A time series analysis of the population genetics of the self-incompatibility polymorphism. 1. Allele frequency distribution of a population with overlapping generations and variation in plant size

ROGER J. BROOKS^{†§}, ANDREW M. TOBIAS[†] & MICHAEL J. LAWRENCE^{‡*}

[†]Operational Research Group, School of Manufacturing and Mechanical Engineering, University of Birmingham, Birmingham B15 2TT and [‡]Wolfson Laboratory for Plant Molecular Biology, School of Biological Sciences, University of Birmingham, Birmingham B15 2TT, U.K.

A mathematical theory of the population genetics of the self-incompatibility polymorphism is developed for the case where self-incompatibility is controlled by a single, multiallelic gene, *S*, expressed gametophytically in the pollen. The theory is based on the probability of a particular allele occurring in a new plant and an approximate equation for this is compared, using computer simulation, with the true value, which depends upon the frequencies of the genotypes of the parental generation. This probability is used to express the frequency of a single established allele, from a population with both overlapping generations and variation in plant size, as a time series. The frequency is shown to follow closely the normal distribution which leads to a simple equation for the variance of the distribution. This variance is approximately equal to the average variance of the frequencies of the different alleles in a single generation and the values closely match the computer simulation results of Brooks *et al.* (1996). The theory allows alternative scenarios to be investigated easily and the effects of varying the parameter values are discussed. The time to steady state can also be calculated.

Keywords: allele frequency distribution, mathematical theory, population genetics, self-incompatibility polymorphism, time series.

Introduction

A mathematical theory of the self-incompatibility polymorphism in flowering plants was first developed by Wright (1939), who gave equations for the expected change in the relative frequency of an allele from one generation to the next and for the probability distribution of allele frequencies in a single generation. This distribution was used to calculate the number of alleles which could be maintained in a population in a number of different circumstances. Wright's work was stimulated by Emerson's (1939) survey of the *Oenothera organensis* population which occurred in canyons in the Organ

Mountains of southern New Mexico. Emerson identified 37 different alleles from seed and cuttings collected from a number of plants in the population and, in subsequent studies, found a further eight alleles (Emerson, 1940). He estimated that the total population probably contained fewer than 500 plants. However, Wright's calculations indicated that a population of this size in steady state would be unable to maintain more than 12 or 13 alleles unless the mutation rate was high or the population consisted of small groups with little interbreeding between them (neither of which was thought to be the case). The nature of the habitat of the Organ Mountains population made an estimation of population size difficult and a more recent survey indicated that the population may contain some 5000 plants (Levin *et al.*, 1979). A population of this size is large enough, assuming a reasonable mutation

*Correspondence.

§Present address: IACR Long Ashton Research Station, Department of Agricultural Sciences, University of Bristol, Long Ashton, Bristol BS18 9AF, U.K.

rate, to maintain 45 alleles according to Wright's theory, which therefore solved the problem originally posed by Emerson's survey.

Fisher (1958) and Moran (1962) questioned some of the approximations made by Wright, to which he replied in 1960 and 1964, respectively, and the validity of the original results was confirmed by the computer simulations of Crosby (1966), Ewens & Ewens (1966), Mayo (1966) and Kimura & Maruyama (1966). Theoretical studies have also been carried out by Ewens (1964), Yokoyama & Nei (1979) and Yokoyama & Hetherington (1982). Nevertheless, the amount of work in this area is quite small considering the widespread occurrence and the importance of the self-incompatibility polymorphism in flowering plants and fungi.

The situation considered by Wright (1939) was a panmictic population in which the parents for each new generation are selected at random from the previous generation. Each of the subsequent studies has considered essentially the same model. A number of other factors occur in natural populations which are likely to affect the population genetics of the polymorphism. In particular, both pollen and seed are not dispersed uniformly, generations overlap and plants vary in size. We recently used computer simulation to investigate the effects of these four factors on the average variance of the allele frequencies of a population in steady state (Brooks *et al.*, 1996). That investigation showed that overlapping generations and variation in plant size considerably decreased and increased the variance, respectively. References to simulations refer to that study and the terminology used in that paper is followed here with the factor for overlapping generations (seed dormancy) being denoted *D* and the factor for variation in plant size (the existence of large plants) being denoted *L*. The simulation study also investigated the effects of limited pollen and seed dispersal (denoted *P* and *S*, respectively), but these factors were found to have only a small effect on the allele frequency variance. The different models are denoted by the factors that they contain, with the model of the panmictic population denoted the NIL model. In order to simplify the discussion, the term generation is used to refer to successive flowering episodes, although, in the case of the models containing seed dormancy, a generation is equal to more than one flowering episode.

