angular function, λ_k, increases with increasing k and $e^{-((\lambda_k+c^2)/4N)t}$ is then close to 0, even for small k. However, the term $\Sigma_{k=1}^{\infty}A_k$ converges very slowly, and finite approximations of this term turn out to be inaccurate if a finite k is used for numerical calculations. An alternative approach is to obtain the probability of fixation, or loss, of an allele after time t by evaluating the integral over time from t to ∞. The probability of fixation, or loss, of a gene before t can then be obtained by using the stationary fixation (loss) probability of an allele under selection and subtracting the fixation (loss) probability of the allele after time t. As was pointed out, one or the other boundary is eventually reached, and the probability of reaching either boundary was derived early on (Ewens 1979). Using this approach, the time-dependent loss, and fixation, probabilities can be written as

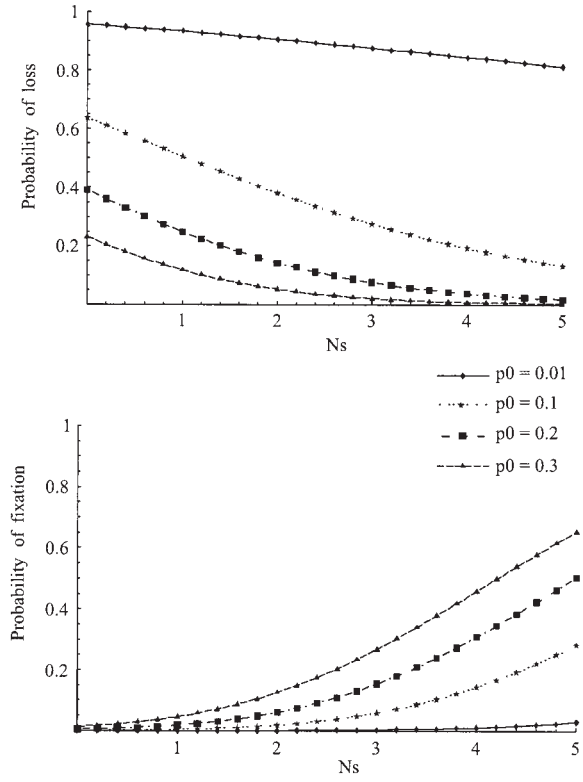

Figure 1.—The probability of loss, or fixation, of an allele at time $t = 10^3$ generations under different selection intensities, s, given various initial frequencies of the allele, p_0, calculated using Equations 9 and 10. The population size $N = 10^3$.

$$f_0(t|p_0) = 1 - \frac{1 - e^{-4cp_0}}{1 - e^{-4c}} - \sum_{k=1}^{\infty}\frac{C_k}{\lambda_k + c^2}e^{-((\lambda_k+c^2)/4N)t}\sum_{n=0,1}^{'}\frac{(n+1)(n+2)}{2}f_n^k, \tag{9}$$

and

$$f_1(t|p_0) = \frac{1 - e^{-4cp_0}}{1 - e^{-4c}} - \sum_{k=1}^{\infty}(-1)^k\frac{C_k}{\lambda_k + c^2}e^{-((\lambda_k+c^2)/4N)t+2c}\sum_{n=0,1}^{'}\frac{(n+1)(n+2)}{2}f_n^k, \tag{10}$$

respectively. Note that in recent years more efficient algorithms have been proposed for calculating the eigenvalues λ_k and coefficient f_n^k of the spheroidal harmonics (Li 1998). Equations 9 and 10 above provide accurate numerical results for these probabilities. The examples in Figure 1, calculated using the above equations, illustrate the effect of natural selection on the loss and fixation probability of an allele at time t for different initial allele frequencies.

Using the solution for the fixation and loss probabilities, one can examine related questions about the effects of natural selection on allele frequency distributions

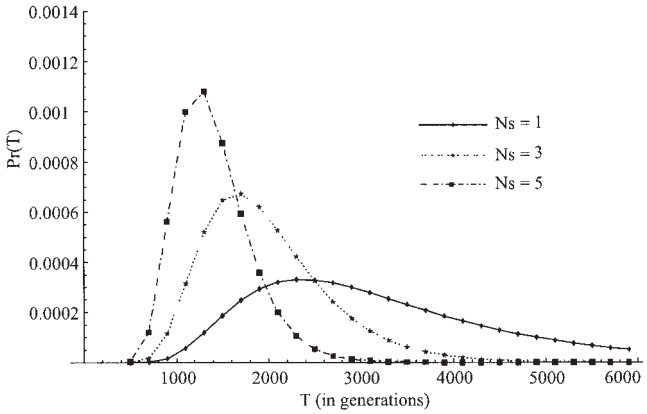

FIGURE 2.—Probability distributions of allele age T, given the initial frequency of the allele $p_0 = 1/(2N)$ and the allele is being fixed, under different selection intensity s, calculated using Equation 11. The population size $N = 10^3$.

