LIMITS TO THE RELATIONSHIP AMONG RECOMBINATION, DISEQUILIBRIUM AND EPISTASIS IN TWO-LOCUS MODELS

ALAN HASTINGS

Department of Mathematics and Division of Environmental Studies, University of California, Davis, California 95616

Manuscript received August 28, 1985
Revised copy accepted January 13, 1986

ABSTRACT

I determine limits to the equilibrium relationship among epistasis, recombination and disequilibrium in two-locus, two-allele models using linear programming techniques. I show that when allele frequencies are one-half at each locus, the symmetric model is the fitness pattern that generates the most disequilibrium for the smallest level of epistasis. When allele frequencies deviate from one-half, much larger levels of epistasis are required to generate similar levels of disequilibrium. I determine the level of epistasis required to generate observed significant levels of disequilibrium in natural populations. The overall implication is that disequilibrium will be large at equilibrium only between strongly interacting, closely linked loci.

As summarized by Karlin (1975) and Ewens (1979), much of our intuition on the equilibrium behavior of multilocus population genetic models, and even two-locus models, has come primarily from a detailed analysis of the classical models—the additive, symmetric and multiplicative cases. To see how well intuition gained from these models carries over to general fitnesses, it is important to understand the equilibrium behavior of multilocus models which do not fall in one of these classical categories. One approach that can be used to obtain general results is to pick fitnesses at random and simulate the resulting multilocus models (e.g., Karlin and Carmelli 1975).

Other approaches to understanding the static behavior of multilocus models include perturbation approaches which allow one to understand the behavior of models which are "close" to the classical ones, as in Hastings (1985a, 1986). In the classical additive model, the only equilibrium solutions have no disequilibrium. Deviations away from the additive model can be measured using epistasis. It is not clear, however, how well the perturbation results (Hastings 1985a, 1986) for weak epistasis carry over to larger levels of epistasis.

Another approach would be to find limits to the relationship among selection, disequilibrium and recombination (Hastings 1981, 1984). If there is no selection, the disequilibrium declines to zero, so unless selection is strong enough, disequilibrium must be small at equilibrium. Similarly, the absence of additive epistasis implies no disequilibrium. Thus, another way to look at the
relationship among selection, recombination and disequilibrium in a general setting is to look at the relationship among epistasis, recombination and disequilibrium. Here, I shall use a linear programming approach (HASTINGS 1981, 1983) to answer the question, how large must epistasis be to lead to a given level of disequilibrium?

The intuition we currently have regarding this question comes primarily from the extensive studies of the symmetric model (LEWONTIN and KOJIMA 1960; KARLIN and FELDMAN 1970; FELDMAN and LIBERMAN 1979; HASTINGS 1985b). Thus, another important question is, how well does the symmetric model reflect the possible relationships among recombination, epistasis and disequilibrium? Note that the symmetric equilibria of the symmetric model, which provide most of the intuition about disequilibrium, require that allele frequencies at both loci be one-half.

A third important question is whether the levels of epistasis required to generate disequilibrium are so high that only very strongly interacting loci would be likely to lead to high levels of disequilibrium. Most attempts to detect linkage disequilibrium in natural populations have been surveys of randomly chosen loci (e.g., LANDLEY, TOBARI and KOJIMA 1974) which have found little evidence for disequilibrium. However, linkage disequilibrium has been detected by ROBERTS and BAKER (1973) and BAKER (1975) in Drosophila montana between esterase loci, which may interact to the extent that epistasis is likely.

Finally, one can use the results of the current paper to draw inferences from data such as that collected for Drosophila montana. How strong must the interaction be in observed cases of linkage disequilibrium? I report answers to this question with regard to Drosophila montana and cats, below.

BACKGROUND AND METHODS

The model to be investigated here is the standard two-locus, two-allele, discrete time formulation [see KARLIN (1975) and EWENS (1979) for reviews]. Let there be two loci with alleles A and a at the first locus and alleles B and b at the second locus. Let the frequency of the four chromosomes AB, Ab, aB and ab be given by \( x_1, x_2, x_3 \) and \( x_4 \), respectively. Define the disequilibrium, \( D \), as \( x_4 - x_2x_3 \).

