Department of Computer & Information Science

Technical Reports (CIS)

University of Pennsylvania *Year* 2008

A Tutorial of the Poisson Random Field Model in Population Genetics

Praveen Sethupathy* Sridhar Hannenhalli†

*University of Pennsylvania, praveens@mail.med.upenn.edu

†University of Pennsylvania, sridharh@pcbi.upenn.edu

This paper is posted at ScholarlyCommons.

http://repository.upenn.edu/cis_reports/807

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

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 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 has been applied to estimate the intensity of selection on synonymous and non-synonymous 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 use the PRF, among other techniques, to provide evidence for purifying selection (even stronger than on non-synonymous coding sites) on a class of regulatory sites known as microRNA target sites [13]. 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:

- Wright-Fisher model
- Diffusion approximation to the Wright-Fisher model
- Derivation, via diffusing theory, of formulas describing evolutionary processes of interest
- Derivation of the PRF using above-mentioned formulas

The first three points 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 non-professionals 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.

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) non-overlapping 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 (fitness = 1) and one derived (fitness = 1+s). 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:

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

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

Assuming no dominance and no recurrent mutation,

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

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 re-written 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 1a provides a pictorial representation of the basic Wright-Fisher model.

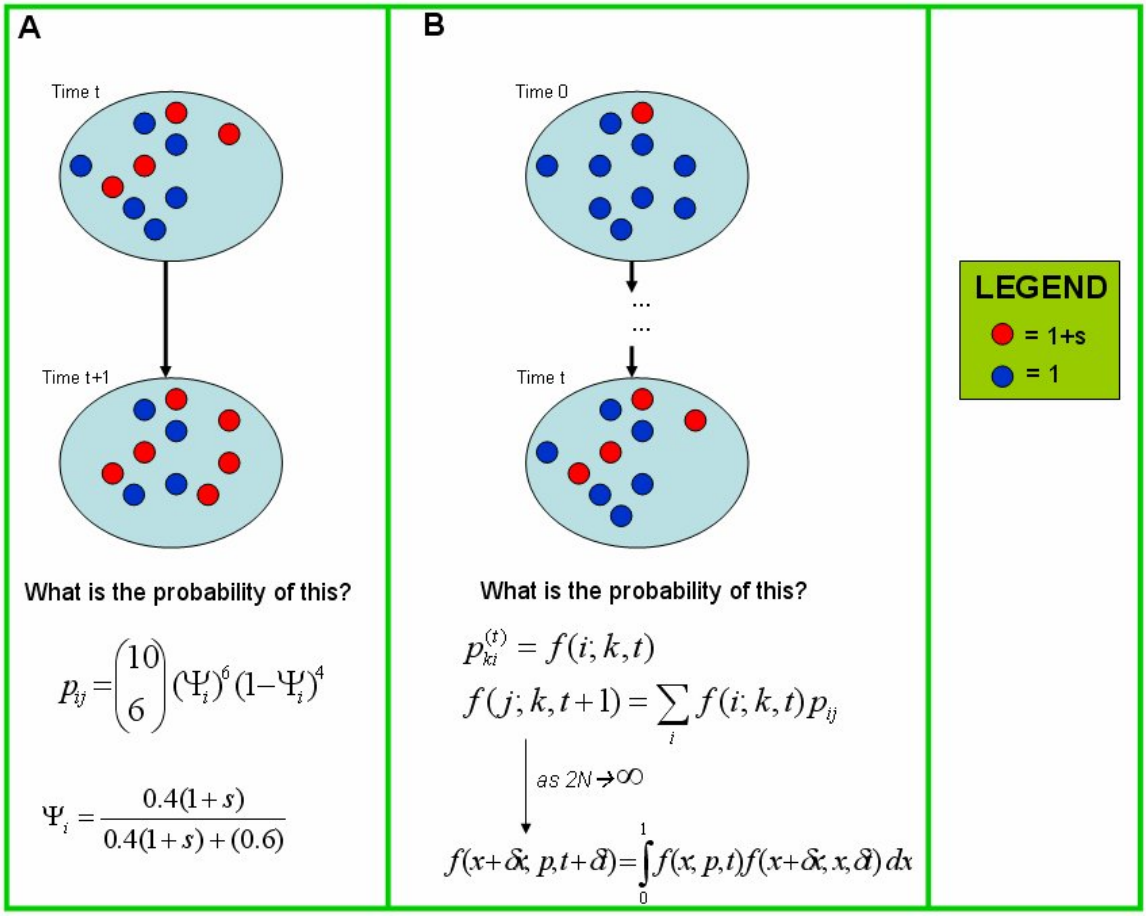

Figure 1. (a) Basic Wright-Fisher model assuming selection, but no dominance or recurrent mutation, **(b)** Diffusion approximation to the basic Wright-Fisher model.