This paper sets out a new mathematical approach to the self-incompatibility polymorphism. The theory is simpler than that of Wright (1939), but, as a result, it is able to include the *D* and *L* factors and is more amenable to further analysis. A second

paper (Brooks *et al.*, 1997) sets out further results, including the number of alleles that can be maintained by a population in equilibrium.

Frequency-dependent selection equations

The theory is based on an equation for the probability of a given allele occurring in a new plant. This section compares alternative equations for this probability for the simplest situation of a panmictic population. Let the relative frequency of a given allele, α , in the previous generation be q_α . Then, the probability, P_α , of either of the two alleles in a new plant being α is given by (Wright, 1960),

$$P_\alpha = q_\alpha + \sum_{\beta \neq \alpha} \sum_{\gamma \neq \alpha} c_{\beta\gamma} \times \frac{q_\alpha}{1 - q_\beta - q_\gamma}, \quad (1)$$

where $c_{\beta\gamma}$ = relative frequency of the genotype $\beta\gamma$ in the previous generation. The first term in the equation is the probability that the allele inherited from the maternal parent is α , which, because each plant has an equal chance of being chosen as the maternal parent, is simply q_α . The second term is the probability that the allele inherited from the paternal parent is α , which requires that neither of the alleles of the maternal parent is α . For each possible maternal genotype $\beta\gamma$, any allele in the population other than those which are β or γ is a feasible paternal allele.

This probability requires that the frequencies of all the genotypes in the previous generation are known. However, if it is assumed that the frequencies of all alleles other than α are equal, P_α can be expressed as a function of the relative frequency of α , q_α :

$$P_\alpha(q_\alpha) = 2q_\alpha \left(1 + \frac{1 - q_\alpha k}{k - 3 + 2q_\alpha} \right), \quad (2)$$

where k = the number of alleles (assumed, in the subsequent analysis, to remain constant). This equation is the same as that of Wright (1939), except that his equation included a variable, R , which was equal to the ratio of successful α paternal alleles to other successful paternal alleles, when mating with a maternal parent without the α allele. The only equation given by Wright (1939) for R assumed that the frequencies of the other alleles are equal, in which case Wright's equation reduces to eqn (2). In his subsequent analysis, however, Wright (1939) considered R to be an unknown constant for which an estimate was required in order to evaluate the particular allele frequency distribution.

The expected relative frequency of α in the next generation is $P_\alpha(q_\alpha)/2$, and so the relative frequency of α is expected to increase if the current frequency, q_α , is less than the mean allele frequency, $1/k$, and to decrease if it is greater than the mean. The frequency tends to move towards the mean frequency and it is this frequency-dependent selection that maintains the polymorphism.

A simulation program was run to compare eqns 1 and 2. The program randomly set up a single generation of 3840 plants containing 16 alleles (the same population size and number of alleles as the previous simulation study). For each of the 16 alleles, the true value of P_α was calculated from eqn (1) using the frequencies of the genotypes in the generation. This probability was then compared with the approximate P_α values from eqn (2) which assumes that the relative frequencies of the other alleles are equal. The 3840 pairs of alleles, representing the genotypes of each plant in the generation, were chosen randomly, using the NAG random number generator G05CAF, based on fixed probabilities of choosing each allele. The second allele in each pair was reselected if it was the same as the first, as individuals must be heterozygotes. Three

versions of the program were run with different allele frequency characteristics in the simulated generation and these are denoted G1, G2 and G3. The allele frequencies varied considerably for G3, less so for G2, and were nearly equal for G1, which was achieved by altering the probabilities of choosing each allele when producing the genotypes of the generation. For G1, these probabilities were equal, whereas, for G2, the probability for allele α (where $\alpha = 1, \dots, 16$) was $(310 + 20\alpha)/7680$ (so that allele 1 will be rare and allele 16 will be common, etc.), and for G3 this was $(-30 + 60\alpha)/7680$. The average standard deviations of the allele frequencies for G1, G2 and G3 were 21, 91 and 270, respectively. Each version of the program was run 63 times (in order to obtain about 1000 values) giving 1008 values for each population.

