Hindawi Publishing Corporation
Advances in Bioinformatics
Volume 2008, Article ID 257864, 9 pages
doi:10.1155/2008/257864

Review Article

A Tutorial of the Poisson Random Field Model in Population Genetics

Praveen Sethupathy[1] and Sridhar Hannenhalli[1, 2]

[1] *Department of Genetics, School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA*
[2] *Department of Computer and Information Sciences, School of Engineering and Applied Sciences, University of Pennsylvania, Philadelphia, PA 19104, USA*

Correspondence should be addressed to Praveen Sethupathy, praveens@mail.med.upenn.edu

Received 6 February 2008; Accepted 15 May 2008

Recommended by Alexander Zelikovsky

Population genetics is the study of allele frequency changes driven by various evolutionary forces such as mutation, natural selection, and random genetic drift. Although natural selection is widely recognized as a bona-fide phenomenon, the extent to which it drives evolution continues to remain unclear and controversial. Various qualitative techniques, or so-called “tests of neutrality”, have been introduced to detect signatures of natural selection. A decade and a half ago, Stanley Sawyer and Daniel Hartl provided a mathematical framework, referred to as the Poisson random field (PRF), with which to determine quantitatively the intensity of selection on a particular gene or genomic region. The recent availability of large-scale genetic polymorphism data has sparked widespread interest in genome-wide investigations of natural selection. To that end, the original PRF model is of particular interest for geneticists and evolutionary genomicists. In this article, we will provide a tutorial of the mathematical derivation of the original Sawyer and Hartl PRF model.

1. Introduction

Selectionists and neutralists have fiercely debated, for the past five decades, the extent to which Darwinian selection has shaped molecular evolution. However, both camps do agree that Darwinian selection is a bona fide natural phenomenon. Therefore, various so-called “tests of neutrality” have been developed to detect natural selection on a particular gene or genomic location (for a review on this topic, see [1]). However, these tests are often qualitative and only provide the directionality of selection. A decade and a half ago, S. Sawyer and D. Hartl provided a mathematical framework with which to determine quantitatively the intensity of selection on a particular gene, which they applied to the *Adh* locus in the *Drosophila* genome [2]. This framework is referred to as the Poisson random field (PRF) model. They then further used this framework to analyze codon bias in enteric bacteria [3]. Owing to the recent availability of whole genome sequences and genome-wide human polymorphism data, it has become increasingly tractable to perform genome-wide scans for signatures of selection. The PRF model has been applied to estimate the intensity of selection on synonymous and nonsynonymous sites throughout mitochondrial and nuclear genomes of a variety of species, including human [4–12]. Very recently, due to the advent of high-throughput experimental and computational identification of genomic regulatory elements, there has been an interest to estimate the intensity of natural selection on regulatory mutations. Chen and Rajewsky [13] use the PRF, among other techniques, to provide evidence for purifying selection (even stronger than on nonsynonymous coding sites) on a class of regulatory sites known as microRNA target sites. Due to the potentially wide range of applications of, and opportunities for theoretical extensions to, the PRF model, it is an increasingly important mathematical framework for quantitative geneticists. In this article, we will provide a tutorial of the mathematical derivation of the basic PRF model that was originally developed in [2]. The tutorial will follow the outline provided below:

(i) Wright-Fisher model,

(ii) diffusion approximation to the Wright-Fisher model,

(iii) derivation, via diffusion theory, of formulas describing evolutionary processes of interest,

(iv) derivation of the PRF using the above-mentioned formulas.

The first three items are discussed in [14], and the last point was originally presented in [2]. In this tutorial, we aim to provide an integrated and comprehensive presentation that is accessible to nonprofessionals or beginners in the field of population genetics. Since the primary purpose is to review mathematical derivations, familiarity with calculus and at least a cursory knowledge of genetics will be helpful for the reader.

2. The Wright-Fisher Model

The Wright-Fisher (WF) model describes the change in frequency of a single mutation (derived allele) in a population over time. The simplest version of the model makes the following assumptions: (1) nonoverlapping generations, (2) constant population size in each generation, and (3) random mating, and is described as follows.

Consider a population of N diploid individuals that has a single polymorphic site with two alleles, one ancestral and one derived. Under this model, the frequency of the derived allele in the current generation is a function of the selection pressure on this allele and the binomial sampling effect with probabilities proportional to the frequency of this allele in the previous generation. The probability, p_{ij}, that there are j genes of the derived allele present at generation G + 1 given i genes of the derived allele present at generation G is given by the following binomial calculation:

$$p_{ij} = \binom{2N}{j} (\Psi_i)^j (1-\Psi_i)^{2N-j}, \tag{1}$$

where Ψ_i depends on the relative fitness of the derived allele.

Assuming no dominance and no recurrent mutation,

$$\Psi_i = \frac{x(1+s)}{x(1+s)+(1-x)}, \tag{2}$$

where $1 + s$ is the fitness of the derived allele relative to 1 for the ancestral allele, and x (which is simply $i/2N$) is the derived allele frequency (daf) in generation G. In the simplest model (no selection and no recurrent mutation), Ψ_i is simply x or $i/2N$.