that have not been studied previously. For example, the probability density function of the length of the time until fixation of a selected allele, when only cases in which an allele is ultimately fixed are taken into account, can be obtained by differentiating the cumulative distribution of the time until fixation with respect to t. This turns out to be

$$y(t|p_0) = \frac{1 - e^{-4c}}{1 - e^{-4cp_0}} \sum_{k=1}^{\infty} (-1)^k \frac{C_k}{4N} e^{-((\lambda_k + c^2)/4N)t + 2c} \sum_{n=0,1}^{'} \frac{(n+1)(n+2)}{2} f_n^k. \quad (11)$$

Numerical examples calculated using the above equation are demonstrated in Figure 2. We can see from this figure that the distribution of the time required for fixation is broad and becomes narrower as selection intensities increase. For a selectively neutral allele, the above equation can be used by simply letting $\lim_{c \to 0} \{y(t|p_0)\}$, and numerical results are compared with results calculated previously for this case by using the solution of the probability distribution of the length of time until fixation of a selectively neutral allele derived by KIMURA (1970). The results calculated using the two equations agree closely.

Previous studies have shown that the loss and fixation probabilities satisfy the Kolmogorov backward equation. The correctness of the novel solutions for the absorption probabilities by generation t presented above was examined by comparing numerical results obtained using these solutions with the results obtained by numerically solving the corresponding backward Kolmogorov equations and with results obtained by computer simulations.

Numerical solution of the PDE: To obtain the probabilities of fixation and loss, one approach is to make use of the backward Kolmogorov equation. The basic difference between the forward and backward equations is the variable that is being fixed. This method was used by Kimura to study the neutral case (KIMURA 1964) and the solutions have been verified by simulation studies. For the selection model we discuss, the probability that a favored allele with initial frequency p_0 is fixed by generation t, denoted by $u(p_0, t)$, satisfies

$$\frac{\partial u(p_0, t)}{\partial t} = \frac{p_0(1 - p_0)}{4N} \frac{\partial^2 u(p_0, t)}{\partial p_0^2} + sp_0(1 - p_0) \frac{\partial u(p_0, t)}{\partial p_0}. \quad (12)$$

Note that $u(p_0, t)$ is equivalent to $f_1(t|p_0)$ but we use different notation to distinguish the backward approach. The boundary conditions are

$$u(1, t) = 1, \quad u(0, t) = 0 \quad (13)$$

for fixation probability. The loss probability also satisfies Equation 12 but with boundary conditions opposite to those of Equation 13. The initial condition for both equations is

$$u(x, t_0|x_0) = \delta(x - x_0), \quad (14)$$

given that the initial frequency of the allele is x_0 at initial time t_0. We solved Equation 12 with boundary conditions (13) and initial condition (14) numerically by the implicit finite difference method (*Crank-Nicolson* scheme).

Simulation studies: To confirm the accuracy of the numerical solutions of the absorption probabilities and compare these with the solutions calculated using the

TABLE 1

Comparison of numerical results of ANA, NUM, SIM, and KIM

c (sN)	p_0	Probability of	ANA	NUM	SIM	KIM
1	0.0005	Fixation	8.07331×10^{-6}	8.14065×10^{-6}	0	−0.03254
		Loss	0.99668	0.99668	0.99684	0.03798
	0.1	Fixation	0.00466	0.00469	0.00481	0.61783
		Loss	0.50691	0.50694	0.50759	0.43346
3	0.0005	Fixation	0.00017	0.00017	0.00011	−1.77029
		Loss	0.99362	0.99362	0.99359	0.03883
	0.1	Fixation	0.05873	0.05903	0.05942	20.8208
		Loss	0.27622	0.27576	0.27547	0.20951

For these results, population size (N) was set to be 10^3, and the age of the allele (t) was 10^3 generations. The number of iterations was 10^5 in simulation tests.