In every case, the true value of P_α from eqn (1) was greater than the value from eqn (2). The expected changes in the absolute allele frequencies are given by $NP_\alpha - 2Nq_\alpha$, where N is the number of plants in the population (3840), and Fig. 1 compares the expected changes for G2 and G3 using the true P_α -values with the expected changes using P_α from eqn (2). The values for G1 lay between the G2

Fig. 1 The expected changes in allele frequency (mean = 480) using the true probability of selection, for randomly generated populations G2 and G3 and the approximate probability eqn (2).

values and the eqn (2) curve (they were very close to the eqn (2) curve). The results show that the approximate probability is close to the true value provided that the variation in the allele frequencies is not too large; they also show that the expected change in the frequency of an allele (the magnitude of the frequency-dependent selection effect) is small unless the frequency of the allele is very high.

The true probability values in Fig. 1 cross the x-axis above the mean allele frequency (480) and so the frequency of an allele will tend towards a value slightly above the mean, with this value depending upon the amount of variation in the allele frequencies. It follows that the distribution of allele frequencies will be slightly negatively skewed with a longer tail of low values. The variation in the allele frequencies for G3 is extreme, with the frequencies being roughly equally spaced and the lowest being close to zero. In the absence of selection other than frequency-dependent selection, the frequencies of most of the alleles are likely to be close to the mean and only rarely will the frequencies be very low or very high (the time series analysis will show that the allele frequency distribution is approximately normal). The expected change in frequencies will therefore lie between the G3 points and the eqn (2) curve and will vary slightly from generation to generation, so that the amount of skewness in the distribution of the frequencies is likely to be small. Wright (1960) compared his formula, including the constant R , as well as Fisher's (1958) alternative formula, with the exact formula, using four numerical examples with four alleles and four examples with five alleles. He also found that the approximate formulae gave values close to the true values, provided that the allele frequencies were not too extreme.

The following mathematical theory uses a time series analysis of the frequency of a single allele and the theory requires the probability to be a linear function of the previous relative frequency of the allele to make the analysis tractable. Even without this restriction, neither of eqns (1) or (2) is suitable. The analysis is based on the frequencies of a single allele and not on those of the genotypes, so eqn (1) cannot be used. The mean frequency calculated from eqn (2) will not be equal to the correct mean of $2N/k$, with the adjustment required depending on the extent of the variation of the allele frequencies which is not known in advance.

A linear function can be obtained from eqn (2) by using the gradient approximation of $P_\alpha(q_\alpha)$ about the mean allele frequency ($q = 1/k$), and this gives (dropping the α subscript for ease of notation):

$$P(q) = 2 \left(\frac{1}{(k-1)(k-2)} + \left(1 - \frac{k}{(k-1)(k-2)} \right) q \right). \quad (3)$$

Subsequent notation is made simpler by defining

$$\lambda = \frac{1}{(k-1)(k-2)}, \quad \text{and} \quad \theta = 1 - \frac{k}{(k-1)(k-2)}, \quad (4)$$

so that

$$P(q) = 2(\lambda + \theta q). \quad (5)$$

This probability will be close to the true probability if the allele frequencies are close to their mean frequency, particularly if the number of alleles is large. The accuracy of the results of the following time series analysis indicates that this is a reasonable approximation.

Time series analysis

Outline of the time series analysis

The analysis considers the frequency of a single established allele. Its frequency over successive generations is a time series, with the frequency in a given generation expressed as an expected value, which is a function of the previous frequencies, plus a random variable because of the randomness in the selection process. Substitution of the expressions for the previous frequencies enables the frequency at any time to be expressed as a sum of the independent random variables and this specifies the probability distribution for the allele.

In our previous study using computer simulation, the output value for each generation, t , was the variance, var_t , of the k allele frequencies. In particular, the mean value of var_t when the population is in steady state, var_{ss} , was compared for the different combinations of the four factors D, P, S and L. The k allele frequencies in each generation must sum to $2N$ and the frequencies of each allele in successive generations are correlated. However, despite this, the frequencies in a given generation can be considered speculatively as k randomly selected values from the probability distribution for a single allele. Therefore, var_t is approximately equivalent to taking the variance of a random sample of k values from the allele frequency distribution, so that its average value for many populations in the same state is approximately equal to the variance of the allele frequency distribution for that state. Therefore the average steady-state variance, var_{ss} , is approximately equal to the variance of the frequency distri-

bution for a single allele in steady state (whichever way the probability distribution is derived). Because the mean allele frequency in the simulated populations was known, the denominator of var_t was k (rather than $k - 1$), and this is also the variance formula that should correspond best to the variance of the allele frequency distribution.

