Stationary Allele Frequency Distributions

Bruce Rannala, *University of California, Davis, California, USA*

Advanced article

Article Contents

Online posting date: 15th April 2013

Forces that determine the allele frequencies in natural populations include genetic drift, natural selection, migration and mutation. A balance of opposing forces can, in some cases, cause allele frequencies to approach a stationary distribution over time. The form of this distribution is not influenced by initial allele frequencies, but instead is determined by the relative magnitudes of different evolutionary forces. Statistical distributions are presented for the stationary allele frequencies under several simple population genetic models (including the k alleles symmetrical mutation model and the Wright island model of migration). In addition, the sampling distribution of allele counts under these models are described. The latter is useful when using genetic marker data to estimate population parameters. A population is at genetic equilibrium if evolutionary forces have persisted long enough for a population to have reached the stationary distribution, this is not often the case in nature.

Introduction

Frequencies of alleles at a genetic locus in a population are the outcome of a complex interplay among several forces including genetic drift, natural selection, gene flow and mutation. Forces such as selection, mutation and gene flow impose systematic pressures, causing allele frequencies to change in a particular direction. Genetic drift causes random (directionless) changes of allele frequency and contributes to the variation of allele frequencies observed within a single population over time, as well as the variation in allele frequencies among populations. **See also**: Genetic Drift in Human Populations; Migration

Allele frequency change in finite populations is a stochastic process. Consequently, the allele frequency in a population at some future generation cannot be exactly predicted, even if current allele frequencies, and the magnitudes of evolutionary forces causing allele frequency changes, are known. At best, probabilities can be assigned to each possible future combination of allele frequencies on the basis of a stochastic model of allele frequency change. The probability distribution of future population allele frequencies, in general, depends on the number of generations of reproduction and the initial allele frequencies, as well as the relative magnitudes of various evolutionary pressures influencing allele frequencies.

In special cases, when opposing forces of evolutionary change balance one another, a 'stationary' probability distribution of allele frequencies may exist. Over time, the allele frequencies in the population invariably approach this distribution, regardless of the initial frequencies of alleles in the population. Once this distribution is achieved, the population remains in this state so long as evolutionary pressures do not change. The stationary allele frequency distribution has long been of interest to population geneticists because it is informative about the long-term outcomes of evolutionary forces and is thought to be of relevance in interpreting gene frequencies in many (but not all) natural populations.

eLS subject area: Evolution & Diversity of Life

How to cite:
Rannala, Bruce (April 2013) Stationary Allele Frequency Distributions. In: eLS. John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0005465.pub3

Theoretical Background

Fisher–Wright model

Mathematical models aimed at predicting population allele frequency changes over time were independently developed by Fisher (1930) and Wright (1931). A fundamental model, considered by both, was of a diploid population of N individuals ($2N$ alleles) with a random mating structure. The so-called Fisher–Wright model assumes that each allele produces a random number of offspring (with mean one) in each generation (subject to the constraint that population size remains constant). The generations are discrete and nonoverlapping. The process of genetic drift is intrinsic to the model. Other processes such as migration, mutation and natural selection have been incorporated. **See also**: Fisher, Ronald Aylmer; Wright, Sewall

Many results obtained by studying the Fisher–Wright model also apply to organisms with more complex mating systems, including humans. Consider a single-genetic locus with two alleles, A and a. If no mutations occur and all other factors, apart from genetic drift, can be neglected, the

probability distribution of the number of copies of a at the next generation, given that there are x copies of a in the current generation, is a binomial (n, p) with parameters $p = x/2N$ and $n = 2N$. The probability distribution of allele frequencies at the next generation can be understood as the frequency with which populations with particular combinations of allele frequencies would be observed if one were to repeat a population mating experiment many times, starting each experiment with the same number, x, of copies of a.

Continuous approximation

The frequency of an allele in a diploid population of size N is an integer value that ranges from 0 to $2N$. To facilitate the study of allele frequency evolution over many generations under the Fisher–Wright model, Fisher (1930), Wright (1931) and Kimura (1955) all made use of an approximation that treats allele frequency as continuous (real valued) rather than discrete (integer valued). The relative frequency of a is $q = x/2N$, where x is the number of copies of a in the population. An increase, by one, of the number of copies of a in the population increases the relative frequency by $\delta q = 1/2N$. For large N (tending to infinity), δq becomes small (infinitesimal) and q takes on an approximately continuous range of values. **See also**: Diffusion Theory

Probability density of allele frequency

Making use of a continuous approximation simplifies the study of the probability distribution of allele frequencies. If $f(q|t)$ is defined to be the probability density function (pdf) of q after t generations of reproduction, the probability that q is in the interval (a, b) is

$$\int_a^b f(q|t)\mathrm{d}q$$

Properties of the process of population allele frequency change over time, such as the expected (or average) population allele frequency at time t, the variance of allele frequencies across replicate populations, etc., can be studied by evaluating statistical moments of the pdf of population allele frequencies.