Let the recombination rate between the two loci be \( r \). Finally, define the marginal gametic fitnesses

\[
\begin{align*}
    w_i &= \sum_{j=1}^{4} x_j w_{ij}, \text{ for } i = 1 \text{ to } 4, \\
    \bar{w} &= \sum_{i=1}^{4} x_i w_i,
\end{align*}
\]

and the mean fitness as

\[
\bar{w} = \sum_{i=1}^{4} x_i w_i,
\]

where \( w_{ij} \) is the fitness (viability) of the individual with gametes for which the frequencies are given by \( x_i \) and \( x_j \). Then, the dynamics of this system is given
EPISTASIS AND DISEQUILIBRIUM

TABLE

<table>
<thead>
<tr>
<th>Genotype</th>
<th>AA</th>
<th>Aa</th>
<th>aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td>$1 - a_1 - b_1 + e_1$</td>
<td>$1 - b_1$</td>
<td>$1 - a_2 - b_1 - e_3$</td>
</tr>
<tr>
<td>Bb</td>
<td>$1 - a_1$</td>
<td>$1$</td>
<td>$1 - a_3$</td>
</tr>
<tr>
<td>bb</td>
<td>$1 - a_1 - b_3 - e_2$</td>
<td>$1 - b_3$</td>
<td>$1 - a_5 - b_3 + e_4$</td>
</tr>
</tbody>
</table>

The four parameters $e_i$, with the $i$ corresponding to the labeling of the variable $x_i$ denoting the frequencies, are measures of the epistatic interaction between alleles at loci $A$ and $B$.

by (assuming $w_{14} = w_{23}$),

$$x_i' = (x_i w_i \pm r D w_{14}) / \bar{w}, \quad (3)$$

where the sign is minus for $i = 1$ or 4 and is plus otherwise.

Next comes the translation, into mathematical terms, of the verbal questions given in the introduction. First, display the fitnesses $w_{ij}$ in the $3 \times 3$ form as shown in Table 1, where the fitness of each genotype consists of an additive component plus an epistatic component. Note, following FISHER (1918), that the four epistatic parameters, $e_i$, can be expressed as

$$e_i = w_{1i} - w_{2i} - w_{3i} + w_{4i}, \quad i = 1 \text{ to } 4. \quad (4)$$

There is no simple way to express these four measures of epistasis as one quantity; an average does not make sense because of sign cancellation. I shall thus give the following two definitions of total epistasis, admittedly motivated partly because they lead to a linear programming problem that I can solve and partly because they allow simple comparison to the analytic results in HASTINGS (1985a).

**Definition 1:** The total epistasis ($e_{tot}$) in the model (1) to (4) is defined as the maximum of the absolute value of the four quantities $e_p q.$, where $p$ and $q.$ are the frequencies of the two alleles in the gamete for which the frequency is $x_i$. The missing subscripts in $p.$ are either $A$ or $a$ and in $q.$ are either $B$ or $b$, as appropriate.

This definition gives a total epistasis in which weights are assigned to gametes according to their frequency when there is no disequilibrium. An alternate possibility without weights is the following.

**Definition 2:** The total epistasis in the model (1) to (4) is defined as the maximum of the absolute value of the four quantities $e_i$.

Using either definition of epistasis just given, the following is a linear programming problem.

**Linear programming problem:** Given the equilibrium values for the disequilibrium and allele frequencies in (1) to (4), minimize the total epistasis.

In solving the linear programming problem, the additive fitness components and the epistatic parameters are allowed to vary, while the equilibrium is held fixed. Also, stability of the equilibrium is not required. However, this simply means that a lower bound on epistasis is obtained for a stable equilibrium. Finally, as expected, the strength of epistasis required increases monotonically.
as the disequilibrium is moved away from zero. Hence, it is sensible to invert
the problem and then express the answers as the maximum disequilibrium
possible for a given level of epistasis.

Answers from this problem can lead to unrealistic values for the fitnesses,
so it is also useful to define the following alternate problem, which constrains
the fitnesses to agree with a priori biological assumptions. One reasonable
constraint might be that the double heterozygote be the most fit genotype.
Another reasonable constraint would be a limit on the strength of selection,
say restricting all fitnesses to lie between 1−s and 1+s. Both of these restric-
tions are linear constraints on the fitnesses. These can be included in the
problem to be solved, leading to the following.