The intuition behind Ψ_i is the following. Consider the scenario where both the ancestral and the derived alleles are neutrally evolving (no or negligible selection pressure). In this case, the probability of sampling a gene of the derived allele from the population in generation G is simply the frequency of the derived allele in generation G, $i/2N$ or x. This can be rewritten as $x/[x+(1-x)]$. Now, suppose that the derived allele is under some selection, s, meaning that the fitness of the derived allele is $1 + s$ relative to 1 for the ancestral allele. In this case, genes are sampled according to their relative fitnesses (as in the equation for Ψ_i above). Figure 1(a) provides a pictorial representation of the basic Wright-Fisher model.

3. Diffusion Theory

We define $p_{ki}^{(t)}$ as the probability that a polymorphic site has i genes of the derived allele at time t, given that it had k genes of the derived allele at time 0. $p_{ki}^{(t)}$ satisfies the following:

$$p_{kj}^{(t+1)} = \sum_i p_{ki}^{(t)} p_{ij}, \tag{3}$$

where p_{ij} is given in (1).

It is convenient to change notation and write $p_{ki}^{(t)}$ as $f(x;p,t)$, so that the above becomes

$$f(j;\, k, t+1) = \sum_i f(i;\, k, t)\, p_{ij}. \tag{4}$$

In this framework, it has been shown to be extremely difficult to explicitly derive formulas for several quantities of evolutionary interest. However, as the size of the population approaches infinity (i.e., $N \to \infty$), and assuming that the scaled selection pressure (Ns) and scaled mutation rate ($N\mu$) remain constant, the discrete Markov process given above can be closely approximated by a continuous-time, continuous-space diffusion process (Figure 1(b)):

$$f(x+\delta x;\, p, t+\delta t) = \int_0^1 f(y;\, p, t) f(x+\delta x;\, y, \delta t)\, dy, \tag{5}$$

where $f(x;\, p, t)$ is the probability distribution of x at time t, x is the daf at time t, p is the daf at time 0, and δx is the daf change in time δt.

We can perform a Taylor series expansion on both sides in δt and δx to derive the forward Kolmogorov equation:

$$\frac{\partial f(x;p,t)}{\partial t} = \frac{\partial^2[b(x)f(x;p,t)]}{2\partial x^2} - \frac{\partial[a(x)f(x;p,t)]}{\partial x}, \tag{6}$$

where

$$\begin{aligned} E(\Delta x) &\approx a(x)\, dt, \\ \text{var}(\Delta x) &\approx b(x)\, dt, \end{aligned} \tag{7}$$

and $a(x)$ and $b(x)$ depend on the genetic model (e.g., see eq (24).

Equation (5) can be represented diagrammatically as in Figure 2. The probability of derived allele frequency $x+\delta x$ at time $t+\delta t$ is the product of the probability of moving from p to x in time t and the probability of moving from x to $x+\delta x$ in time δt, summed over all possible values of x.

The frequency trajectory of a derived allele can also be depicted as in Figure 3, which illustrates that the probability of frequency x at time $t+\delta t$ is the product of the probability of moving from p to $p+\delta p$ in time δt and the probability of moving from $p+\delta p$ to x in time t, summed over all possible values of δp. This is formalized as follows:

$$f(x;\, p, t+\delta t) = \int_0^1 f(p+\delta p;\, p, \delta t) f(x;\, p+\delta p, t)\, d(\delta p). \tag{8}$$

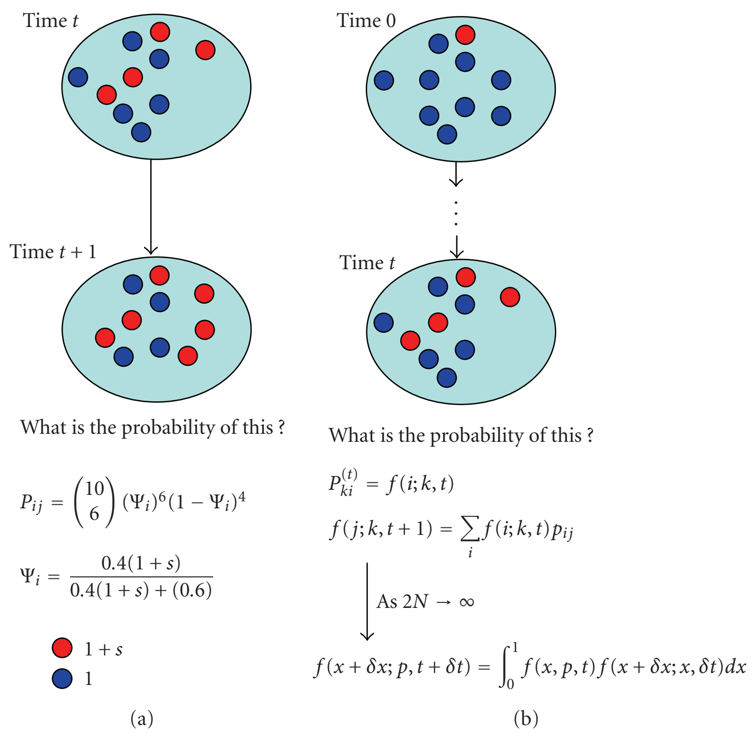
Time t
Time $t+1$
What is the probability of this ?
$P_{ij} = \binom{10}{6} (\Psi_i)^6 (1-\Psi_i)^4$
$\Psi_i = \dfrac{0.4(1+s)}{0.4(1+s)+(0.6)}$
● $1+s$
● 1
(a)
Time 0
Time t
What is the probability of this ?
$P_{ki}^{(t)} = f(i;k,t)$
$f(j;k,t+1) = \sum_i f(i;k,t)p_{ij}$
As $2N \to \infty$
$f(x+\delta x;p,t+\delta t) = \int_0^1 f(x,p,t)f(x+\delta x;x,\delta t)dx$
(b)

FIGURE 1: Pictorial representation of the Wright-Fisher process and its diffusion approximation: (a) basic Wright-Fisher model assuming selection, but no dominance or recurrent mutation and (b) diffusion approximation to the basic Wright-Fisher model.