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:

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

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:

$$f(j;\, k, t+1) = \sum_i f(i;\, k, t)\, p_{ij} \qquad \{2\}$$

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 1b):

$$f(x+\delta x;\ p,t+\delta t) = \int_0^1 f(x;\ p,t) f(x+\delta x;\ x,\delta t)\ dx \qquad \{3\}$$

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:

$$\frac{\partial f(x;p,t)}{\partial t} = \frac{\partial^2[b(x)f(x;p,t)]}{2\partial x^2} - \frac{\partial[a(x)f(x;p,t)]}{\partial x} \qquad \{4\}$$

where

$$E(dx) \approx a(x)\ dt$$
$$\mathrm{var}(dx) \approx b(x)\ dt$$

and *a(x)* and *b(x)* depend on the genetic model (for an example, see page 8).

Equation {3} 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.

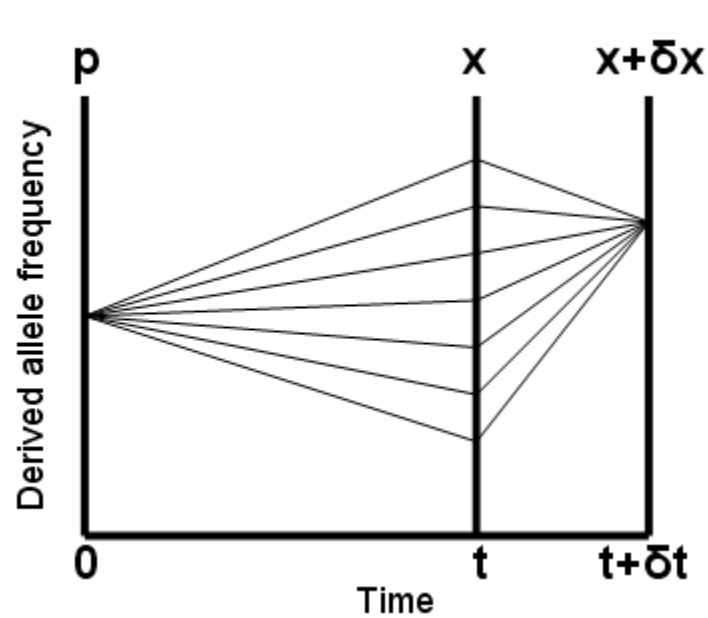

Figure 2. A diagrammatic intuition for equation {3}.

The frequency trajectory of a derived allele can also be depicted in the following manner:

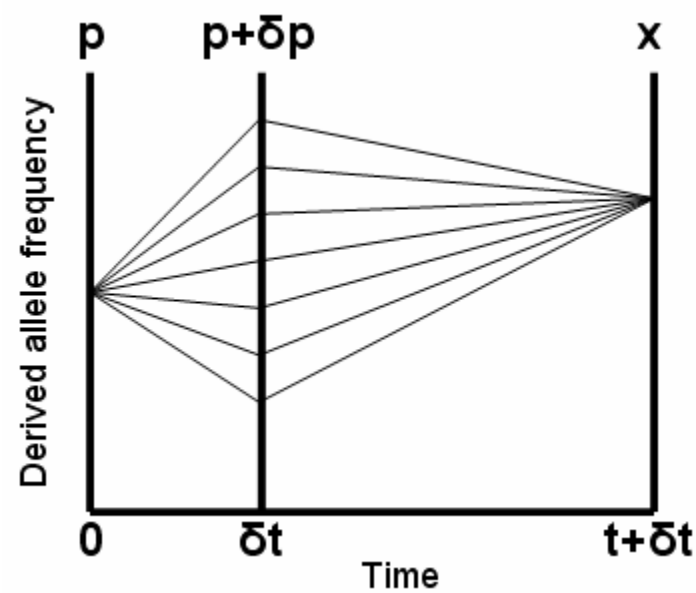

Figure 3. A diagrammatic intuition for equation {5}.

