Heredity (1984), **52** (3), 373–382

A MODEL FOR THE INCORPORATION OF EPISTASIS INTO A COMPUTER SIMULATION FOR THREE EXPERIMENTAL DESIGNS

M. J. KEARSEY AND S. L. STURLEY

Department of Genetics, University of Birmingham, Birmingham B15 2TT, England

Received 27.x.83

SUMMARY

Using a sigmoid relationship between gene dosage and phenotype, a computer model is presented that accurately simulates the effects of epistasis for quantitative traits in three experimental designs; the basic generations (*i.e.*, parents, F_1's, F_2's and backcrosses), inbred families produced by single seed descent, and the triple test cross. It is shown that the classical expectations for components of generation means and variances are fulfilled when the genetical control is additive or interactive. Furthermore departures from the classical situation found in practice were also exhibited by our model. It seems likely, therefore, that in future studies, this inherently more flexible model for predicting the effect of epistasis may replace other methods of simulating epistasis.

1. INTRODUCTION

Predicting the consequences of various breeding strategies for a quantitative trait is relatively straightforward, providing the genetic effects are largely additive. Unfortunately, such ideal situations rarely exist in nature and departures from additivity, due to the independent or joint action of epistasis, genotype-environment interaction and linkage disequilibrium, are common.

It is possible in principle to remove all but the last effect by rescaling the data, although in practice it is often the case that removing the effect of, say, genotype-environmental interaction by such means, increases the effect of epistasis or vice-versa. So called macro-environmental genotype-environmental interaction can be avoided by raising all the material in the same environment. The effect of linkage disequilibrium is frequently negligible and the biases are such that they affect observed and predicted distributions to a similar extent (Jinks and Pooni, 1976). Epistasis can be incorporated into our genetical models but such models rapidly become complex as the number of loci and hence the numbers and types of interactions increase. Nonetheless, attempts have been made to investigate the consequences of various types of epistasis on estimates of genetical components and the predictions made from them. They have depended on ascribing individual numerical values to every possible digenic combination and summing in accordance with the relevant genetic algebra (Pooni and Jinks, 1979).

The purpose of the present paper is to explore a different model of epistasis, which we believe is easier to apply in simulation studies and which more closely mimics the underlying nature of gene action and interaction. In this model we assume a sigmoid relationship between gene dosage and phenotype. Thus a primarily additive gene action is designated by the

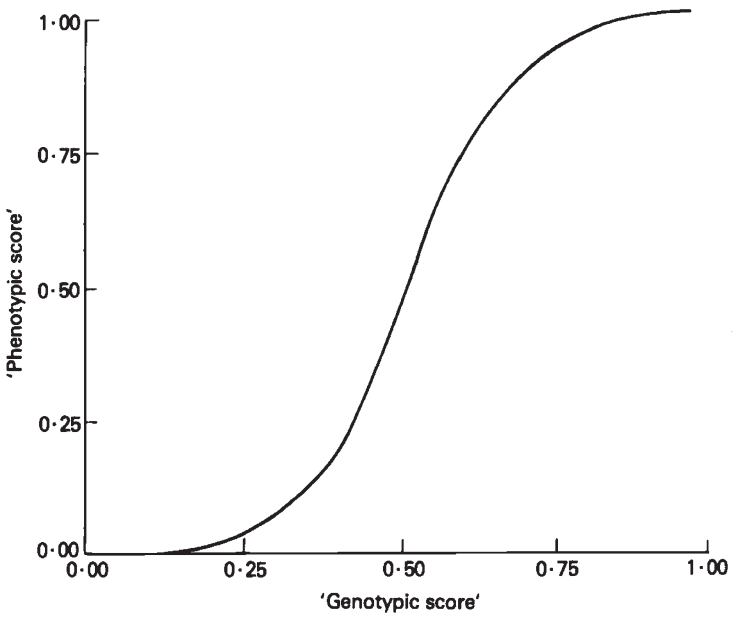

FIG. 1. The sigmoid function used to simulate epistasis.

middle linear portion of the relationship whereas interactions of varying complexity are given by the high and low non-linear extremes of the function (see fig. 1).