Stationary Allele Frequency Distributions

Stationary distributions

If a stationary allele frequency distribution exists, the pdf of the stationary allele frequency distribution can be obtained by taking the limit

$$\varphi(q) = \lim_{t\to\infty} f(q|t)$$

Wright (1969) gives analytical expressions for the stationary distribution of allele frequencies under several models of mutation, migration and selection. Here, results for the allele frequency distribution under mutation models are briefly outlined with finite numbers of possible alleles, and models of subdivided populations with symmetrical migration patterns. The models presented here assume neutrality of alleles.

Two alleles: mutation or migration

Two-state mutation model

A simple model of mutation considered by Wright (1931) assumes that only two alleles are possible at a locus, a and A. The mutation process is reversible with a mutation rate v from allele a to A and rate u from A to a. Let q be the frequency of allele a. Then the stationary pdf of q is

$$\phi(q) = \frac{\Gamma(4Nu + 4Nv)}{\Gamma(4Nu)\Gamma(4Nv)} q^{4Nu-1}(1-q)^{4Nv-1}$$

where Γ denotes the gamma function. This is a beta distribution with parameters $\alpha = 4Nu$ and $\beta = 4Nv$ (see Johnson *et al.*, 1995). The expectation (mean) of the allele frequency at stationarity is $u/(u+v)$ and the variance (across populations that are at stationarity and experiencing identical evolutionary pressures) is $uv/[(u+v)^2(4N(u+v)+1)]$.

Wright island model

A simple model of migration, the so-called 'Wright island model', assumes that a fraction m of the alleles in a population are replaced by migrant alleles in each generation. The allele frequency among migrants is assumed to be constant, $\bar{q}$. The stationary pdf of allele frequency in the population is

$$\phi(q) = \frac{\Gamma(4Nm)}{\Gamma(4Nm\bar{q})\Gamma[4Nm(1-\bar{q})]} q^{4Nm\bar{q}-1}(1-q)^{4Nm(1-q)-1}$$

This is a beta distribution with parameters

$$\alpha = 4Nm\bar{q} \quad \beta = 4Nm(1-\bar{q})$$

The mean allele frequency at stationarity is $\bar{q}$.

Multiple alleles: mutation or migration

Mutation model (k alleles)

A general formula is available for the stationary allele frequency distribution only under a (quite unrealistic) model of mutation in which all alleles mutate to allele j with the same rate v_j. This has been called the parent

independent mutation model. In that case, the stationary joint pdf of allele frequencies is

$$\phi(q_1, q_2, \ldots q_k) = \Gamma\left(4N\sum_{i-1}^{k} v_i\right)\prod_{j-1}^{k}\frac{q_j^{4Nv_j-1}}{\Gamma(4Nv_j)}$$

where q_j is the frequency of allele j.

Wright island model (k alleles)

The Wright island model can be extended to k alleles (see Wright, 1969). If q_j is defined to be the frequency of the jth allele on an island, and $\bar{q}_j$ to be the frequency of the allele among migrants, the stationary joint pdf of allele frequencies is

$$\phi(q_1, q_2, \ldots q_k) = \Gamma(4Nm)\prod_{j=1}^{k}\frac{q_j^{4Nm\bar{q}_j-1}}{\Gamma(4Nm\bar{q}_j)}$$

The above distributions are special cases of a general distribution known as the Dirichlet distribution (Kotz *et al.*, 2000), which arises in many population genetic models.