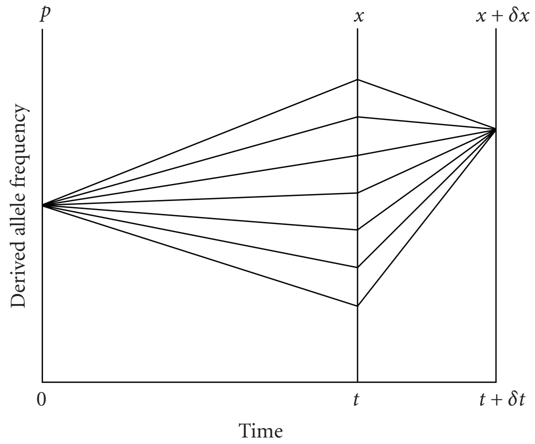

FIGURE 2: A diagrammatic intuition for (3) illustrates that the probability of derived allele frequency $x + \delta x$ at time $t + \delta t$ is the product of the probability of moving from p to x in time t and the probability of moving from x to $x + \delta x$ in time δt, summed over all possible values of x.

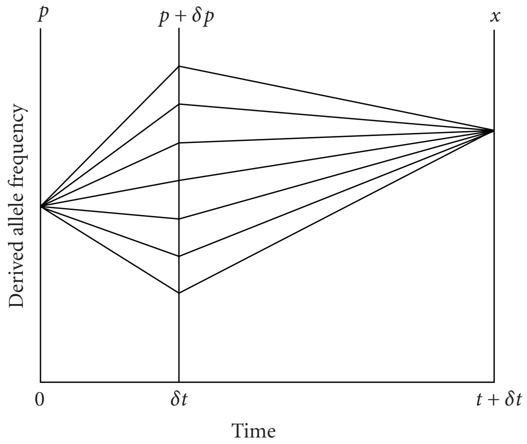

FIGURE 3: A diagrammatic intuition for (5) illustrates that the probability of frequency x at time $t + \delta t$ is the product of the probability of moving from p to $p + \delta p$ in time δt and the probability of moving from $p + \delta p$ to x in time t, summed over all possible values of δp.

We can again perform a Taylor series expansion on both sides to derive the backward Kolmogorov equation:

$$\frac{\partial f(x;p,t)}{\partial t} = b(p)\frac{\partial^2[f(x;p,t)]}{2\partial p^2} + a(p)\frac{\partial[f(x;p,t)]}{\partial p}. \tag{9}$$

The forward and backward Kolmogorov equations have played a central role in theoretical population genetics since 1922. For details regarding their derivation, we refer the reader [15, Chapter 4]. Next, we will discuss how they are utilized to derive formulas for various quantities of evolutionary interest (yellow boxes in Figure 4).

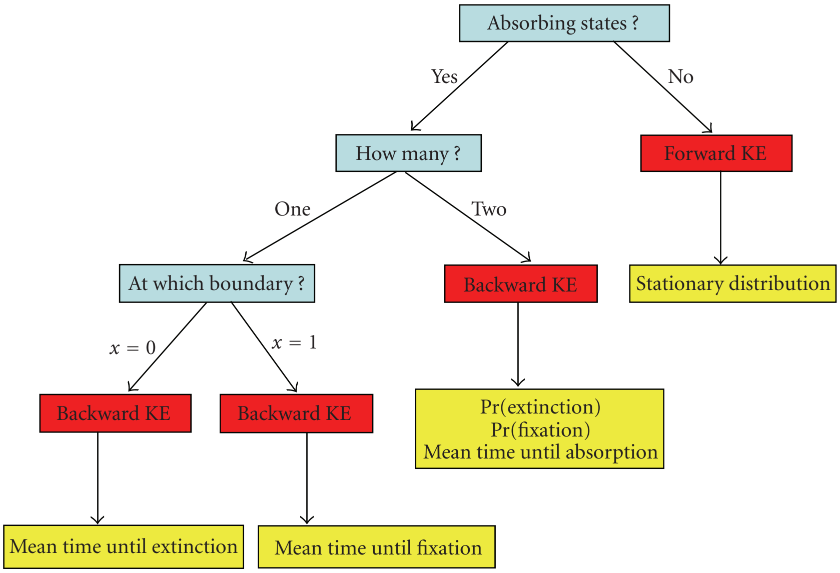

Figure 4: The utility of the Kolmogorov equations for studying evolutionary processes. Blue boxes correspond to questions that clarify the assumptions of the genetic model being used, the red boxes correspond to when the Kolmogorov equations (KEs) are utilized, and yellow boxes correspond to quantities of evolutionary interest.