above analytic equations, the sample path (over time) of the population frequency of a selected allele was simulated by a modified "pseudo-sampling variable (PSV)" method. The PSV method was suggested by Kimura and was used for the simplest neutral models (Kimura 1980; Kimura and Takahata 1983). The method can be easily extended to models taking account of selection and mutation. The gist of the method is to simulate the diffusion process itself rather than simulating the entire population of alleles at each generation under a binomial distribution for the case of two alleles or a multinomial distribution for three or more alleles. For the selection model discussed above, $E[\delta x] = sx(1 - x)$, $\mathrm{Var}[\delta x] = x(1 - x)/(2N)$, and the frequency of the allele at the next generation x', given the current frequency x, is simulated by

$$x' = x + \xi_{\mathrm{PSV}}, \tag{15}$$

where ξ_{PSV} is a uniform random variable with mean $sx(1 - x)$ and variance $x(1 - x)/(2N)$. We fixed the age of the allele and simulated the sample path of allele frequency for a large number of iterations, tabulating the proportion of simulated populations in which the allele was fixed, or lost, by time t (SIM), and comparing this with the expected probabilities obtained by numerically solving the PDE (NUM) and with the results obtained using Equations 9 and 10 (ANA). The results calculated using Kimura's solutions (KIM) were also included. For both ANA and KIM, the sums of the equations were evaluated to $k = 8$. Results are listed in Table 1. The examples in Figure 2 were also verified by the simulations.

The probability distribution of an allele, taking account of random drift and natural selection, was studied by Kimura under diffusion theory. The complexity of Kimura's analytical solutions make them difficult to use, except in the case that only random drift is considered. By calculating the probabilities of fixation and loss of an allele using the equations given by Kimura (*e.g.*, Equation 8), we found the results were invariably quite inaccurate even though the sum of Equation 8 was evaluated to a relatively large k, and $t/N \geq 1$ was used. An alternative solution is proposed, which provides much more accurate results even when a small k is chosen. For example, in Table 1, when $t/N = 1$ and k was set to be 8, the numerical results agree very closely with those obtained numerically and by simulation.

We thank Warren Ewens for helpful comments. Support was provided for this research by Canadian Institutes of Health Research grant MOP 44064 and National Institutes of Health grant HG01988 to B.R.

LITERATURE CITED

Bharucha-Reid, A. T., 1960 *Elements of the Theory of Markov Processes and Their Applications,* McGraw-Hill, New York.

Dean, M., M. Carrington and S. J. O'Brien, 2002 Balanced polymorphism selected by genetic versus infectious human disease. Annu. Rev. Genomics Hum. Genet. **3:** 263–292.

Ewens, W. J., 1979 *Mathematical Population Genetics.* Springer-Verlag, Berlin.

Feller, W., 1952 The parabolic differential equations and the associated semi-groups of transformations. Ann. Math. **55:** 468–519.

Kimura, M., 1954 Process leading to quasi-fixation of genes in natural populations due to random fluctuation of selection intensities. Genetics **39:** 280–295.

Kimura, M., 1955a Solution of a process of random genetic drift with a continuous model. Proc. Natl. Acad. Sci. USA **41:** 144–150.

Kimura, M., 1955b Stochastic processes and distribution of gene frequencies under natural selection. Cold Spring Harbor Symp. Quant. Biol. **20:** 33–53.

Kimura, M., 1957 Some problems of stochastic processes in genetics. Ann. Math. Stat. **28:** 882–901.

Kimura, M., 1964 Diffusion models in population genetics. J. Appl. Probab. **1:** 177–232.

Kimura, M., 1970 The length of time required for a selectively neutral mutant to reach fixation through random frequency drift in a finite population. Genet. Res. **15:** 131–133.

Kimura, M., 1978 Change of gene frequencies by natural selection under population number regulation. Proc. Natl. Acad. Sci. USA **75:** 1934–1937.

Kimura, M., 1980 Average time until fixation of a mutant allele in a finite population under continued mutation pressure: studies by analytical, numerical, and pseudo-sampling methods. Proc. Natl. Acad. Sci. USA **77:** 522–526.

Kimura, M., and N. Takahata, 1983 Selective constraint in protein polymorphism: study of the effectively neutral mutation model by using an improved pseudosampling method. Proc. Natl. Acad. Sci. USA **80:** 1048–1052.

Li, L. W., M. S. Leong, T. S. Yeo, P. S. Kooi and K. Y. Tan, 1998 Computations of spheroidal harmonics with complex arguments: a review with an algorithm. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. **58:** 6792–6806.

Slatkin, M., and B. Rannala, 2000 Estimating allele age. Annu. Rev. Genomics Hum. Genet. **1:** 225–249.

Stratton, J. A., P. M. Morse, L. J. Chu and R. A. Hutner, 1941 *Elliptic Cylinder and Spheroidal Wave Functions.* John Wiley, New York.

Stratton, J. A., P. M. Morse, L. J. Chu, J. Little and F. J. Corbato, 1956 *Spheroidal Wave Functions.* The Technology Press of MIT and John Wiley, New York.

Wright, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97–159.

Wright, S., and W. E. Kerr, 1954 Experimental studies of the distribution of gene frequencies in very small populations of *Drosophila Melanogaster.* II. Bar. Evolution **8:** 225–240.

Communicating editor: M. K. Uyenoyama