Panmictic population

The basic time series equation is based on the probability, $P(q)$, of a given allele being selected as one of the two alleles of a new plant, using eqn (5). Each new generation consists of N plants and the process of choosing each plant is identical because uniform dispersal distributions are assumed. The probability distribution for the absolute frequency of the allele in the new generation is therefore the binomial distribution with mean $NP(q)$ and variance $NP(q)(1 - P(q))$. For any sizeable population, the normal distribution provides a very close approximation to this distribution, and a continuous distribution is acceptable.

A panmictic population is the simplest model in which the parents are chosen at random from the previous generation. For any one of the k alleles in the population, let its absolute frequency in the t th generation be f_t . Then, the NIL model time series equation is:

$$f_t = NP \left(\frac{f_{t-1}}{2N} \right) + z_t, \quad (\text{for } t > 0), \quad (6)$$

where

z_t = normal random variable with mean 0 and

$$\text{variance } NP \left(\frac{f_{t-1}}{2N} \right) \left(1 - P \left(\frac{f_{t-1}}{2N} \right) \right).$$

The frequency, f_t , for a real population, must lie between 0 and N (and must be an integer), but, for realistic values of N and $P(q)$ the probability of eqn (6) yielding a value outside this range is very small.

The variance of z_t depends on the value of $P(f_{t-1}/2N)$, but is approximately linearly related to it for small $P(f_{t-1}/2N)$. Therefore, using the average value of $P(f_{t-1}/2N)$ which is $2/k$, the mean variance is approximately:

$$\frac{2N(k-2)}{k^2}. \quad (7)$$

Overlapping generations

The effect of overlapping generations is that the parents of a plant may come from a generation that is several generations previous to that of the offspring. The appropriate probability distribution for seed dormancy, assuming that the probability of a given seed germinating each time a new generation arises is a constant, is the geometric distribution. Denoting the constant probability by d , the probability of a seed germinating after i generations is $d(1-d)^{i-1}$. On average, this is the proportion of the new generation that are progeny of the i th previous generation and so the expected frequency of the allele in the new generation is the sum of expected frequencies for each of these proportions. Therefore, for the D model, eqn (6) becomes:

$$f_t = z_t + \sum_{i=1}^{\infty} d(1-d)^{i-1} NP \left(\frac{f_{t-i}}{2N} \right). \quad (8)$$

Variation in plant size

The L factor for variation in plant size in the simulation study assumed only two sizes, large and small, for simplicity. The analysis that follows also assumes only two plant sizes but could be extended to more than two. Large plants have a greater probability of contributing to the next generation than small plants because they produce more seed and pollen, and the relative frequency of the allele in the previous generations needs to be adjusted to take account of this. Because work on *Papaver dubium* indicated that the variation in plant size, at least in that species, did not have a genetic basis (Arthur *et al.*, 1973), it is assumed that the probability of a plant being large is the same for every plant. Let the proportion of large plants be l_p and let the ratio of the quantity of seed and pollen of large plants compared to that of small plants be l_q (large plants have l_q times the probability of contributing to the next generation). Each large plant is therefore equivalent to l_q small plants and so the population contains an equivalent of:

$$N(1-l_p) + Nl_p l_q = N(1 + l_p(l_q - 1)) \text{ small plants.} \quad (9)$$

The frequency of a particular allele in the entire population, f_t , is the total number of plants in the population that contain the allele. The probability of any plant being large is l_p , and so the frequency of the allele in large plants, $f_{l,t}$, follows the binomial distribution with mean $f_t l_p$ and variance $f_t l_p(1 - l_p)$. This again can be approximated by the normal distribution and therefore,

$$f_{i,t} = f_i l_p + z_{i,t}, \quad (10)$$

where $z_{i,t}$ = normal random variable with mean 0 and variance $f_i l_p (1 - l_p)$. The variance of $z_{i,t}$ varies linearly with f_i , which has a mean of $2N/k$, so the mean variance is:

$$\frac{2Nl_p(1-l_p)}{k}. \quad (11)$$

Because each large plant is equivalent to l_q small plants, the frequency of the allele in the generation is equivalent to a small plant frequency of $f_i + f_{i,t}(l_q - 1)$. The effective relative frequency of the allele is therefore:

$$\frac{f_i + f_{i,t}(l_q - 1)}{2N(1 + l_p(l_q - 1))}. \quad (12)$$

Substituting in the expression for $f_{i,t}$ from eqn (10), the effective relative frequency of the allele becomes:

$$\frac{f_i}{2N} + \frac{z_{i,t}(l_q - 1)}{2N(1 + l_p(l_q - 1))} = \frac{f_i}{2N} + \frac{Lz_{i,t}}{2N}, \quad (13)$$

where

$$L = \frac{l_q - 1}{1 + l_p(l_q - 1)}.$$

The time series equation for the allele frequency for the DL model is then obtained by taking eqn (8) for the D model and using eqn (13) to give the relative frequency of the allele in the probability function, P , and so:

$$f_t = z_t + \sum_{i=1}^{\infty} d(1-d)^{i-1} NP \left(\frac{f_{t-i}}{2N} + \frac{Lz_{i,t-i}}{2N} \right), \quad (14)$$

with initial conditions, to match the simulations, $f_t = 2N/k$ for $t \leq 0$ (the initial generation was denoted generation 1 in the simulations but the subsequent notation is simpler if the initial generation is taken as $t = 0$).

Derivation of the allele frequency distribution

The probability function, P , given by eqn (5) can be substituted into eqn (14). The allele frequency at generation t is then

$$f_t = z_t + 2N\lambda + \sum_{i=1}^{\infty} d\theta(1-d)^{i-1}(f_{t-i} + Lz_{i,t-i}). \quad (15)$$

This also holds at generation $t-1$,

$$f_{t-1} = z_{t-1} + 2N\lambda + \sum_{i=1}^{\infty} d\theta(1-d)^{i-1}(f_{t-1-i} + Lz_{i,t-1-i}). \quad (16)$$

By comparing these two equations, f_t can be expressed just in terms of f_{t-1} ,

$$\begin{aligned} f_t &= (1-d)(f_{t-1} - z_{t-1} - 2N\lambda) + z_t + 2N\lambda \\ &\quad + d\theta(f_{t-1} + Lz_{i,t-1}) \\ &= (1-d+d\theta)f_{t-1} + z_t - (1-d)z_{t-1} + d\theta Lz_{i,t-1} + 2Nd\lambda. \end{aligned} \quad (17)$$

The general equation for the allele frequency is then obtained by substituting for $f_{t-1}, \dots, f_0$ each time using eqn (17), and noting that $f_0 = 2N/k$, to give

$$\begin{aligned} f_t &= z_t + d\theta \sum_{i=1}^{t-1} (1-d+d\theta)^{i-1}(z_{t-i} + Lz_{i,t-i}) \\ &\quad + 2Nd\lambda \sum_{i=1}^{t-1} (1-d+d\theta)^{i-1} + \frac{2N}{k} (1-d+d\theta)^t. \end{aligned} \quad (18)$$

However, using the expressions for θ and λ (eqn 4),

$$d\lambda \sum_{i=1}^{t-1} (1-d+d\theta)^{i-1} + \frac{1}{k} (1-d+d\theta)^t = \frac{1}{k}, \quad \forall t > 0, \quad (19)$$

so that,

$$f_t = \frac{2N}{k} + z_t + d\theta \sum_{i=1}^{t-1} (1-d+d\theta)^{i-1}(z_{t-i} + Lz_{i,t-i}). \quad (20)$$

This frequency consists of the mean allele frequency, $2N/k$, plus a sum of independent normal random variables (each with a zero mean). The probability distribution of f_t is, therefore, the normal distribution (although it should be noted that successive values of f_t are correlated) with a mean of $2N/k$ and a variance (using eqns 7 and 11 for the variance of z)

$$\begin{aligned} \text{var}(f_t) &= \frac{2N(k-2)}{k^2} \left(1 + d^2\theta^2 \sum_{i=1}^{t-1} (1-d+d\theta)^{2(i-1)} \right) \\ &\quad + \frac{2Nl_p(1-l_p)}{k} d^2\theta^2 L^2 \sum_{i=1}^{t-1} (1-d+d\theta)^{2(i-1)}. \end{aligned} \quad (21)$$

The variance in steady state is given by $\text{var}(f_{\infty})$ and var_{ss} will approximately equal this value, so that:

$$\text{var}_{ss} \approx \text{var}(f_{\infty}) = \frac{2N(k-2)}{k^2} \left(1 + \frac{d^2\theta^2}{1-(1-d+d\theta)^2} \right)$$

$$+ \frac{2Nl_p(1-l_p)}{k} \frac{d^2\theta^2L^2}{1-(1-d+d\theta)^2}. \quad (22)$$