The efficacy of such a model can be tested in several ways. It must of course supply data that show the expected relationships between components of first and second degree statistics (Mather and Jinks, 1982). In previous work on *Nicotiana rustica* for example, anomalies have occurred, principally concerning the sign or magnitude of certain components of means and variances (Pooni and Jinks, 1981; Pooni, 1976). These deviations from the classical duplicate and complementary situations probably result from an oversimplification inherent in a digenic model and a failure to take full account of the true degree of gene dispersion involved. Thus any model that purports to mimic more closely the situation in nature should also produce these anomalies.

With this in mind, three experimental designs have been simulated. Firstly inbred lines have been produced by single seed descent from an F_2. In the presence of epistasis, the distribution of these is expected to show both skewness and kurtosis, the sign and magnitude of which depends on the nature and degree of the interaction present (Pooni, Jinks and Cornish, 1977). Secondly, the Triple test cross (Kearsey and Jinks, 1968) has been generated, again from predetermined parents. This yields three estimates of D, the additive genetic variance, one from the standard orthogonal comparison, ($L_1+L_2+L_3$, Jinks and Perkins, 1970) a second from the North Carolina Model III comparison, (L_1+L_2, Comstock and Robinson, 1952), while the analysis of variance of the families resulting from the crosses between the F_2 male testers and the F_1 females (L_3 families, Pooni and Jinks, 1979) provides the third estimate. Because these estimates each have

different epistatic biases, their relative magnitudes can be used to indicate the type of epistasis involved (Pooni and Jinks, 1979) and our simulation should reflect this.

The final experimental design used involves the so called basic generations *i.e.*, the inbred parents, F_1's, F_2's and back-crosses. Estimates of the components of first degree statistics are obtained by weighted least squares analysis and can be used to determine the genetical control of the character (Mather and Jinks, 1982). In particular, classical digenic interactions can be defined according to the sign of $[h]$, the dominance component, relative to $[l]$, the heterozygous × heterozygous interaction component. Deviations from these classical situations have occurred in certain association crosses of *N. rustica*. The relative magnitudes of all the first degree components follow a pattern atypical of both classical duplicate and complementary epistasis. It appears that it is the sign of $[l]$ that is aberrant, this being negative indicating duplicate interactions when all other components are positive as would be the case for complementary epistasis (Pooni and Jinks, 1982). Furthermore, it has long been realised that limitations in this type of analysis have probably been responsible for the possible spuriously high incidence of duplicate epistasis (Jinks and Jones, 1958, Jinks, Perkins and Pooni, 1973). Estimates of the components of second degree statistics of these generations are highly correlated with each other and biased by epistasis to the extent that they are considered useless for most practical applications. However, estimates of D are found to follow a pattern which depends on the degree of gene dispersion shown by the original parents. For a character known to display duplicate epistasis, association crosses give a significantly smaller estimate of D than dispersion crosses (Pooni, 1976). Our model should also show this previously unpredicted result.

2. The computer simulation

A computer program was used which simulates a quantitative trait controlled by 16 loci of equal effect. The program allows progeny to be produced by combining gametes generated from defined parents by a "random walk" procedure. This allows varying degrees of linkage to be accommodated between adjacent loci, although in the present study all loci were unlinked.

Scoring the progeny was a three stage process. Firstly, for each individual progeny the numbers of homozygous increasing (n_1), homozygous decreasing (n_2) and heterozygous loci ($n_3 = 16 - n_1 - n_2$) were determined. Using the notation of Mather and Jinks (1983), in which d is the additive and h the dominance deviation at any locus, the genotypic value of a given progeny is then

$$g_i = m + (n_1 - n_2)d + n_3 h.$$

Since all loci have the same effect and there is no epistasis at this level, the dominance ratio (b) is given by $b = h/d$ (or $h = bd$).

Thus the genotypic value can now be rewritten as

$$g_i = m + d(n_1 - n_2 + bn_3).$$

By specifying values for m, d and b, the genotypic score can be obtained for every zygote as it is produced.

The second stage is to rescale this score relative to the extreme scores that might be obtained. Since we have confined our attention to the situation in which there is no overdominance (*i.e.*, $-1 \leqq b \leqq 1$), the extreme genotypes will be those which are homozygous for all the increasing and decreasing alleles respectively and hence we have the genotypic values

$$g_{max} = m + 16d$$

$$g_{min} = m - 16d.$$

The position of a particular genotype within this range can then be determined as

$$g_i^* = \frac{g_i - g_{min}}{g_{max} - g_{min}}$$

and hence $0 \leqq g_i \leqq 1$.