**Alternate linear programming problem:** Given the equilibrium values for
the disequilibrium and allele frequencies in (1) to (4), and any linear constraints
on the fitnesses, minimize the total epistasis.

Either of these linear programming problems can be solved using standard
techniques [see HASTINGS (1981) for more details in the context of a similar
problem].

**RESULTS**

The answers obtained actually do not depend critically on which definition
of epistasis is used, nor on the form of the constraints; therefore, I shall limit
the discussion to definition 1 for epistasis and to the linear programming
problem in which all fitnesses are constrained to be less than one (and greater
than zero), so the double heterozygote is the most fit. I solved the linear
programming problem for allele frequencies ranging from 0.5 to 0.9 (in steps
of 0.1) at each locus and for disequilibrium ranging from the minimum to the
maximum value possible, in ten equally spaced steps, omitting the endpoints
(since epistasis would be infinite there). As noted in HASTINGS (1981), answers
for other allele frequencies can be obtained from these. I also used two values
for r, namely 0.1 and 0.01; however, except when disequilibrium was so large
that the constraint that all fitnesses be positive came into play, the ratio of the
minimum epistasis to r was independent of r. Minimum epistasis was propor-
tional to r; this also held for a number of other cases that I examined. Thus,
in the figures, I shall deal only with the case r = 0.1. First, I shall summarize
information about the entire array of solutions and then examine two param-
eter combinations in more detail.

I shall first compare the relationships between epistasis and disequilibrium
as estimated in HASTINGS (1985a), using perturbation results valid for small
levels of epistasis with that obtained here. Equation (26) in HASTINGS (1985a)
provides the following estimate for weak epistasis:

\[ rD(p_a p_b q_a q_b)^{-1} \approx 2x_1 - 3x_2 + 3x_3 - 2x_4. \]

(5)

From this equation, using definition 1 for total epistasis, \( e_{tot} \), one obtains

\[ e_{tot} \approx rD(p_a p_b q_a q_b)^{-1/4}. \]

(6)

Note that, even in this estimate, not only does the epistasis go up markedly as
Figure 1.—Minimum levels of epistasis, $e_{min}$, as a function of disequilibrium, at equilibrium, for allele frequencies of 0.5 at each locus, indicated by boxes (□). The epistasis is calculated using definition 1. The recombination rate, $r$, is 0.1. The line is the estimate (6) obtained using a perturbation approach valid for weak epistasis.

the allele frequencies deviate from one-half, but so does the correlation coefficient (Hastings 1985a).

Result 1: For all the parameter values here, the estimate (6) provides an upper limit to the level of disequilibrium obtained for a given level of epistasis (or, correspondingly, a lower limit to the level of epistasis necessary to generate a given level of disequilibrium). How good this estimate is will be discussed below.

I shall now turn to two specific, illustrative choices for allele frequencies. Initially, I shall discuss the case with all alleles at frequency one-half. In this case, two important results emerge (see Figure 1). First, the result from the weak epistasis case, (6), provides a good estimate for the relationship between epistasis and disequilibrium, unless disequilibrium and epistasis are large. More importantly, the fitness patterns and equilibria that give the smallest level of epistasis for a given level of disequilibrium correspond to the symmetric equilibrium of a symmetric model [see Karlin and Feldman (1970) for a review]. Let the fitnesses be given as in Table 2, and let us assume that the fitnesses satisfy the further constraint that

$$2(\beta + \gamma) = \alpha + \delta.$$  

(7)

Then, letting

$$m = \delta - \alpha$$

(8)

the following solution is obtained (see Karlin and Feldman 1970):

$$D = r/m - 0.25 (1 + 16r^2/m^2)^{1/4}.$$  

(9)
### TABLE 2

Fitnesses for the symmetric two-locus, two-allele model

<table>
<thead>
<tr>
<th></th>
<th>$AA$</th>
<th>$Aa$</th>
<th>$aa$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$BB$</td>
<td>$1-\delta$</td>
<td>$1-\gamma$</td>
<td>$1-\alpha$</td>
</tr>
<tr>
<td>$Bb$</td>
<td>$1-\beta$</td>
<td>$1$</td>
<td>$1-\beta$</td>
</tr>
<tr>
<td>$bb$</td>
<td>$1-\alpha$</td>
<td>$1-\gamma$</td>
<td>$1-\delta$</td>
</tr>
</tbody>
</table>

**Figure 2.**—Minimum levels of epistasis, $e_{\text{min}}$, as a function of disequilibrium, at equilibrium, for allele frequencies of 0.8 at each locus, indicated by boxes (□). The epistasis is calculated using definition 1. The recombination rate, $r$, is 0.1. The line is the estimate (6) obtained using a perturbation approach valid for weak epistasis.