Without seed dormancy, $d = 1$, and, with a single plant size, $l_p = 0$ or 1 , so that for the NIL model (panmictic population), for example, this equation simplifies to:

$$\text{var}_{ss} = \frac{2N(k-2)}{k^2} \cdot \frac{1}{1-\theta^2}. \quad (23)$$

Comparison with the simulation results

Table 1 shows a comparison of the var_{ss} time series values with those obtained from the simulation of the NIL, D, L and DL models. Three extra runs of the NIL simulation model were carried out with larger values of N and k in order to extend this comparison (the simulation model requires the population size to be a multiple of the number of genotypes because its initial conditions are an equal number of each genotype). It is apparent that, in each case, there is good agreement between the value of var_{ss} given by the simulation study and that obtained from the time series eqn (22).

In the simulation study, each model was run 25 times and the average variance for these 25 runs calculated for each generation. The initial generation for each run contained an equal number of each possible genotype (and hence equal allele frequencies) and so the average variance during the approach to steady state can be compared with the time series $\text{var}(f_i)$ values from eqn (21). This was carried out for each of the models in Table 1 and, in each case, the time series values were close to those of the simulation model. Figure 2 shows the time

series and simulation values for two of the NIL models for the first 100 generations.

The time series analysis predicts that the allele frequencies will be approximately normally distributed. The greatest approximation in the time series analysis is probably that for the probability function, particularly if the variance of allele frequencies is high. A single run of the L simulation model, which has the largest variance of those models without the P and S factors, for 2200 generations was carried out and the frequencies of the 16 alleles recorded to test whether they are approximately normally distributed. Figure 3 shows the proportion of the steady-state values (generations 201–2200, 32 000 values) in intervals of 10 as well as the corresponding normal distribution curve with the same mean and variance (a mean of 480 and a variance, from the data of 9222). It is apparent that there is good agreement with the normal distribution. The skewness of the data is very small with a value of -0.19 .

Distribution of the generation variance values

The variance values for each generation, var_i , for a population in steady state will fluctuate around var_{ss} and the simulation study showed that these fluctuations are substantial. Equation (20) gives the probability distribution for the frequency of a single allele and, as discussed earlier, the frequencies of the k alleles in one generation are approximately equivalent to taking k values at random from this distribution. The frequency of an allele is normally distributed and the variance of a random sample from a normal distribution follows the chi-squared distribution given by,

$$\frac{(k-1)S^2}{\sigma^2} \sim \chi_{k-1}^2, \quad (24)$$

Table 1 The simulation and time series estimates of the average allele frequency variance, var_{ss} . For the D factor, $d = 0.7$ and for the L factor, $l_p = 0.2$ and $l_q = 20$ (see text for further details)

Model	Population size, N	No. of alleles, k	Simulation var_{ss}	Time series var_{ss}	Time series T_{ss}
NIL	3840	16	2914	2865	30
NIL	3968	32	3659	3438	66
NIL	7680	16	5683	5731	30
NIL	7936	32	7229	6875	66
D	3840	16	2080	2112	41
L	3840	16	9725	9871	30
DL	3840	16	6658	6959	43

Fig. 2 The time series (lines) and simulation (circles and triangles) average variance values during the approach to steady state for two of the NIL models.

Fig. 3 A comparison of the distribution of the allele frequencies of the L model in steady state with the normal distribution.

where S^2 is the variance of a sample of size k from a normal distribution with variance σ^2 . The variance of the sample is var_t and, in steady state, $\sigma^2 = \text{var}_{ss}$ so that the distribution of the steady state variance values is given by,

$$\text{var}_t \sim \frac{\text{var}_{ss}}{k-1} \chi_{k-1}^2 \quad (25)$$

and so the standard deviation of the steady state var_t values is:

$$SD(\text{var}_t) = \sqrt{\frac{2}{k-1}} \times \text{var}_{ss}. \quad (26)$$

In the simulation experiments, $k = 16$, so that, $SD(\text{var}_t) = 0.365\text{var}_{ss}$. A close linear relationship was found in the simulation study between these two variables with the regression equation obtained ($R^2 = 0.996$) being $SD(\text{var}_t) = 0.370\text{var}_{ss}$, which is very close to the predicted relationship.