Sampling Distributions of Alleles

The stationary distributions of allele frequencies outlined earlier apply to a population. Experimental studies of natural populations usually characterise the number of copies of each distinct allele observed in a sample of n diploid individuals ($2n$ copies in total). This is referred to as the sampling distribution of alleles and can be used to estimate parameters of population genetic models. For example, one can derive a maximum-likelihood estimator of $4Nm$ under the Wright island model (Rannala and Hartigan, 1996). Assuming that individuals are sampled at random (with respect to genotype), and that genotypes are in Hardy–Weinberg equilibrium in a population, the probability distribution of allelic sample configurations (conditional on the allele frequencies) is either binomial (two alleles) or multinomial (more than two alleles)

$$\Pr(x_1, x_2, \ldots x_k|\mathrm{q}) = \binom{2n}{x_1, x_2, \ldots, x_k}\prod_{j=1}^{k} q_j^{xj}$$

where $\mathrm{q} = q_j$ is a vector of the population allele frequencies, q_j the frequency of allele j and x_j the number of copies of allele j in the sample. If the population is at equilibrium, and the stationary allele frequency distribution is specified under a particular model, the sampling distribution is

$$\Pr(x_1, x_2, \ldots x_k) = \int \Pr(x_1, x_2, \ldots x_k|\mathrm{q})\phi(\mathrm{q})\mathrm{dq}$$

where integration is over the multidimensional simplex of population allele frequencies. For most models of finite numbers of alleles that have been a subject of analysis, the sampling distribution is a multinomial Dirichlet distribution (Johnson *et al.*, 1997). In the special case of two alleles, this simplifies to the beta binomial distribution.

Mutation model (k alleles)

For the k-allele symmetrical mutation model, the stationary sampling distribution of alleles is

$$\Pr(x_1, x_2, \ldots x_k) = \left[\Gamma(2n+1)\Gamma\left(4N\sum_{i=1}^{k} v_i\right)\right] \times \left[\Gamma\left(2n + 4N\sum_{i=1}^{k} v_i\right)\right]^{-1} \times \prod_{j=1}^{k}\frac{\Gamma(x_j + 4Nv_j)}{\Gamma(x_j+1)\Gamma(4Nv_j)}$$

Wright island model (k alleles)

For the island model, the stationary sampling distribution of alleles is

$$\Pr(x_1, x_2, \ldots x_k) = \frac{\Gamma(2n+1)\Gamma(4Nm)}{\Gamma(2n+4Nm)} \times \prod_{j=1}^{k}\frac{\Gamma(x_j + 4Nm\bar{q}_j)}{\Gamma(x_j+1)\Gamma(4Nm\bar{q}_j)}$$

Discussion

Stationary allele frequency distributions have been derived for a broad range of models, in particular models with selection and k alleles parent independent mutation (see the introduction to Buzbas and Joyce, 2009). Here the author has outlined results for the stationary frequency distributions and sampling distributions of several of the simplest models with migration, genetic drift and mutation. More recently, simulation-based methods have been extensively used to study frequency distributions of alleles under more complex models for which analytical solutions are not available. Modern approaches often focus on nonstationary sampling distributions using the coalescent process which models the genetic evolution of a sample of chromosomes backwards through time, rather than projecting frequencies forward in time as is done in the classical Wright–Fisher–Markov process or the diffusion approximation to this process. **See also**: Coalescent Theory; Evolution: Neutralist View; Mutational Change in Evolution; Population Genetics: Overview

References

Buzbas EO and Joyce P (2009) Maximum likelihood estimates under K-allele models can be numerically unstable. *Annals of Applied Statistics* **3**: 1147–1162.

Fisher RA (1930) *The Genetical Theory of Natural Selection*. Oxford, UK: Oxford University Press.

Johnson NL, Kotz S and Balakrishnan N (1995) *Continuous Univariate Distributions*, vol. 2. New York: Wiley.

Johnson NL, Kotz S and Balakrishnan N (1997) *Discrete Multivariate Distributions*. New York: Wiley.

Kimura M (1955) Solution of a process of random genetic drift with a continuous model. *Proceedings of the National Academy of Sciences of the USA* **41**: 144–150.

Kotz S, Balakrishnan N and Johnson NL (2000) *Continuous Multivariate Distributions*, vol. 1. New York: Wiley.

Rannala B and Hartigan JA (1996) Estimating gene flow in island populations. *Genetical Research* **67**: 147–158.

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

Wright S (1969) Evolution and the genetics of populations. In: *The Theory of Gene Frequencies*, vol. 2. Chicago, IL: University of Chicago Press.

Further Reading

Ewens WJ (1979) *Mathematical Population Genetics*. New York: Springer.

Fearnhead P (2006) The stationary distribution of allele frequencies when selection acts at unlinked loci. *Theoretical Population Biology* **70**: 376–386.

Gale JS (1990) *Theoretical Population Genetics*. London, UK: Unwin Hyman.

Medhi J (1994) *Stochastic Processes*, 2nd edn. New Delhi, India: Wiley Eastern.

Rice JA (1995) *Mathematical Statistics and Data Analysis*, 2nd edn. Belmont, CA: Duxbury Press.