Figure 3 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:

$$f(x;\, p, t+\delta t) = \int_0^1 f(p+\delta p;\, p, \delta t)\, f(x;\, p+\delta p, t)\; d(\delta p) \qquad \{5\}$$

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

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

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 to chapter 4 of [15]. Next we will discuss how they are utilized to derive formulas for various quantities of evolutionary interest (yellow boxes in Figure 4).

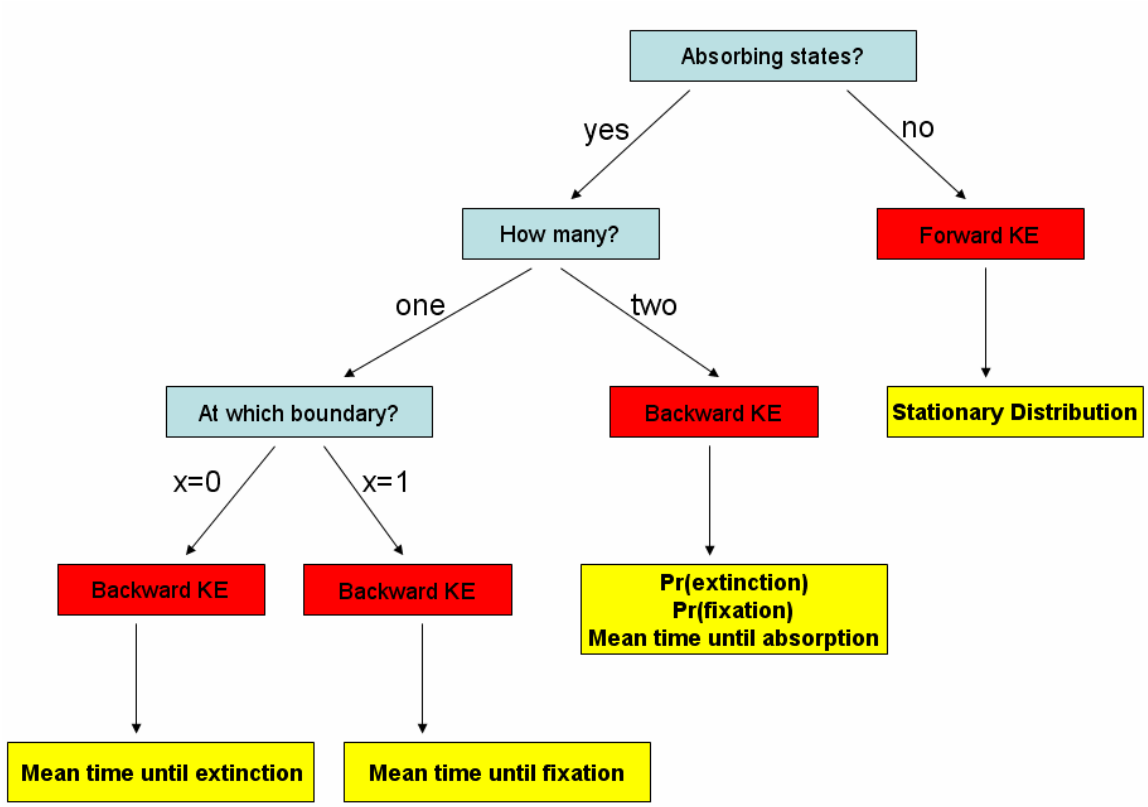

Figure 4. Blue boxes correspond to questions that clarify the assumptions of the genetic model being used, the red boxes correspond to when the Kolmogorov equations (KE) are utilized, and yellow boxes correspond to quantities of evolutionary interest.

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*:

$$\frac{\partial F(x,t)}{\partial t} = \frac{\partial [b(x) f(x,t)]}{2\partial x} - [a(x) f(x,t)] \qquad \{7\}$$

$$F(x,t) = \int_0^x f(y,t)\,dy \qquad \{8\}$$

F(x,t) is the probability of the derived allele assuming a frequency between 0 and *x* at time *t*. Therefore, the derivative of *F(x,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.