In accomplishing this scoring and rescaling no allowance for the possibility of epistasis has been made. The third step involves the introduction of epistasis through a non-linear relationship between gene dosage and phenotype. Here we have used a sigmoid function, since this generates complementary interactions at low dosage, additive (non epistatic) effects at intermediate dosage and duplicate interactions at high gene dosage.

If X_i is a measure of gene dosage and Y_i the phenotype, we have used the function

$$Y_i = [0{\cdot}5(1 - \sin(X_i))]^2 - 0{\cdot}25$$

where X_i is expressed in radians (see fig. 1).

In any given simulation we can allow the range of genotypes to extend over any part of the range of X. For instance g_{min} could correspond to $X = 0$ and g_{max} to $X = 1$; this would simulate a situation in which the 16 genes segregating represent all the genes controlling the character. Alternatively the range could be restricted such that g_{min} and g_{max} correspond to intermediate values of X, thus yielding X_{min} and X_{max} and hence, by computation, Y_{min} and Y_{max}. Clearly any progeny derived from parents with rescaled score g_i^* can now be assigned an X value within this "epistatic range" and hence Y_i, its rescaled "phenotypic value" now including the effect of epistasis, can be determined.

Clearly simulations using different parts of the range will produce different ranges of Y. In order to produce some comparability between different simulations, the Y's have been linearly rescaled so that they always range between 40 and 140. The shape of the distribution of Y's within these limits will vary with the range of X's chosen for study. Finally, an environmental deviation is added by drawing a number at random from a normal distribution with mean zero and variance E_1.

3. The experimental design

Essentially three different genetical situations have been studied, namely complementary epistasis (C), no epistasis (A) and duplicate epistasis (D). These were produced by restricting attention to a narrow range of abscissa values (X) at the lower, middle and upper end respectively of the sigmoid function (see table 1 and fig. 1). In every case the dominance ratio was set at 0·5, as was E_1.

TABLE 1

Abscissa values used in the sigmoid function to produce 3 types of genetical control

Model	X_{min}	X_{max}
Additive	0·495	0·505
Duplicate Epistasis	0·80	1·00
Complementary Epistasis	0·00	0·15

For each of these 3 situations, two extreme pairs of parents were chosen which differed in both cases at all 16 loci. The one pair of parents was in association, the other in dispersion.

Using all 6 combinations of genetical controls (C, A, D with parents in association and dispersion) the following material was simulated:

(a) The basic generations of two inbred parents and their F_1, F_2 and first backcrosses, including reciprocals. These were produced as if they had been raised in a randomised plot design involving 4 plots per generation.
(b) Inbred families, 1050 of size 50, generated by single seed descent. (Clearly in the absence of linkage the degree of dispersion of the two original parents is irrelevant here.)
(c) The triple test cross. Here the experiment was designed to yield estimates of parameters with standard errors of around 10 per cent. Thus 260 males from the F_2 were crossed to the usual testers (P_1, P_2, F_1 and RF_1) with family sizes of 100, considerably in excess of the numbers required to detect epistatic variation (Pooni and Jinks, 1976).

4. THE CLASSICAL SITUATION

(i) *Basic generations, F_1, F_2 and backcrosses*

Using the joint scaling test (Cavalli, 1952) weighted least squares estimates of the components of means of these generations were obtained (table 2). In the absence of epistasis (situation A) but with association, a model involving m, $[d]$ and $[h]$ was found to be adequate, while in the dispersed cross, although m and $[h]$ maintained their values, $[d]$ was not significant due to the effect of dispersion (*i.e.*, $r_d \simeq 0$, Mather and Jinks, 1982). Estimates of D from the second degree statistics of both crosses were approximately equivalent and very close to an expected value (156·24) derived from the estimate of $[d]$. In this case, the potence ratio ($[h]/[d]$) is a true dominance ratio and yields an estimate (0·5034) which is not significantly different from the given value (0·5).

If we now consider complementary epistasis (C), the dispersion cross requires an m, $[h]$, $[i]$ and $[l]$ model. Both $[h]$ and $[l]$ were positive, a pattern typical of classical complementary epistasis. Due to the effect of r_i, $[i]$ was negative while $[d]$ and $[j]$ were zero because r_d and r_j are close to zero. In the dispersion cross from situation D (*i.e.*, duplicate epistasis) an m, $[h]$, $[i]$ and $[l]$ model again fitted, although this time $[l]$ was of the opposite sign (negative) to $[h]$, a characteristic of classical duplicate epistasis.