It is this solution and these fitnesses which provide the optimal solution to the linear programming problem.

**Result 2:** For allele frequencies which are one-half at each locus, the symmetric model is the model that leads to the most disequilibrium for the least epistasis.

For a case in which the most common allele at each locus has a frequencies of 0.8 at both loci, a different picture emerges, which is typical of cases in which the allele frequencies are not one-half (see Figure 2), as given in the following two results.

**Result 3:** When allele frequencies are not one-half at each locus, the symmetric model is no longer the model that generates the most disequilibrium for a given level of epistasis.

**Result 4:** When allele frequencies are not one-half at each locus, the level of epistasis required to generate a given level of disequilibrium is greatly underestimated by the formula obtained for a weak epistasis. Since even this formula suggests that levels of disequilibrium and correlation coefficients will be lower.
TABLE 3
Minimum epistasis calculated for significant nonzero values of $D$ reported in ROBERTS and BAKER (1973) and BAKER (1975) for esterase loci in Drosophila montana

<table>
<thead>
<tr>
<th>Pair of loci (<em>A</em> listed first)</th>
<th>Year</th>
<th>$p_A$</th>
<th>$p_B$</th>
<th>$D$</th>
<th>$\varepsilon_{\text{min}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1$ and $2$</td>
<td>1970</td>
<td>0.274</td>
<td>0.637</td>
<td>-0.109</td>
<td>0.006</td>
</tr>
<tr>
<td>$1$ and $2$</td>
<td>1973</td>
<td>0.236</td>
<td>0.627</td>
<td>-0.081</td>
<td>0.004</td>
</tr>
<tr>
<td>$3$ and $4$</td>
<td>1970</td>
<td>0.568</td>
<td>0.346</td>
<td>-0.163</td>
<td>0.017</td>
</tr>
<tr>
<td>$3$ and $4$</td>
<td>1973</td>
<td>0.537</td>
<td>0.398</td>
<td>-0.162</td>
<td>0.011</td>
</tr>
</tbody>
</table>

Definition 2 was used for total epistasis, so $\varepsilon_{\text{min}}$ represents the largest value of $e_i$. The allele frequencies are $p_A$ and $p_B$. Fitnesses were constrained to all be less than one, and a value of 0.002 was used for $r$ in all cases.

TABLE 4
Minimum epistasis calculated for significant nonzero values of $D$ reported in HEDRICK (1985) for coat colors of cats

<table>
<thead>
<tr>
<th>Pair of loci (<em>A</em> listed first)</th>
<th>Locality</th>
<th>$p_A$</th>
<th>$p_B$</th>
<th>$D$</th>
<th>$\varepsilon_{\text{min}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$ and $d$</td>
<td>A</td>
<td>0.762</td>
<td>0.255</td>
<td>0.040</td>
<td>NFS*</td>
</tr>
<tr>
<td>$a$ and $S$</td>
<td>A</td>
<td>0.321</td>
<td>0.255</td>
<td>-0.027</td>
<td>0.637</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>0.495</td>
<td>0.248</td>
<td>-0.058</td>
<td>0.937</td>
</tr>
<tr>
<td>$d$ and $A$</td>
<td>A</td>
<td>0.328</td>
<td>0.765</td>
<td>0.036</td>
<td>NFS</td>
</tr>
<tr>
<td>$l$ and $S$</td>
<td>M</td>
<td>0.296</td>
<td>0.578</td>
<td>0.032</td>
<td>0.419</td>
</tr>
<tr>
<td>$l$ and $S$</td>
<td>P</td>
<td>0.495</td>
<td>0.856</td>
<td>0.070</td>
<td>NFS</td>
</tr>
<tr>
<td>$S$ and $t$</td>
<td>P</td>
<td>0.365</td>
<td>0.410</td>
<td>-0.072</td>
<td>0.965</td>
</tr>
</tbody>
</table>

Definition 2 was used for total epistasis, so $\varepsilon_{\text{min}}$ represents the largest value of $e_i$. Fitnesses were constrained to all be less than one, and a value of 0.5 was used for $r$ in all cases. The loci are $a$ (agouti or nonagouti), $d$ (nondilute or dilute), $l$ (short or long hair), $S$ (spotted or notspotted) and $t$ (striped or blotched tabby). The $p_A$ and $p_B$ allele frequencies listed are for the dominant allele. The localities are Amsterdam (A), Montreal (M) and Portsmouth (P). [See HEDRICK (1985) for more details.]