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].

Probability of extinction

Using {8} we arrive at an equation parallel to {6}:

$$\frac{\partial F(x,t)}{\partial t} = b(p)\frac{\partial^2[F(x,t)]}{2\partial p^2} + a(p)\frac{\partial[F(x,t)]}{\partial p} \qquad \{9\}$$

The probability that the derived allele frequency, *x*, reaches 0 at time *t* follows from {8} and is given by:

$$P_0(p,t) = \int_0^{0^+} f(y,t)\,dy = F(0^+,t) \qquad \{10\}$$

where *p* is the initial frequency of the derived allele and 0^+ indicates 0+ε where ε is very small.

Replacing $F(0^+,t)$ with $P_0(p,t)$, {9} can be written as:

$$\frac{\partial P_0(p,t)}{\partial t} = b(p)\frac{\partial^2[P_0(p,t)]}{2\partial p^2} + a(p)\frac{\partial[P_0(p,t)]}{\partial p} \qquad \{11\}$$

As $t \to \infty$, $P_0(p,t)$ can be interpreted as the probability that extinction ever occurs (independent of time) and can be re-written in the form $P_0(p)$. From {11} it is evident that $P_0(p)$ satisfies the following equation:

$$0 = b(p)\frac{\partial^2[P_0(p)]}{2\partial p^2} + a(p)\frac{\partial[P_0(p)]}{\partial p} \qquad \{12\}$$

Solving {12} we arrive at the following:

$$P_0(p) = \int_p^1 \psi(y)\,dy \Big/ \int_0^1 \psi(y)\,dy \qquad \{13\}$$

where

$$\psi(y) = e^{-2\int^{y}[a(z)/b(z)]\,dz} \qquad \{14\}$$

and where *a(z)* and *b(z)* are defined as in {4}.

Probability of fixation

The probability that the derived allele frequency, *x*, reaches 1 at time *t* follows from {8} and is given by:

$$P_1(p,t) = \int_{1^-}^{1} f(y,t)\,dy = 1 - \int_{0}^{1^-} f(y,t)\,dy = 1 - F(1^-,t) \qquad \{15\}$$

where *p* is the initial frequency of the derived allele and 1^- indicates 1-ε where ε is very small.

In equation {9}, *F(x,t)* can be replaced by *1-F(x,t)* without any loss of generality. Also, by replacing *1-F(1⁻,t)* with $P_1(p,t)$, {9} can be re-written as:

$$\frac{\partial P_1(p,t)}{\partial t} = b(p)\frac{\partial^2[P_1(p,t)]}{2\partial p^2} + a(p)\frac{\partial[P_1(p,t)]}{\partial p} \qquad \{16\}$$

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

$$P_1(p) = \int_0^p \psi(y)\,dy \Big/ \int_0^1 \psi(y)\,dy \qquad \{17\}$$

where $\psi(y)$ has been defined in {14} and *a(z)* and *b(z)* have been defined in {4}.

The probability of fixation and the probability of extinction must sum to 1. Using {13} and {17} we can verify that this is indeed the case.

Consider a genetic model that assumes the presence of selection, but no recurrent mutation, where $a(x) = \gamma x(1-x)$, $b(x) = x(1-x)$, and $\gamma = 2Ns$. Starting from {17}, we can express the probability of fixation under this genetic model in the following manner:

$$\begin{aligned} P_1(p) &= \int_0^p e^{-2\int^{y}[a(z)/b(z)]\,dz}\,dy \Big/ \int_0^1 e^{-2\int^{y}[a(z)/b(z)]\,dz}\,dy \\ &= \int_0^p e^{-4Nsy}\,dy \Big/ \int_0^1 e^{-4NSy}\,dy \\ &= \left[-e^{-4Nsy}\right]_0^p \Big/ \left[-e^{-4Nsy}\right]_0^1 \end{aligned}$$

$$= \frac{1 - e^{-4Nsp}}{1 - e^{-4Ns}} \qquad \{18\}$$

Mean time until either extinction or fixation

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

$$P_0(p,t) + P_1(p,t) = \int_0^t \phi(p,t)\, dt \qquad \{19\}$$

Equations {11}, {16}, and {19} show that $\phi(p,t)$ satisfies the following equation:

$$\frac{\partial \phi(p,t)}{\partial t} = b(p) \frac{\partial^2 [\phi(p,t)]}{2 \partial p^2} + a(p) \frac{\partial [\phi(p,t)]}{\partial p} \qquad \{20\}$$

