

A note on the symmetry constraints imposed by dominance in multiple locus genetic models

BY MONTGOMERY SLATKIN

*Department of Zoology, NJ-15, University of Washington,
Seattle, WA 98195, U.S.A.*

(Received 22 September 1978)

SUMMARY

A method is introduced that allows the simplification of the calculation of equilibrium solutions in multiple locus genetic models of a single infinite population. The method can be applied when the number of different fitnesses is equal to or less than one more than the number of independent allelic frequencies. The results are in terms of relationships – the symmetry constraints – between the gametic frequencies that must be satisfied at any boundary or internal equilibrium. The symmetry constraints are independent of the fitness values and of the recombination fractions. This can lead to some understanding of the equilibrium structure of a model when the full equilibrium solution is not obtained and reduces the number of independent variables in the calculations of the full equilibrium solutions. Examples of two locus models with two alleles at each locus and with two alleles at one locus and three at the other are discussed.

The inherent complexity of multiple locus genetic models that is imposed by mendelian inheritance precludes their complete description by a useful and general analytic theory. Only when many simplifying assumptions are made is it possible to understand either the transient or equilibrium structure of multiple locus models. The purpose of this note is to introduce and apply a new method for simplifying the computation of the equilibrium gametic frequencies when there are relatively few distinct phenotypic fitnesses. This method makes it possible to find the constraints on the gametic frequencies that must be satisfied at any equilibrium of a model without finding the complete analytic solution for the equilibrium gametic frequencies. When the method can be applied the results are very general in that they do not depend on either the fitness values or on the recombination fractions between the loci. In addition, information is obtained on both ‘internal’ equilibria, in which all gametes are present, and ‘boundary’ equilibria in which one or more of the gametes are absent.

Consider a dioecious species that has discrete, non-overlapping generations and that is effectively infinite in size, and let x_1, \dots, x_m be the frequencies in a generation of gametic types 1, ..., m being considered. Using notation introduced elsewhere (Slatkin, 1979), assume that there are n distinct phenotypes with different fitnesses, and let $u_{i, jk}$ be the fraction of the individuals of genotype jk ($j, k = 1, \dots, m$)

that are of phenotype i . The $u_{i,jk}$ provide a simple way of expressing all the dominance relationships among the gametes. In most applications the $u_{i,jk}$ are either 0 or 1, but more general cases with fractional values describing 'somatic polymorphisms' can be used (Slatkin, 1979). We will assume throughout that the $u_{i,jk}$ are constant and, since the list of phenotypes is exhaustive,

$$\sum_{i=1}^n u_{i,jk} = 1 \quad (1)$$

for all pairs jk .

Assume that the relative (or absolute) fitnesses of the different phenotypes in the population are W_i ($i = 1, \dots, n$). Usually the W_i are assumed to be constants, but here we need to assume only that, if the W_i depend on population density or genotypic frequencies, they be different for all feasible densities or frequencies.

With the above notation, the basic recursion equations for the gametic frequencies have the following form:

$$\bar{W}x'_j = x_jF_j + R_j \quad (2)$$

for $j = 1, \dots, m$, where the prime indicates the succeeding generation, F_j is the marginal fitness of the j th gametic type, and R_j represents all the terms that appear from the various types of recombination that can occur for the number of loci and numbers of alleles being considered. The recombination terms will not be a part of the results, so we will not be more specific about the details of their form.

While the basic equations for a multiple locus model depend on the recombination fractions, there are certain combinations of those equations that do not and that can be expressed as homogeneous functions of the marginal fitnesses alone. One such equation is obtained by summing (2) over all j to obtain.

$$\bar{W} \sum_{j=1}^m x'_j = \sum_{j=1}^m x_j F_j = \bar{W}. \quad (3)$$

Other equations can be found by adding the gametic frequencies in such a way that recursion equations for each of the allelic frequencies are obtained. The actual combinations of the equations depend on the number of loci and the number of alleles at each locus, but there is the same number of such linearly independent equations as there are independent allele frequencies. The set of equations that form the starting point for the analysis contains one more equation (eqn. 3) than the number of independent allelic frequencies. The resulting set of equations is not sufficient to determine the equilibrium allelic or gametic frequencies, but we will see that some information can be obtained about the equilibrium frequencies from that set.