* NFS, no feasible solution, indicates that there is no solution in which all fitnesses lie between 0 and 1 that can produce the observed disequilibrium.

For allele frequencies which differ from one-half, the conclusion is that disequilibrium is much harder to generate for allele frequencies which deviate from one-half. (This conclusion holds if disequilibrium is replaced by correlation coefficient or a normalized disequilibrium, such as the disequilibrium divided by the maximum disequilibrium for the given allele frequencies.)

To understand more about the role of epistasis in generating disequilibrium in natural populations, I computed the epistasis necessary (see Tables 3 and 4) to maintain observed levels of disequilibrium which differed significantly from zero as reported in two studies—one of Drosophila montana by ROBERTS and BAKER (1973) and BAKER (1975) and one of cats by HEDRICK (1985). Even in the case of Drosophila montana, fairly large levels of epistasis are required. A comparison with the results of HASTINGS (1981), which gave levels of selection required in this case, shows that it really is the level of epistasis required that determines the level of selection necessary.

For the cats studied by HEDRICK (1985), the levels of epistasis required are
extremely, and probably unrealistically, large. It is unlikely that these levels of disequilibrium are maintained by epistatic selection leading to an equilibrium. Also, the values for epistasis obtained are almost identical to those reported in Table 4, if the constraint that all fitnesses lie between 0 and 1 is replaced by another realistic constraint; namely, requiring all fitnesses to lie between 0 and 2. This is in agreement with the general discussion above.

DISCUSSION

I shall begin by considering further the relationship between the results reported here and the extensive analyses of the symmetric model. Since the symmetric model gives the largest disequilibrium for the smallest epistasis of any fitness pattern for allele frequencies of one-half at each locus, the use of the symmetric model to provide limits to the relationship between epistasis and disequilibrium is further justified.

However, it is every important to note that as allele frequencies vary from one-half at each locus in a two-locus model, a very different picture emerges. No longer is the symmetric model the model that gives the most disequilibrium for a given level of epistasis. However, the level of disequilibrium—and, more importantly, the correlation coefficient—between alleles at two loci is much lower for a given level of epistasis of allele frequencies which differ from one-half than it is for equal allele frequencies. Thus, the intuition gained from a study of the symmetric model may overestimate the level of disequilibrium expected for a given level of epistasis.

The view that disequilibrium requires very large levels of epistasis is corroboration by the estimates of the epistasis required to produce the levels of disequilibrium observed in Drosophila montana. Such high levels of epistasis are understandable in this case because the two loci are both esterase loci, and the frequencies of null alleles are used. The extremely large levels of epistasis required to maintain the observed levels of disequilibrium observed in cats by Hedrick (1985) provide still further evidence that large values for disequilibrium are unlikely.

This paper thus provides another view as to why large-scale surveys of randomly chosen pairs of loci have failed to turn up large levels of linkage disequilibrium (reviewed in Hedrick, Jain and Holden 1978). In Hastings (1981), I suggested that one reason for this outcome was the strong selection required to maintain disequilibrium. Here, I have shown that not only is strong selection required to maintain significant levels of linkage disequilibrium, but strong epistasis is required as well. This corroborates and extends the implications of my earlier analysis of systems with weak epistasis (Hastings 1985a). Thus, it is not at all surprising that one notable case discussed above, in which linkage disequilibrium was found among outcrossers, involved closely linked loci that were likely to have strong epistatic interactions. Moreover, the allele frequencies at these loci were close to one-half. Finally, these loci were closely linked. I would suggest that further investigations of similar pairs of loci might turn up other cases of linkage disequilibrium in outcrossing species.
I thank P. Hedrick for sending me his paper before publication. I also thank P. Hedrick and the referees for their helpful comments. Supported by PHS grant RO1 GM32130.

LITERATURE CITED