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

$$-1 = -\int_0^\infty \phi(p,t)\, dt \qquad \{21\}$$

Performing integration-by-parts, we get the following:

$$-1 = -[t\phi(p,t)]_0^\infty + \int_0^\infty t \frac{\partial \phi(p,t)}{\partial t}\, dt \qquad \{22\}$$

Using equation {20} and the fact that $\phi(p,t)$ approaches 0 faster than *t* approaches ∞, we can re-write {22} as:

$$-1 = 0 + \int_0^\infty t\, [b(p) \frac{\partial^2 [\phi(p,t)]}{2 * \partial p^2} + a(p) \frac{\partial [\phi(p,t)]}{\partial p}]\, dt \qquad \{23\}$$

After interchanging the order of integration and differentiation we get:

$$-1 = b(p) \frac{d^2 \bar{t}(p)}{2 * dp^2} + a(p) \frac{d\bar{t}(p)}{dp} \qquad \{24\}$$

where

$$\bar{t}(p) = mean\ time\ until\ absorption = \int_0^\infty t\phi(p,t)\, dt = \int_0^1 t(p,x)\, dx \qquad \{25\}$$

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:

$$t(p,x) = \frac{2P_1(p)\int_x^1 \psi(y)\,dy}{b(x)\psi(x)} \qquad \{26\}$$

and $\psi(x)$ is defined in {14}.

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

Under the simplest genetic model that assumes no selection and no recurrent mutation, we can set $s=0$ in {14} and {18} and show that $P_1(p)$ reduces to p and $\psi(y)$ reduces to 1. It follows from this that {26} can be reduced to:

$$t(p,x) = \frac{2p(1-x)}{x(1-x)} = \frac{2p}{x} \qquad \{27\}$$

Under a genetic model where $s \neq 0$, using $\gamma = 2Ns$, {26} can be re-written as:

$$t(p,x) = \frac{\frac{2(1-e^{-2\gamma p})}{1-e^{-2\gamma}}\int_x^1 e^{-2\gamma y}\,dy}{x(1-x)(e^{-2\gamma x})}$$

After integrating and simplifying the terms, we obtain:

$$t(p,x) = \frac{(1-e^{-2\gamma p})(1-e^{-2\gamma(1-x)})}{[\gamma(1-e^{-2\gamma})]\,[x(1-x)]}$$

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:

$$f(x) = t(p,x) \approx \frac{1-e^{-2\gamma(1-x)}}{[N(1-e^{-2\gamma})]\,[x(1-x)]} \qquad \{28\}$$

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.

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 {28}, that

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

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). *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.

$$g(x) = 2N\mu \frac{1 - e^{-2\gamma(1-x)}}{[N(1-e^{-2\gamma})]\,[x(1-x)]}$$

$$= 2\mu \frac{1 - e^{-2\gamma(1-x)}}{(1-e^{-2\gamma})\; x(1-x)} \qquad \{29\}$$

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)*:

$$F(i) = \int_0^1 g(x)\; P(i \mid x)\, dx = \int_0^1 g(x) \binom{n}{i} x^i\; (1-x)^{n-i}\; dx \qquad \{30\}$$

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 {30}. 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*. Each random variable X_i is assumed to follow an independent Poisson distribution (and therefore, *X* is referred to as a Poisson Random Field) with mean equal to *F(i)*. 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:

$$P(X_i = x_i \mid \theta, \gamma) = \frac{e^{-F(i)} F(i)^{x_i}}{x_i!} \qquad \{31\}$$

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

$$P(X) = L(\theta,\gamma) = \prod_{i=1}^{n-1} P(X_i = x_i \mid \theta,\gamma) \qquad \{32\}$$

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 [16].