We will proceed by considering some examples that illustrate both the method itself and the type of the results obtained. Consider two linked loci with two alleles at each. Using the standard notation (Karlín & Feldman, 1970), let the set of frequencies (x_1, x_2, x_3, x_4) correspond to the set of gametes $(A_1B_1, A_1B_2, A_2B_1,$

A_2B_2) (Karlin & Feldman, 1970). With this notation, the basic recursion equations, (2), are

$$\left. \begin{aligned} \bar{W}x'_1 &= x_1F_1 - rD\bar{W}, \\ \bar{W}x'_2 &= x_2F_2 + rD\bar{W}, \\ \bar{W}x'_3 &= x_3F_3 + rD\bar{W}, \\ \bar{W}x'_4 &= x_4F_4 - rD\bar{W}, \end{aligned} \right\} \quad (4)$$

where \bar{W} is the fitness of the double heterozygote, $D (= x_1x_4 - x_2x_3)$ is the linkage disequilibrium between the two loci, and r is the recombination fraction (Lewontin & Kojima, 1960). If we add the equations in (4) in the way described above, we obtain the three equations

$$\left. \begin{aligned} \bar{W}(x'_1 + x'_2 + x'_3 + x'_4) &= x_1F_1 + x_2F_2 + x_3F_3 + x_4F_4, \\ \bar{W}(x'_1 + x'_2) &= x_1F_1 + x_2F_2, \\ \bar{W}(x'_1 + x'_3) &= x_1F_1 + x_3F_3, \end{aligned} \right\} \quad (5)$$

where the first equation is (3) and the second two are the recursion equations for the frequencies of the A_1 and B_1 alleles. There are obviously other combinations of (4) that will have the effect of eliminating the recombination term but all such linear combinations can be derived from (5).

To consider the equilibrium solutions for the gametic frequencies, we assume that $x'_j = x_j$ for $j = 1, \dots, 4$, and rewrite (5) as

$$\left. \begin{aligned} x_1(F_1 - \bar{W}) + x_2(F_2 - \bar{W}) + x_3(F_3 - \bar{W}) + x_4(F_4 - \bar{W}) &= 0, \\ x_1(F_1 - \bar{W}) + x_2(F_2 - \bar{W}) &= 0, \\ x_1(F_1 - \bar{W}) + x_3(F_3 - \bar{W}) &= 0. \end{aligned} \right\} \quad (6)$$

We write the marginal gametic fitnesses, F_j , in terms of the phenotypic fitnesses, W_i , and the $u_{i,jk}$, as

$$F_j = \sum_{i=1}^n \sum_{k=1}^m x_k u_{i,jk} W_i. \quad (7)$$

Finally, we can substitute (7) in (6) to obtain

$$\left. \begin{aligned} \sum_{i=1}^n \left\{ \sum_{j,k=1}^4 x_j x_k u_{i,jk} \right\} (W_i - \bar{W}) &= 0, \\ \sum_{i=1}^n \left\{ x_1 \sum_{k=1}^4 x_k u_{i,1k} + x_2 \sum_{k=1}^4 x_k u_{i,2k} \right\} (W_i - \bar{W}) &= 0, \\ \sum_{i=1}^n \left\{ x_1 \sum_{k=1}^4 x_k u_{i,1k} + x_3 \sum_{k=1}^4 x_k u_{i,3k} \right\} (W_i - \bar{W}) &= 0, \end{aligned} \right\} \quad (8)$$

where we have used (1) and the fact that the gametic frequencies sum to one in order to express each equation in terms of $(W_i - \bar{W})$.