In a model where there is two-way recurrent mutation (i.e., there are no absorbing states, either extinction or fixation), stationarity is achieved when the probability of change in the derived allele frequency is no longer dependent on time t. We solve for the stationary distribution, *f(x)*, in the following manner. First, we integrate through the forward Kolmogorov equation with respect to x:

$$\frac{\partial F(x;p,t)}{\partial t} = \frac{\partial\left[b(x)f(x;p,t)\right]}{2\partial x} - \left[a(x)f(x;p,t)\right], \quad (10)$$

$$F(x;p,t) = \int_0^x f(y;p,t)\,dy, \quad (11)$$

where $F(x;p,t)$ is the probability of the derived allele assuming a frequency between 0 and x at time t. Therefore, the derivative of $F(x;p,t)$ with respect to t can be interpreted as the probability flux (change in probability over time) of the diffusion process. The stationary distribution, *f(x)*, can be solved by setting the probability flux equal to zero.

4. Derivation of Formulas Describing Evolutionary Processes of Interest

Let us now focus on a genetic model that assumes no recurrent mutation (i.e., two absorbing states, one at $x = 0$ and another at $x = 1$). As depicted by Figure 4, in such a model, it is possible to determine the probability of extinction ($x = 0$), the probability of fixation ($x = 1$), and the mean time until absorption (either at $x = 0$ or $x = 1$) by using the Kolmogorov backward equation (Figure 4). It is also possible to derive the mean time until absorption conditioned on always eventually reaching only one of the two states. Since this quantity is not directly applicable to the PRF, we do not review its derivation here, but instead refer the reader to [14].

4.1. Probability of Extinction

Using (11), we arrive at an equation parallel to (9):

$$\frac{\partial F(x;p,t)}{\partial t} = b(p)\frac{\partial^2\left[F(x;p,t)\right]}{2\partial p^2} + a(p)\frac{\partial\left[F(x;p,t)\right]}{\partial p}. \quad (12)$$

The probability that the derived allele frequency, x, reaches 0 at or before time t follows from (11) and is given by

$$P_0(p,t) = \int_0^{0^+} f(y;p,t)\,dy = F(0^+;p,t), \quad (13)$$

where p is the initial frequency of the derived allele and 0^+ indicates $0 + \varepsilon$, where ε is very small.

Replacing $F(0^+;p,t)$ with $P_0(p,t)$, (12) can be written as

$$\frac{\partial P_0(p,t)}{\partial t} = b(p)\frac{\partial^2\left[P_0(p,t)\right]}{2\partial p^2} + a(p)\frac{\partial\left[P_0(p,t)\right]}{\partial p}. \quad (14)$$

As $t \to \infty$, $P_0(p,t)$ can be interpreted as the probability that extinction ever occurs (independent of time) and can be rewritten in the form $P_0(p)$. From (14), it is evident that $P_0(p)$ satisfies the following equation:

$$0 = b(p)\frac{\partial^2\left[P_0(p)\right]}{2\partial p^2} + a(p)\frac{\partial\left[P_0(p)\right]}{\partial p}. \quad (15)$$

Solving (15), we arrive at the following:

$$P_0(p) = \frac{\int_p^1 \psi(y)\,dy}{\int_0^1 \psi(y)\,dy}, \quad (16)$$

where

$$\psi(y) = e^{-2\int^y [a(z)/b(z)]\,dz} \quad (17)$$

and where *a(z)* and *b(z)* are defined as in (6).

4.2. Probability of Fixation

The probability that the derived allele frequency, x, reaches 1 at time t follows from (11) and is given by

$$\begin{aligned} P_1(p,t) &= \int_{1^-}^{1} f(y;p,t)\,dy \\ &= 1 - \int_{0}^{1^-} f(y;p,t)\,dy = 1 - F(1^-;p,t), \end{aligned} \tag{18}$$

where p is the initial frequency of the derived allele and 1^- indicates $1-\varepsilon$, where ε is very small.

In (12), $F(x;p,t)$ can be replaced by $1-F(x;p,t)$ without any loss of generality. Also, by replacing $1-F(1^-;p,t)$ with $P_1(p,t)$, (12) can be rewritten as

$$\frac{\partial P_1(p,t)}{\partial t} = b(p)\frac{\partial^2[P_1(p,t)]}{2\partial p^2} + a(p)\frac{\partial[P_1(p,t)]}{\partial p}. \tag{19}$$

By letting $t\to\infty$ and solving for $P_1(p)$, we arrive at the following:

$$P_1(p) = \frac{\int_0^p \psi(y)\,dy}{\int_0^1 \psi(y)\,dy}, \tag{20}$$

where $\psi(y)$ has been defined in (17) and *a(z)* and *b(z)* have been defined in (6).

The probability of fixation and the probability of extinction must sum to 1. Using (16) and (20), we can verify that this is indeed the case.