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, M. McDonald and R. Kreitman devised a 2-by-2 contingency table test of neutrality that was later named the MK test [17]. In the traditional MK test, a 2-by-2 contingency table is formed in order to compare the number of non-synonymous 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 non-synonymous sites may be under selective pressure (i.e. synonymous sites are assumed to be neutrally evolving). If non-synonymous sites are evolving according to a neutral model, then the expectation is that $P_n/P_s = D_n/D_s$. However, if non-synonymous sites are under negative selection, then the expectation is that $P_n/P_s > D_n/D_s$, and if under positive selection, then $P_n/P_s < D_n/D_s$. The 2-by-2 table is given below:

MK Table	*# of polymorphic sites*	*# of fixed substitutions*
Synonymous	SP	SF
Replacement (Non-Syn)	RP	RF

Table 1. 2-by-2 contingency table introduced by (McDonald and Kreitman, 1991) for the inference of natural selection on non-synonymous coding sites.

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.

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 equation {27}):

$$
\begin{aligned}
g_{neutral}(x) &= \theta \frac{2p}{x} \\
&= 2N\mu \frac{2(1/2N)}{x} \\
&= \frac{2\mu}{x}
\end{aligned}
\qquad \{33\}
$$

where μ is the per-locus mutation rate per generation.

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:

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

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 ($g_{neutral}(x)$) 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 $g_{neutral}(x)$ and $P_m(x)$ over the range of possible values for *x*:

$$
\begin{aligned}
L(m) &= \int_0^1 g_{neutral}(x)\ P_m(x)\, dx \\
&= 2\mu \int_0^1 \frac{1 - x^m - (1-x)^m}{x}\, dx \\
&= 2\mu \sum_{k=1}^{m-1} \frac{1}{k}
\end{aligned}
\qquad \{34\}
$$

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

$$
SP = L(m) + L(n) \qquad \{35\}
$$

Expected number of replacement polymorphic sites

The derivation of the expected value of RP follows the same logic. As described in {29}, the expected number of polymorphic sites with population daf *x* given some average selection pressure γ is given by *g(x)*. Similar to {34}, 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:

$$H(m) = \int_0^1 g(x)\, P_m(x) dx$$
$$= \int_0^1 g(x)[1 - x^m - (1-x)^m]\, dx \qquad \{36\}$$

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

$$RP = H(m) + H(n) \qquad \{37\}$$

Expected number of synonymous fixed substitutions

When $s=0$, the expected number of fixed substitutions in one species relative to another that diverged $t_{div}2N$ 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 number of mutations per generation and the number of generations since divergence:

$$\mu\, t_{div} 2N \qquad \{38\}$$

The probability of fixation is given in equation {18}. As s approaches 0 (i.e. neutral evolution), the probability of fixation can 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:

$$(t_{div}2N)\,(1/2N) = \mu\, t_{div} \qquad \{39\}$$

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:

$$\int_o^1 T_m(x)\, g_{neutral}(x)\, dx = \mu \int_0^1 \left(x^m \ \frac{2}{x} \right) dx = \mu \left| \frac{x^m}{m} \right|_0^1 = \mu \frac{2}{m} \qquad \{40\}$$

where $T_m(x)$ = *Pr(a derived allele with daf x<1 is observed with x=1 in a size m sample)* and $g_{neutral}(x)$ is given in {33}.

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

$$SF = \mu(t_{div} + 2/m) + \mu(t_{div} + 2/n)$$
$$= 2\mu(t_{div} + 1/m + 1/n) \qquad \{41\}$$

Expected number of replacement fixed substitutions

Similar to the calculation of {39}, given some selection pressure, γ, the expected number of fixed substitutions in one species relative to another that diverged $t_{div}2N$ generations ago is given as the product of {38} and {18}:

$$(\mu t_{div} 2N)\left(\frac{1-e^{-4Nsp}}{1-e^{-4Ns}}\right) \qquad \{42\}$$

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

$$(\mu t_{div} 2N)\left(\frac{2s}{1-e^{-2\gamma}}\right)$$

$$= \mu t_{div} \frac{2\gamma}{1-e^{-2\gamma}} \qquad \{43\}$$

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:

$$Q(m) = \int_o^1 T_m(x)\, g(x)\, dx$$
$$= 2\mu \int_0^1 x^{m-1} \frac{1-e^{-2\gamma(1-x)}}{(1-e^{-2\gamma})(1-x)}\, dx \qquad \{44\}$$