The set of equations in (8) are of the form

$$Ty = 0, \quad (9)$$

where T is a $3 \times n$ matrix and y is a n -vector with elements $(W_i - \bar{W})$. From the standard theory of linear, algebraic equations (e.g. Hoffman & Kunze, 1961) we

know that, if $n \leq 3$, the system (9) has non-trivial solutions (i.e. not all elements of y equal to zero) only if the rank of T is less than n . That is, if $n = 3$, then T is a square matrix and its determinant must be zero, because we have assumed that the W_i are different and thus cannot all equal \bar{W} . Similarly, if $n = 2$, then T is a 3×2 matrix and the determinants of the three 2×2 submatrices it contains must be zero. As we can see from (8), the elements of T are functions only of the x_j and the $u_{i,jk}$ so we have obtained conditions that must be satisfied at the equilibrium values of the x_j that depend only on the dominance relationships and not on the values of the W_i or the recombination fraction. Thus, if the number of distinct phenotypic classes is less than or equal to one plus the number of independent allele frequencies, there are constraints on the equilibrium gametic frequencies.

Table 1. Summary of results for 2-locus model with two alleles at each locus

Case	Genotypic fitnesses			Subscripts (i, jk) of the $u_{i,jk}$ for which $u_{i,jk} = 1$. All other $u_{i,jk} = 0$	Constraints on x_i , the equilibrium gametic frequencies	
	B_1B_1	B_1B_2	B_2B_2			
i	A_1A_1	W_1	W_2	W_1	(1, 11) (1, 22) (1, 33) (1, 44) (2, 12) (2, 21) (2, 13) (2, 31) (2, 24) (2, 42) (2, 34) (2, 43) (3, 14) (3, 41) (3, 23) (3, 32)	Either $x_1 = x_4, x_2 = x_3$ or $x_1 + x_4 = x_2 + x_3 = \frac{1}{2}$
	A_1A_2	W_2	W_3	W_2		
	A_2A_2	W_1	W_2	W_1		
ii	A_1A_1	W_1	W_2	W_3	(1, 11) (1, 44) (2, 12) (2, 13) (2, 14) (2, 21) (2, 23) (2, 24) (2, 31) (2, 32) (2, 34) (2, 41) (2, 42) (2, 43) (3, 22) (3, 33)	Either $x_1 = x_4$ or $x_2 = x_3$
	A_1A_2	W_2	W_2	W_2		
	A_2A_2	W_3	W_2	W_1		
iii	A_1A_1	W_1	W_2	W_1	(1, 11) (1, 22) (1, 33) (1, 44) (2, 12) (2, 13) (2, 14) (2, 21) (2, 23) (2, 24) (2, 31) (2, 32) (2, 34) (2, 41) (2, 42) (2, 43)	$x_1 = x_4$ and $x_2 = x_3$
	A_1A_2	W_2	W_2	W_2		
	A_2A_2	W_1	W_2	W_1		
iv	A_1A_1	W_1	W_2	W_1	Same $u_{i,jk}$ as case (i) except $u_{1,44} = f$ and $u_{2,44} = 1 - f$, where $1 - f : W_2 \quad 0 \leq f \leq 1$	Either $x_2 = x_3$ or $x_1 = \frac{1}{4}\{1 \pm \sqrt{(1+8[(1+f)x_4^2 - x_4]}\}$
	A_1A_2	W_2	W_3	W_2		
	A_2A_2	W_1	W_2	$f : W_1$		