Consider a genetic model that assumes the presence of selection, but no recurrent mutation, where $a(x) = sx(1-x)$ and $b(x) = x(1-x)/2N$. Starting from (20), we can express the probability of fixation under this genetic model in the following manner:

$$\begin{aligned} P_1(p) &= \frac{\int_0^p e^{-2\int^y [a(z)/b(z)]\,dz} dy}{\int_0^1 e^{-2\int^y [a(z)/b(z)]\,dz} dy} \\ &= \frac{\int_0^p e^{-4Nsy} dy}{\int_0^1 e^{-4Nsy} dy} = \frac{1-e^{-4Nsp}}{1-e^{-4Ns}}. \end{aligned} \tag{21}$$

4.3. Mean Time Until Either Extinction or Fixation

We define $\phi(p,t)$ to be the density function of the time t at which absorption occurs. The probability that absorption occurs, at either boundary $x=0$ or $x=1$, by time t, is

$$P_0(p,t) + P_1(p,t) = \int_0^t \phi(p,t)\,dt. \tag{22}$$

Furthermore, since absorption must happen by $t = \infty$, we know that

$$\int_0^\infty \phi(p,t)\,dt = 1. \tag{23}$$

Performing integration by parts, we get the following:

$$-1 = -[t\phi(p,t)]_0^\infty + \int_0^\infty t\frac{\partial\phi(p,t)}{\partial t}\,dt. \tag{24}$$

Equations (14), (19), and (22) show that $\phi(p,t)$ satisfies the following equation:

$$\frac{\partial\phi(p,t)}{\partial t} = b(p)\frac{\partial^2[\phi(p,t)]}{2\partial p^2} + a(p)\frac{\partial[\phi(p,t)]}{\partial p}. \tag{25}$$

Using (25) and the fact that $\phi(p,t)$ approaches 0 faster than t approaches ∞, we can rewrite (24) as

$$-1 = 0 + \int_0^\infty t\left[b(p)\frac{\partial^2[\phi(p,t)]}{2{*}\partial p^2} + a(p)\frac{\partial[\phi(p,t)]}{\partial p}\right]dt. \tag{26}$$

After interchanging the order of integration and differentiation we get

$$-1 = b(p)\frac{d^2\bar{t}(p)}{2{*}dp^2} + a(p)\frac{d\,\bar{t}(p)}{dp}, \tag{27}$$

where

$$\begin{aligned} \bar{t}(p) &= \text{mean time until absorption} \\ &= \int_0^\infty t\phi(p,t)\,dt = \int_0^1 t(p,x)\,dx \end{aligned} \tag{28}$$

and $t(p,x)dx$ is the mean time that the daf spends in the interval $(x, x+\delta x)$ before absorption occurs.

We are interested in the case, where $p = 1/2N$, since this is the initial frequency of the derived allele. In this case, we are interested only in values of x greater than $1/2N$, and for these values we can write

$$t(p,x) = \frac{2P_1(p)\int_x^1 \psi(y)\,dy}{b(x)\psi(x)}, \tag{29}$$

and $\psi(x)$ is defined in (17).

Under the simplest genetic model that assumes no selection and no recurrent mutation, we can set $s=0$ in (17) and (21) and show that $P_1(p)$ reduces to p and $\psi(y)$ reduces to 1. It follows from this that (29) can be reduced to

$$t(p,x) = \frac{2p(1-x)}{x(1-x)/2N} = \frac{4Np}{x}. \tag{30}$$

Under a genetic model where $s \neq 0$, using $\gamma = 2Ns$, (29) can be rewritten as

$$t(p,x) = 2N\frac{\left(2\left(1-e^{-2\gamma p}\right)/\left(1-e^{-2\gamma}\right)\right)\int_x^1 e^{-2\gamma y}\,dy}{x(1-x)\left(e^{-2\gamma x}\right)}. \tag{31}$$

After integrating and simplifying the terms, we obtain

$$t(p,x) = 2N\frac{\left(1-e^{-2\gamma p}\right)\left(1-e^{-2\gamma(1-x)}\right)}{\left[\gamma\left(1-e^{-2\gamma}\right)\right]\left[x(1-x)\right]}. \tag{32}$$

Finally, substituting $\gamma = 2Ns$ and $p = 1/2N$, and invoking the approximation $e^{-a} = (1-a)$ for small values of a, $t(p,x)$ reduces approximately to

$$f(x) = t(p,x) \approx \frac{2\left(1-e^{-2\gamma(1-x)}\right)}{\left[\left(1-e^{-2\gamma}\right)\right]\left[x(1-x)\right]}, \tag{33}$$

where $f(x)dx$ is a notation common in the literature to represent the expected time for which the population frequency of a derived allele is in the range $(x, x+dx)$ before eventual absorption.

5. Poisson Random Field Theory

S. Sawyer and D. Hartl expanded the modeling of site evolution to multiple sites. Their model makes the following assumptions: (1) mutations arise at Poisson times, (2) each mutation occurs at a new site (infinite sites, irreversible), and (3) each mutant follows an independent WF process (no linkage). Sawyer and Hartl noticed from $f(x)$ in (33), that

$$\int_{x_1}^{x_2} \theta f(x)\,dx = \int_{x_1}^{x_2} g(x)\,dx \tag{34}$$

is the expected number of sites in the population with derived allele frequency between x_1 and x_2 (where θ equals $2N\mu$, the per-locus mutation rate). The function $g(x)$, for which the full expression is given below, is also referred to in the literature as the limiting, equilibrium, or expected density function for derived allele frequencies.