Values of var_t for the L model in steady state (7500 values) were plotted to investigate whether they approximately followed the χ_{k-1}^2 distribution. Figure 4 shows the proportion of values in intervals of 250, as well as the χ_{k-1}^2 distribution with the same mean and standard deviation (a mean of 9725 and a standard deviation of 3438, see Table 1 in Brooks *et al.*, 1996). These values closely follow the χ_{k-1}^2 distribution as expected.

Equations (22) and (25), therefore, completely specify the distribution of the allele frequencies, var_t , for a population in steady state. The probability of the variance of a population in steady state falling

within any particular interval can therefore be calculated.

Discussion

Time to steady state and sensitivity analysis

The time series analysis allows a precise investigation of the time to steady state, T_{ss} , because the equations give an exact value for $\text{var}(f_t)$ which can be compared with $\text{var}(f_\infty)$, whereas an estimate is difficult with a simulation model because of its stochasticity. For example, T_{ss} could be considered as the number of generations for $\text{var}(f_t)$ to reach 99 per cent of $\text{var}(f_\infty)$ and these values are shown in Table 1. The time series values indicate that including overlapping generations and increasing the number of alleles in the population both increase the time for the model to reach steady state, whereas changes in the population size and the inclusion of variation in plant size do not appear significantly to affect T_{ss} . The effect of the number of alleles can be seen in Fig. 2. The time series equations could be generated for initial conditions other than equal frequency to investigate further the approach to steady state.

The average steady-state variance of allele frequency, var_{ss} , can be easily obtained for any set of parameter values from eqn (22), which allows the effects of varying the parameter values to be assessed. The value of var_{ss} increases linearly as N increases and approximately linearly as d increases. As k increases, var_{ss} increases (unless d is small when the approach is from above) towards an asymptote value of $Nd(1 + l_p(1 - l_p)L^2)$. The value of var_{ss} also

Fig. 4 A comparison of the distribution of the L model variance values with the chi-squared distribution with $k - 1$ degrees of freedom.

increases towards an asymptote as l_q increases, whereas as l_p varies, var_{ss} is a maximum when

$$l_p = \frac{1}{l_q + 1}. \quad (27)$$

At this value of l_p the total contribution of the large plants to future progeny equals the total contribution of the small plants.

The effect of overlapping generations and variation in plant size

Equation 20 allows a greater understanding of the effects of overlapping generations and variation in plant size. The allele frequency in the t th generation is the mean allele frequency plus the weighted sum of past random perturbations with a larger weight for recent perturbations. The dormancy factor, d , reduces the weight of recent perturbations and increases the weight of old perturbations so that the frequency is a more even sum of the past perturbations and so less variable. The effect of the variation in plant size is to add an additional source of randomness, namely the proportion of the plants containing the allele that are large, and so it increases the allele frequency variance.

Mutation

Past studies (e.g. Wright, 1939) have derived a probability distribution for the frequencies of all alleles which includes the effect of mutations. Such a probability distribution gives the distribution of the values of all the alleles in the population and has an increased proportion of alleles with low frequency because of the inclusion of recently mutated alleles. However, if the mutation rate u is small (certainly if $u \leq 10^{-6}$) this difference is negligible, and the effect of mutations on the overall allele frequency distribution can be ignored for most natural populations. Mutations could, perhaps, be added to the time series analysis by dividing the alleles in the population into classes, according to the length of time they have been present in the population, and generating a separate time series for recent mutations.

Limited pollen and seed dispersal

Limited pollen and seed dispersal introduces a spatial element to the system and causes frequency-dependent selection to operate locally rather than globally (Brooks *et al.*, 1996). It is difficult to incorporate these two factors into the time series analysis

because this is based on global allele frequencies. However, the simulation results indicated that the effect of the P and S factors on the variance of the allele frequencies is small, so the time series results for the DL model are close to those of the most realistic simulation model (the DPSL model), which includes limited seed and pollen dispersal.