To illustrate this result, we consider the first example (case i) in Table 1, in which the fitnesses are determined by the number of heterozygous loci. The completely symmetric, multiplicative model is a special case of this. The locations of the W_i in the 3×3 table of genotypic fitnesses completely specify the values of the $u_{i,jk}$. The subscripts of the non-zero $u_{i,jk}$ are included to show the relationship between the usual notation and notation used here. With those values of the $u_{i,jk}$, the matrix T can be found from (8) to be

$$T = \begin{pmatrix} x_1^2 + x_2^2 + x_3^2 + x_4^2 & 2x_1x_2 + 2x_1x_3 + 2x_2x_4 + 2x_3x_4 & 2(x_1x_4 + x_2x_3) \\ x_1^2 + x_2^2 & 2x_1x_2 + x_1x_3 + x_2x_4 & x_1x_4 + x_2x_3 \\ x_1^2 + x_3^2 & x_1x_2 + 2x_1x_3 + x_3x_4 & x_1x_4 + x_3x_2 \end{pmatrix}. \quad (10)$$

While the direct evaluation of the determinant of (10) would be a difficult and tedious problem in elementary algebra, there are some simplifying properties of T that follow from its basic form and that make the problem less difficult than it first appears to be. If we consider (8), equation (1) implies that if we add the columns of T , the sum of each row is simply the sum of gametic frequencies added to obtain (6). Since we can add any column of a matrix to any other without changing the determinant, we can replace any column of T by the sum of the appropriate gametic frequencies. This can be verified by adding the columns in (10). To simplify the calculations, we should always choose the column with the most terms in the sum. Thus, with (10) we would replace the middle column. Also, we can subtract any row from the first without altering the determinant, so T can be simplified further. Clearly, the choice of the row would depend on the algebraic structure of T . In the present example, either the second or third row could be used.

With these two simplifications, which can be used for any selection model, we are left with the problem of finding the conditions under which the determinant of

$$T' = \begin{pmatrix} x_3^2 + x_4^2 & x_3 + x_4 & x_1x_4 + x_2x_3 \\ x_1^2 + x_2^2 & x_1 + x_2 & x_1x_4 + x_2x_3 \\ x_1^2 + x_3^2 & x_1 + x_3 & x_1x_4 + x_2x_3 \end{pmatrix} \tag{11}$$

is zero. We can further simplify (11) by using more elementary row operations to obtain

$$\det T' = (x_1 - x_4)(x_2 - x_3) \det \begin{pmatrix} x_2 + x_3 & 1 & 0 \\ x_1^2 + x_2^2 & x_1 + x_2 & x_1x_4 + x_2x_3 \\ x_1 + x_4 & 1 & 0 \end{pmatrix} \tag{12}$$

which can be easily evaluated as

$$\det T' = (x_1 - x_4)(x_2 - x_3)(x_1 + x_4 - x_2 - x_3)(x_1x_4 + x_2x_3) = 0. \tag{13}$$

From (13) we can obtain all of the possible constraints on the gametic frequencies at an equilibrium under a selection model of the form in Table 1, case i. At equilibrium, at least *one* of the following conditions must be satisfied:

$$\begin{aligned} x_1 &= x_4, & x_2 &= x_3, \\ x_1 + x_4 &= x_2 + x_3 = \frac{1}{2}. \end{aligned} \tag{14}$$

In addition, we find the trivial equilibria in which one locus or both loci are fixed (e.g. $x_1 = x_2 = 0, x_3 = x_4 = 0$, etc.). There is no guarantee that there is always a solution satisfying one of the constraints in (14) or that such a solution is stable, but we can be sure that any equilibrium solution, including any solution for which one or more of the x_j are zero, must satisfy at least one of the constraints. If that is not the case, then the condition imposed by the structure of (9) is violated. The constraints in (14) do not provide the values of the equilibrium solutions, which usually depend on r and the W_i , but they do provide a way to reduce the number of independent variables in the equations for the equilibrium solutions from three to two with the certainty that no possible solutions will be missed.