$$g(x) = \theta\frac{2\left(1-e^{-2\gamma(1-x)}\right)}{\left[\left(1-e^{-2\gamma}\right)\right]\left[x(1-x)\right]} = 4N\mu\frac{1-e^{-2\gamma(1-x)}}{\left(1-e^{-2\gamma}\right)x(1-x)}. \tag{35}$$

In a sample of size n, the expected number of sites with i (which ranges from 1 to $n-1$) copies of the derived allele is defined as a function of $g(x)$:

$$F(i) = \int_0^1 g(x)\,P(i \mid x)\,dx = \int_0^1 g(x)\binom{n}{i}x^i(1-x)^{n-i}\,dx. \tag{36}$$

The intuition behind *F(i)* is the following. The expected number of polymorphic sites with population daf x that have i copies of the derived allele out of n samples is given by the product of the expected number of sites with population daf x, $g(x)$, and the probability that each of those sites has i copies in the sample, which is given by the binomial calculation in the right-hand side of (36). To determine the expected number of sites with *any* population daf that have i copies of the derived allele, this product must be integrated over all possible values of x (resulting in *F(i)* above).

Consider the sample data $X = (X_1, X_2, X_3, \ldots, X_{n-1})$, where X_i is the observed number of sites with i copies of the derived allele out of n. Sawyer and Hartl showed that the number of derived alleles in the entire population at a particular frequency is a PRF with mean density given by (35) [2]. It follows, from the marking theorem on Poisson processes [16], that each random variable X_i is an independent Poisson distribution with mean equal to *F(i)* [2]. This framework allows us to define the probability of observing x_i sites that have i copies of the derived allele (and $n-i$ copies of the ancestral allele) as the following:

$$P(X_i = x_i \mid \theta, \gamma) = \frac{e^{-F(i)}F(i)^{x_i}}{x_i!}. \tag{37}$$

Since the X_i's are independent, the probability of observing $X = (X_1, X_2, X_3, \ldots, X_{n-1})$ is given as

$$P(X) = L(\theta, \gamma) = \prod_{i=1}^{n-1} P(X_i = x_i \mid \theta, \gamma). \tag{38}$$

The likelihood equation above provides a convenient means of estimating the values of the parameters θ and γ. The use of the PRF theory leads directly to a likelihood-ratio test of neutrality. Λ is defined as the ratio of the likelihood value under the maximum likelihood estimate of γ to the likelihood value under the neutral value of γ. It is a standard result that $2\ln\Lambda$ is asymptotically chi-square distributed with one degree of freedom [17].

Sawyer and Hartl further extended the PRF model in order to calculate the ratio of expected number of polymorphisms within species to expected number of fixed differences between species.In 1991, McDonald and Kreitman devised a 2-by-2 contingency table test of neutrality that was later named the MK test [18]. In the traditional MK test, a 2-by-2 contingency table is formed in order to compare the number of nonsynonymous and synonymous sites that are polymorphic within a species (RP and SP) and diverged between species (RF and SF) (Table 1). The central assumption of the MK test is that only nonsynonymous sites may be under selective pressure (i.e., synonymous sites are assumed to be neutrally evolving). If nonsynonymous sites are evolving according to a neutral model, then the expectation is that $P_n/P_s = D_n/D_s$. However, if nonsynonymous sites are under negative selection, then the expectation is that $P_n/P_s > D_n/D_s$, and if under positive selection, then $P_n/P_s < D_n/D_s$. Sawyer and Hartl derived the formulas for the expected values of SP, SF, RP, and RF using their PRF theory [2]. Below are the derivations of each of these formulas. For all of the derivations, assume that the data consists of samples of size m and n from two different species.

5.1. Expected Number of Synonymous Polymorphic Sites

Under neutral evolution ($s = 0$), the expected number of polymorphic sites with population daf x can be computed by taking the product of the per-locus mutation rate ($\theta = 2N\mu$) and the probability under a neutral model of a single mutation having a frequency of x (from (30)):

$$g_{\text{neutral}}(x) = \theta\frac{4Np}{x} = 2N\mu\frac{4N(1/2N)}{x} = \frac{4N\mu}{x} = \frac{2\theta}{x}. \tag{39}$$

Table 1: *McDonald-Kreitman contingency table.* 2-by-2 contingency table introduced by [18] for the inference of natural selection on nonsynonymous coding sites.

MK Table	No. of polymorphic sites	No. of fixed substitutions
Synonymous	SP	SF
Replacement (nonSynonymous)	RP	RF

Now, consider species 1 with sample size m. The probability that a polymorphic site, with population daf equal to x, is detected as polymorphic in a sample of size m is given as

$$\begin{aligned} P_m(x) &= 1 - (\text{all } m \text{ are derived}) - (\text{all } m \text{ are ancestral}) \\ &= 1 - x^m - (1-x)^m. \end{aligned} \tag{40}$$

The expected number of synonymous polymorphic sites, with population daf x, in the species 1 sample is the product of the expected number of synonymous polymorphic sites with daf x in the population ($g_{\text{neutral}}(x)$) and the fraction of those that are expected to be detected in a sample of size $m(P_m(x))$. It follows then that the total expected number of synonymous polymorphic sites, with any population daf, in the species 1 sample is computed by integrating the product of $g_{\text{neutral}}(x)$ and $P_m(x)$ over the range of possible values for x:

$$\begin{aligned} L(m) &= \int_0^1 g_{\text{neutral}}(x)\, P_m(x)\, dx \\ &= 2\theta \int_0^1 \frac{1 - x^m - (1-x)^m}{x}\, dx \\ &= 2\theta \sum_{k=1}^{m-1} \frac{1}{k}. \end{aligned} \tag{41}$$