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

$$RF = \mu\left(\frac{2\gamma t_{div}}{1-e^{-2\gamma}} + 2G(m)\right) + \mu\left(\frac{2\gamma t_{div}}{1-e^{-2\gamma}} + 2G(n)\right)$$
$$= 2\mu\left(\frac{2\gamma t_{div}}{1-e^{-2\gamma}} + G(m) + G(n)\right) \qquad \{45\}$$
$$where\ G(m) = Q(m)/2\mu$$

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 {35}, {37}, {41}, and {45}, respectively, and solving for the parameters. It has been shown that these estimates are equivalent to maximum-likelihood estimates [2, 18]. Bustamante et al. also eloquently describe and implement a hierarchical Bayesian model for parameter estimation [9].

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. Recent theoretical work has focused on relaxing some of the 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 [19]. On-going theoretical and empirical work in this area will undoubtedly continue to extend the power of a PRF based approach for population genetic inference.

Acknowledgements

We 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 has been instrumental to the successful completion of this tutorial.

References

1. Biswas, S and Akey, JM (2006) Genomic insights into positive selection. **Trends Genet** *22*: 437-446.

2. Sawyer, SA and Hartl, DL (1992) Population genetics of polymorphism and divergence. **Genetics** *132*: 1161-1176.

3. Hartl, DL, Moriyama, EN and Sawyer, SA (1994) Selection Intensity for Codon Bias. **Genetics** *138*: 227-234.

4. Akashi, H (1995) Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* dna. **Genetics** *139*: 1067-1076.

5. Nachman, MW (1998) Deleterious mutations in animal mitochondrial dna. **Genetica** *102/103*: 61-69.

6. Rand, DM and Kann, LM (1998) Mutation and selection at silent and replacement sites in the evolution of animal mitochondrial dna. **Genetica** *102/103*: 393-407.

7. Akashi, H (1999) Inferring the fitness effects of dna mutations from polymorphism and divergence data: Statistical power to detect directional selection under stationarity and free recombination. **Genetics** *151*: 221-238.

8. Weinreich, DM and Rand, DM (2000) Contrasting patterns of nonneutral evolution in proteins encoded in nuclear and mitochondrial genomes. **Genetics** *156*: 385-399.

9. Bustamante, CD, Nielsen, R, Sawyer, S, Olsen, KM, Puruggannan, MD and Hartl, DL (2002). The cost of inbreeding in *Arabidopsis*. **Nature** *416*: 531-534.

10. Sawyer, SA, Kulathinal, RJ, Bustamante, CD and Hartl, DL (2003) Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. **Journal of Molecular Evolution** *57*: S154-S164.

11. Bartolome, C, Maside, X, Yi, S, Grant, AL, and Charlesworth, B (2005) Patterns of selection on synonymous and nonsynonymous variants in *Drosophila* miranda. **Genetics** *169*:1495-1507.

12. Bustamante, CD, Fledel-Alon, A, Williamson, S, Nielsen, R, Hubisz, MT, Glanowski, S, Tanenbaum, DM, White, TJ, Sninsky, JJ, Hernandez, RD, Civello, D, Adams, MD, Cargill, M and Clark, AG (2005) Natural selection on protein-coding genes in the human genome. **Nature** *437*: 1153-1157.

13. Chen, K and Rajewsky, N (2006) Natural selection on human microRNA binding sites inferred from SNP data. **Nat Genet** *38*:1452-1456.

14. Ewens, WJ (2004) *Mathematical Population Genetics: I. Theoretical Introduction.* Springer, New York, NY.

15. Ewens, WJ (1979) *Mathematical Population Genetics*. Springer, New York, NY.

16. Wilks, SS (1962) *Mathematical Statistics*. John Wiley & Sons.

17. McDonald, J and Kreitman, M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. **Nature** *351*: 652-654.

18. Williamson, S, Fledel-Alon, A and Bustamante, CD (2004) Population Genetics of Polymorphism and Divergence for Diploid Selection Models With Arbitrary Dominance. **Genetics** *168*: 463-475.

19. Desai, M and Plotkin, JB (2008) Detecting Directional Selection from the Polymorphism Frequency Spectrum. In submission.