It is important to emphasize what is actually being claimed for this method. The algebraic manipulations and the form of the terms in (13) are similar to those in Karlin & Feldman (1970) in which the full analytic solutions for a more general class of fitness models is obtained by using a set of transformations on the gametic frequencies. For the class of symmetric viability models considered in that paper, those transformations yield the complete set of equilibrium solutions, and it would not be necessary to use the method described here. However, this method still has some value for the symmetric models by showing in a simple way what features of the fitness models give rise to different sets of equilibrium solutions. This is illustrated by comparing cases i and ii in Table 1. Both of these models are in the class considered by Karlin & Feldman (1970), but we can see that a slightly different arrangement of the fitnesses changes the types of equilibrium solutions that are possible. In particular, one type of solution possible in case i is not in case ii. Of course, this can be obtained from the complete analysis in Karlin & Feldman (1970), but it is much easier to do so using the present method. Also, this method guarantees that the exclusion of one class of solutions in case ii does not depend on the values of r or the W_i , and it can be applied to two-locus models that do not fall in the class of symmetric models of Karlin & Feldman (1970) (e.g. case iv, Table 1). However, this method does not replace a full analysis of the equilibrium solutions. Instead it can simplify that analysis by reducing the number of independent variables and revealing some features of the solutions that are due only to the dominance relationships.

The results for some other selection schemes are also shown in Table 1. Case iii illustrates the kinds of result that are obtained when there are fewer values of W_i than independent equations. The matrix T in (8) for this case is now a 3×2 matrix which can be put in the form

$$T' = \begin{pmatrix} x_3^2 + x_4^2 & x_3 + x_4 \\ x_1^2 + x_2^2 & x_1 + x_2 \\ x_1^2 + x_3^2 & x_1 + x_3 \end{pmatrix}. \quad (15)$$

Since the rank of T' in (15) must be one, the determinants of all three 2×2 submatrices must be zero at the equilibrium values of the x_j . That condition can be satisfied only when $x_1 = x_4$ and $x_2 = x_3$. The final case, case iv, shows the increase in complexity of the results when even a slight deviation from a simple symmetric viability scheme is used. The genotype fitnesses are the same as in case i except that a fraction f of the $A_2A_2B_2B_2$ individuals are in phenotypic class 1 and a fraction $1-f$ are in class 2. The constraints on the equilibrium solution are more complex and probably would not be apparent from an inspection of the basic recursion formulae. They do, of course, reduce to the constraints found in case i when $f = 1$.

This technique can be used for models of more alleles at one or more loci. The greater algebraic complexity of such problems usually means that it is not possible to obtain complete algebraic solutions for all of the symmetry constraints, but this method can still provide some information about the solutions for models

that are inherently difficult to analyse. As an illustration of both the use and the possible limitations of this method, we can consider the model analysed by Feldman *et al.* (1975) of a two-locus system with three alleles (A_1, A_2, A_3) at the first locus and two alleles (B_1, B_2) at the second. Following their notation, we let the set $(x_1, x_2, x_3, x_4, x_5, x_6)$ denote the frequencies of the gametes ($A_1B_1, A_1B_2, A_2B_1, A_2B_2, A_3B_1, A_3B_2$). We will consider a generalization of the case analysed in that paper and assume that the genotypic fitnesses depend only on the number of homozygous loci and not on the identity of the alleles. We can denote the fitness of the double heterozygotes as W_1 , of the single heterozygotes as W_2 , and of the complete homozygotes as W_3 . In the multiplicative fitness model analysed by Feldman *et al.* (1975, equation 1), $W_i = w^{(i-1)}$ where w is a given constant.

This specification of the fitnesses determines the subscripts of all 36 non-zero $u_{i,jk}$. There are three independent allele frequencies so we can obtain four independent equations for the gametic frequencies. Since there are only three phenotypic classes, we expect there to be several symmetry constraints imposed on the equilibrium frequencies. The set of independent equations that appears to be the easiest to manipulate is obtained by adding the recursion equations for x_1 and x_2 ; x_3 and x_4 ; x_5 and x_6 ; and x_1, x_3 and x_5 . This provides the recursion formulae for A_1, A_2, A_3 and B_1 . The matrix T' that is analogous to (11) is