Finally, the total number of expected synonymous polymorphic sites in both species' sample data is given as

$$\text{SP} = L(m) + L(n). \tag{42}$$

5.2. Expected Number of Replacement Polymorphic Sites

The derivation of the expected value of RP follows the same logic. As described in (35), the expected number of polymorphic sites with population daf x given some average selection pressure γ is given by $g(x)$. Similar to (41), the total expected number of replacement polymorphic sites in the species 1 sample is computed by integrating the product of $g(x)$ and $P_m(x)$ from 0 to 1:

$$\begin{aligned} H(m) &= \int_0^1 g(x)\, P_m(x) dx \\ &= \int_0^1 g(x) \left[1 - x^m - (1-x)^m\right] dx. \end{aligned} \tag{43}$$

Finally, the total expected number of replacement polymorphic sites in both species' sample data is given as

$$\text{RP} = H(m) + H(n). \tag{44}$$

5.3. Expected Number of Synonymous Fixed Substitutions

When $s = 0$, the expected number of fixed substitutions in one species relative to another that diverged $t_{\text{div}}2N$ generations ago is given as the product of the number of total mutations and the probability of fixation of each mutation. The number of total mutations is the product of the mutation rate per generation and the number of generations since divergence is

$$\theta t_{\text{div}} 2N. \tag{45}$$

The probability of fixation is given in (21). As s approaches 0 (i.e., neutral evolution), the probability of fixationcan be reduced to p using the approximation $e^{-a} = (1-a)$ for small values of a. Thus, for a newly derived neutral allele that has an initial frequency of $1/2N$, the probability of fixation is also $1/2N$.

Therefore, the total expected number of fixed substitutions in species 1 is

$$(\theta t_{\text{div}} 2N)(\frac{1}{2N}) = \theta t_{\text{div}}. \tag{46}$$

However, given that the data are samples of the populations from both species, not all sites identified as fixed substitutions in the sample are truly fixed substitutions in the entire population. The expected number of sites in the species 1 sample that fall into this category is given by

$$\int_0^1 T_m(x)\, g_{\text{neutral}}(x)\, dx = \theta \int_0^1 \left(x^m \, \frac{2}{x}\right) dx = \theta \left| \frac{x^m}{m} \right|_0^1 = \theta\, \frac{2}{m}, \tag{47}$$

where $T_m(x) = \Pr(\text{a derived allele daf } x < 1 \text{ is observed with } x = 1 \text{ in a size } m \text{ sample})$ and $g_{\text{neutral}}(x)$ is given in (39).

Therefore, the total expected number of synonymous fixed substitutions in both species' sample data is given as

$$\text{SF} = \theta\left(t_{\text{div}} + \frac{2}{m}\right) + \theta\left(t_{\text{div}} + \frac{2}{n}\right) = 2\theta\left(t_{\text{div}} + \frac{1}{m} + \frac{1}{n}\right). \tag{48}$$

5.4. Expected Number of Replacement Fixed Substitutions

Similar to the calculation of (46), given some selection pressure, γ, the expected number of fixed substitutions in one species relative to another that diverged $t_{\text{div}}2N$ generations ago is given as the product of (45) and (21):

$$(\theta\, t_{\text{div}} 2N)\left(\frac{1 - e^{-4Nsp}}{1 - e^{-4Ns}}\right). \tag{49}$$

Substituting $1/2N$ for p and invoking the approximation that $e^{-a} = (1-a)$ for small values of a, we arrive at the following:

$$(\theta\, t_{\mathrm{div}} 2N)\left(\frac{2s}{1-e^{-2\gamma}}\right) = \theta\, t_{\mathrm{div}} \frac{2\gamma}{1-e^{-2\gamma}}. \tag{50}$$

However, again, given that the data are samples of the populations from both species, not all sites identified as fixed substitutions in the sample are truly fixed substitutions in the entire population. The expected number of sites in the species 1 sample that fall into this category is given by

$$Q(m) = \int_0^1 T_m(x)\, g(x)\, dx = 2\theta \int_0^1 x^{m-1} \frac{1-e^{-2\gamma(1-x)}}{(1-e^{-2\gamma})(1-x)}\, dx. \tag{51}$$

Therefore, the total expected number of replacement fixed substitutions in both species' sample data is given as

$$\begin{aligned} \mathrm{RF} &= \theta\left(\frac{2\gamma\, t_{\mathrm{div}}}{1-e^{-2\gamma}} + 2G(m)\right) + \theta\left(\frac{2\gamma\, t_{\mathrm{div}}}{1-e^{-2\gamma}} + 2G(n)\right) \\ &= 2\theta\left(\frac{2\gamma\, t_{div}}{1-e^{-2\gamma}} + G(m) + G(n)\right), \\ &\qquad \text{where } \; G(m) = Q(m)/2\theta. \end{aligned} \tag{52}$$

5.5. Estimating Parameters

It is possible to obtain estimates of θ and γ by setting each of the observed values SP, RP, SF, and RF (Table 1) to their PRF expectations given by (42), (44), (48), and (52), respectively, and solving for the parameters. It has been shown that these estimates are equivalent to maximum-likelihood estimates [2, 19]. Bustamante et al. also eloquently describe and implement a hierarchical Bayesian model for parameter estimation [9].