It would appear from the computer simulation study that the most important effect of limited pollen and seed dispersal is to cause the alleles to cluster and that this is likely to have two opposing effects. The number of alleles existing locally is likely to be less than k which will increase the strength of frequency-dependent selection and, hence, decrease the allele frequency variance. However, the area occupied by a particular allele will vary randomly from one generation to the next, which adds a further random element to the system, causing greater variation in allele frequency. It would be possible to incorporate such an effect in the time series analysis by changing the number of selections made from N (in eqn 6) to $\hat{N}_t$, where $\hat{N}_t$ is itself a time series of a form, say,

$$\hat{N}_t = \omega \hat{N}_{t-1} + (1 - \omega) \frac{N\hat{k}}{k} + \hat{z}_t, \quad (28)$$

where $\hat{k}$ is the number of alleles that can be maintained locally ($N\hat{k}/k$ is an estimate of the area occupied by the allele, with the probability equation then based on $\hat{k}$ rather than k), ω represents the strength of the tendency of the size of the area to move towards the average and $\hat{z}_t$ is the random change in the size of the area. This leads to an additional term in the time series analysis. However, eqn (28) only represents one possible way in which the P and S factors affect the allele frequency. It is also very hard to determine realistic values for the variables in eqn (28), certainly without running a simulation model, which makes it inappropriate, at present, to include this mechanism in the analysis.

Conclusion

The frequency distribution for a single, established allele in a population with overlapping generations and variation in plant size has been derived as the normal distribution, with the variance able to be easily calculated. This enables alternative scenarios to be investigated and gives a greater understanding of the effects of different factors. The effects of changing the parameter values have been discussed. The time to steady state can be precisely calculated with overlapping generations and the number of

alleles in the population increasing the time to steady state but changes in the population size and the inclusion of variation in plant size having little effect. The following paper (Brooks *et al.*, 1997) sets out further results. The general approach used here can be applied to any variable that can be expressed linearly as a time series and so may have applications in other areas of population genetics.

References

- ARTHUR, A. E., GALE, J. S., AND LAWRENCE, M. J. 1973. Variation in wild populations of *Papaver dubium*. VII. Germination time. *Heredity*, **30**, 189–197.
- BROOKS, R. J., TOBIAS, A. M. AND LAWRENCE, M. J. 1996. The population genetics of the self-incompatibility polymorphism in *Papaver Rhoeas*. XI. The effects of limited pollen and seed dispersal, overlapping generations and variation in plant size on the variance of *S*-allele frequencies in populations at equilibrium. *Heredity*, **76**, 367–376.
- BROOKS, R. J., TOBIAS, A. M. AND LAWRENCE, M. J. 1997. A time series analysis of the population genetics of the self-incompatibility polymorphism. 2. Frequency equivalent population and the number of alleles that can be maintained in a population. *Heredity*, **79**, 361–364.
- CROSBY, J. L. 1966. Self-incompatibility alleles in the population of *Oenothera organensis*. *Evolution*, **20**, 567–579.
- EMERSON, S. 1939. A preliminary survey of the *Oenothera organensis* population. *Genetics*, **24**, 524–537.
- EMERSON, S. 1940. Growth of incompatible pollen tubes in *Oenothera organensis*. *Bot. Gaz.*, **101**, 890–911.
- EWENS, W. J. 1964. On the problem of self-sterility alleles. *Genetics*, **50**, 1433–1438.
- EWENS, W. J. AND EWENS, P. M. 1966. The maintenance of alleles by mutation — Monte Carlo results for normal and self-sterility populations. *Heredity*, **21**, 371–378.
- FISHER, R. A. 1958. *The Genetical Theory of Natural Selection*, 2nd edn. Dover Publications, New York.
- KIMURA, M. AND MARUYAMA, T. 1966. Personal communication. Mayo, H. 1966.
- LEVIN, D. A., RITTER, K. AND ELLSTRAND, N. C. 1979. Protein polymorphism in the narrow endemic *Oenothera organensis*. *Evolution*, **33**, 534–542.
- MAYO, O. 1966. On the problem of self-incompatibility alleles. *Biometrics*, **22**, 111–120.
- MORAN, P. A. P. 1962. *The Statistical Processes of Evolutionary Theory*. Clarendon Press, Oxford.
- WRIGHT, S. 1939. The distribution of self-sterility alleles in populations. *Genetics*, **24**, 538–552.
- WRIGHT, S. 1960. On the number of self-incompatibility alleles maintained in equilibrium by a given mutation rate in a population of a given size: a reexamination. *Biometrics*, **16**, 61–85.
- WRIGHT, S. 1964. The distribution of self-incompatibility alleles in populations. *Evolution*, **18**, 609–619.
- YOKOYAMA, S. AND HETHERINGTON, L. E. 1982. The expected number of self-incompatibility alleles in finite plant populations. *Heredity*, **48**, 299–303.
- YOKOYAMA, S. AND NEI, M. 1979. Population dynamics of sex-determining alleles in honey bees and self-incompatibility alleles in plants. *Genetics*, **91**, 609–626.