$$T' = \begin{pmatrix} x_1x_4 + x_1x_6 + x_2x_3 + x_2x_5 & x_1 + x_2 & x_1^2 + x_2^2 \\ x_1x_4 + x_2x_3 + x_3x_6 + x_4x_5 & x_1 + x_3 & x_1^2 + x_3^2 \\ x_1x_6 + x_2x_5 + x_3x_6 + x_4x_5 & x_5 + x_6 & x_5^2 + x_6^2 \\ x_1x_4 + x_1x_6 + x_3x_2 + x_3x_6 & x_1 + x_3 + x_5 & x_1^2 + x_3^2 + x_5^2 \\ & + x_5x_2 + x_5x_4 \end{pmatrix}, \tag{16}$$

and the general theory tells us that its rank must be 2 or less at the equilibrium values of the x_j . Clearly, the evaluation of the four 3×3 determinants contained in (16) is not trivial, but some information can be obtained from (16) without evaluating all the determinants. By first subtracting the second and third column from the first and using the fact that the frequencies sum to one, we can see that the rank of (16) is the same as the rank of

$$\begin{pmatrix} (x_2 - x_1)(x_2 + x_3 + x_5) & x_1 + x_2 & x_1^2 + x_2^2 \\ (x_4 - x_3)(x_1 + x_4 + x_5) & x_3 + x_4 & x_3^2 + x_4^2 \\ (x_6 - x_5)(x_1 + x_3 + x_6) & x_5 + x_6 & x_5^2 + x_6^2 \\ -x_1(x_2 + x_3 + x_5) - x_3(x_1 + x_4 + x_5) & x_1 + x_3 + x_5 & x_1^2 + x_3^2 + x_5^2 \\ & -x_6(x_1 + x_3 + x_6) \end{pmatrix}. \tag{17}$$

From (17) we can easily establish that there are certain kinds of symmetric solutions that might be expected but are not permitted under this selection model. For example, there is no solution of the form $x_1 = x_2, x_3 = x_4$ and $x_5 = x_6$, except for the 'central' equilibrium at which all six frequencies equal 1/6. This was asserted without proof by Feldman *et al.* (1975) who said it followed from the structure of the model. This method shows how it follows from the structure. With somewhat more difficulty, it can be shown that there are no solutions with

any two of those pairs of frequencies equal but the third unequal. It is also possible to investigate the various kinds of boundary equilibria that are possible. However, a complete analysis would depend on the evaluation of subdeterminants that do not appear to possess any simple form although computer languages exist that could simplify the calculations. As a consequence, the power of this method depends on the actual dominance relationships among the alleles and in some cases, such as the one considered above, all of the symmetry constraints cannot easily be found. Nevertheless, the analysis of multiple locus genetic models is sufficiently difficult that this technique can be useful in some problems, and, when it is, it provides very general results that can lead to significant simplification of the problem of finding equilibrium gametic frequencies.

I thank J. Felsenstein, R. Lande and G. J. Thomson for helpful discussions of this topic. This research is supported by a NIH Research Career Development Award No. K04-GM00118.

REFERENCES

- FELDMAN, M. W., LEWONTIN, R. C., FRANKLIN, I. R. & CHRISTIANSEN, F. B. (1975). Selection in complex genetic systems, III: An effect of allele multiplicity with two loci. *Genetics* **79**, 333–347.
- HOFFMAN, K. & KUNZE, R. (1961). *Linear Algebra*. Englewood Cliffs, New Jersey: Prentiss Hall.
- KARLIN, S. & FELDMAN, M. W. (1970). Linkage and selection: Two locus symmetric viability model. *Theoretical Population Biology* **1**, 39–71.
- LEWONTIN, R. C. & KOJIMA, K. (1960). The evolutionary dynamics of complex polymorphisms. *Evolution* **14**, 458–472.
- SLATKIN, M. (1979). The evolutionary response to frequency and density dependent interactions. *American Naturalist*. (In the Press.)