6. Concluding Remarks

Sawyer and Hartl's seminal presentation of the PRF in 1992 provided an innovative mathematical framework for estimating selection pressures and mutation rates, which are critical parameters that influence molecular evolution. However, it is worth noting that the model does harbor certain limitations. Foremost among these is the assumption of site independence, which is equivalent to the assumption of free recombination among mutations (i.e., no linkage). Thus, the model may not be appropriate for many data wherein strong linkage is present. Another limitation is the assumption of infinite sites (i.e., each mutation is at a new site). Although this assumption allows for a simpler model, it is not always biologically appropriate, especially for organisms that experience a higher mutation rate. Indeed, recent work has shown that the assumption of infinite sites can underestimate selection pressures and mutation rates and even infer positive selection, when in fact there is weak negative selection [20]. Recent theoretical work has focused on relaxing these and other assumptions of the original PRF model, so as to make it more appropriate for diverse biological contexts. For a brief list of such studies, we refer the reader to [20]. Ongoing theoretical and empirical work in this area will undoubtedly continue to extend the power of a PRF-based approach for population genetic inference.

Acknowledgments

The authors would like to thank Professors Joshua B. Plotkin and Warren J. Ewens for many thoughtful comments on the manuscript, discussions about the material, and suggestions about the presentation. Their support and expert advice have been instrumental to the successful completion of this tutorial. They also thank two anonymous reviewers for their helpful suggestions in improving the manuscript.

References

[1] S. Biswas and J. M. Akey, "Genomic insights into positive selection," *Trends in Genetics*, vol. 22, no. 8, pp. 437–446, 2006.

[2] S. A. Sawyer and D. L. Hartl, "Population genetics of polymorphism and divergence," *Genetics*, vol. 132, no. 4, pp. 1161–1176, 1992.

[3] D. L. Hartl, E. N. Moriyama, and S. A. Sawyer, "Selection intensity for codon bias," *Genetics*, vol. 138, no. 1, pp. 227–234, 1994.

[4] H. Akashi, "Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA," *Genetics*, vol. 139, no. 2, pp. 1067–1076, 1995.

[5] M. W. Nachman, "Deleterious mutations in animal mitochondrial DNA," *Genetica*, vol. 102-103, pp. 61–69, 1998.

[6] D. M. Rand and L. M. Kann, "Mutation and selection at silent and replacement sites in the evolution of animal mitochondrial DNA," *Genetica*, vol. 102-103, pp. 393–407, 1998.

[7] H. Akashi, "Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination," *Genetics*, vol. 151, no. 1, pp. 221–238, 1999.

[8] D. M. Weinreich and D. M. Rand, "Contrasting patterns of nonneutral evolution in proteins encoded in nuclear and mitochondrial genomes," *Genetics*, vol. 156, no. 1, pp. 385–399, 2000.

[9] C. D. Bustamante, R. Nielsen, S. A. Sawyer, K. M. Olsen, M. D. Purugganan, and D. L. Hartl, "The cost of inbreeding in *Arabidopsis*," *Nature*, vol. 416, no. 6880, pp. 531–534, 2002.

[10] S. A. Sawyer, R. J. Kulathinal, C. D. Bustamante, and D. L. Hartl, "Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection," *Journal of Molecular Evolution*, vol. 57, supplement 1, pp. S154–S164, 2003.

[11] C. Bartolomé, X. Maside, S. Yi, A. L. Grant, and B. Charlesworth, "Patterns of selection on synonymous and nonsynonymous variants in *Drosophila miranda*," *Genetics*, vol. 169, no. 3, pp. 1495–1507, 2005.

[12] C. D. Bustamante, A. Fledel-Alon, S. Williamson, et al., "Natural selection on protein-coding genes in the human genome," *Nature*, vol. 437, no. 7062, pp. 1153–1157, 2005.

[13] K. Chen and N. Rajewsky, "Natural selection on human microRNA binding sites inferred from SNP data," *Nature Genetics*, vol. 38, no. 12, pp. 1452–1456, 2006.

[14] W. J. Ewens, *Mathematical Population Genetics: I. Theoretical Introduction*, Springer, New York, NY, USA, 2004.

[15] W. J. Ewens, *Mathematical Population Genetics*, Springer, New York, NY, USA, 1979.

[16] J. F. C. Kingman, *Poisson Processes*, Oxford University Press, Oxford, UK, 1993.

[17] S. S. Wilks, *Mathematical Statistics*, John Wiley & Sons, New York, NY, USA, 1962.

[18] J. H. McDonald and M. Kreitman, "Adaptive protein evolution at the *Adh* locus in *Drosophila*," *Nature*, vol. 351, no. 6328, pp. 652–654, 1991.

[19] S. Williamson, A. Fledel-Alon, and C. D. Bustamante, "Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance," *Genetics*, vol. 168, no. 1, pp. 463–475, 2004.

[20] M. Desai and J. B. Plotkin, "Detecting directional selection from the polymorphism frequency spectrum," Genetics, In press.