TABLE 2

Estimates of components of means and variances from the basic generations arising from dispersed and associated crosses with 3 genetical models

Genetical model		m	$[d]$	$[h]$	$[i]$	$[j]$	$[l]$	D
C.	Dispersed	50·4168	—	7·5059	−3·9728	—	14·2473	164·5716
	Associated	35·9802	50·0072	50·5022	54·0106	13·3228	−14·3051	224·9131
A.	Dispersed	89·9870	—	25·1067	—	—	—	151·8937
	Associated	89·9722	49·9981	25·1629	—	—	—	161·3351
D.	Dispersed	132·0776	—	12·7853	1·3416	—	−5·2812	20·3974
	Associated	114·5483	49·9914	65·1273	−24·5471	−63·9848	−40·0791	−183·8476

In the association cross from the same situation (D), [*d*] and [*j*] now appear in the model. Here [*i*], [*j*] and [*l*] were all negative, [*h*] being positive again as would be expected for classical duplicate epistasis.

(ii) *Inbred families*

The additive model (situation A) produced a sample of inbred families which were normally distributed, exhibited no skewness and only slight, possibly spurious kurtosis (table 3). The mean of all the inbreds ($\bar{F}_\infty$) gives a direct estimate of *m* and this was in accord with the estimate from the basic generations. Likewise the variance between true inbred family means is an estimate of *D* and this too was a close approximation to that obtained from the selfing backcrossing series.

TABLE 3

Analysis of inbred families for 3 different genetical models (C, A and D)

Genetical Model	D	E	Coefficient of skewness	Coefficient of kurtosis	Inbred Mean (F_∞)
Complementary epistasis (C)	71·1091	0·4955	1·9749*	8·9951**	49·1939
No epistasis (A)	148·8082	0·5012	$0·0219^{NS}$	2·7998*	89·7951
Duplicate epistasis (D)	65·0431	0·4993	−2·1913	11·6789	131·3147

** $0 \leqq 0·01$, * $0·01 \leqq P \leqq 0·05$, NS $P > 0·05$.

The inclusion of complementary epistasis (C) produced a sample of inbreds which showed the expected positive skewness and kurtosis. In the presence of epistasis, the inbred mean and between family variance yield estimates of $m+[i]$ and $D+I$ respectively (where *I* is the variance due to *i* type interactions). For the former, comparison with the estimate obtained from the basic generations of a dispersion cross was favourable. This did not apply for the latter, probably because no estimate of *I* was available.

A sample of inbreds from situation D revealed the considerable negative skewness and positive kurtosis predicted in the presence of classical duplicate epistasis. Again the $\bar{F}_\infty$ estimate of $m+[i]$ is only comparable to those estimates from a dispersion cross of the basic generations.

(iii) *Triple test cross*

Triple test crosses from situation A with parents in association and dispersion gave no evidence of epistasis. Estimates of *D* from the 3 comparisons mentioned previously for both crosses were all homogeneous (table 4). These values were also very similar to the estimates obtained from both the previous experimental designs. Similarly, all estimates of the dominance ratio were close to the expected value of 0·5.

In the presence of either type of non allelic interaction for both dispersion and association crosses, the orthogonal comparison testing for epistasis was highly significant. If we consider only those triple test crosses in which the parents were in association, it is clear that the magnitude of the L_3 estimate

TABLE 4

Estimates of the additive genetic variance from simulated triple test cross experiments on material showing different genetical controls (C, A and D)

Genetical model		$D(L_1+L_2+L_3+L_4)$	$D(L_1+L_2)$	$D(L_3)$
C.	Dispersed	146·8278	148·8646	144·5571
	Associated	145·8834	120·3326	173·9411
A.	Dispersed	146·6990	144·3102	149·1634
	Associated	157·3554	155·9926	158·2599
D.	Dispersed	9·7949	8·6976	10·9558
	Associated	46·8664	102·8926	12·5903

of D relative to the others does indicate the type of epistasis involved. Thus for situation C, the L_3 estimate was the largest of the three, indicating complementary epistasis. Similarly for situation D, the L_3 estimate was the smallest of the three indicating duplicate epistasis.

5. ANOMALIES

If a digenic model for these interactions is correct then estimates of m, $[h]$ and $[l]$ from the basic generations should be the same regardless of the degree of gene dispersion involved. This was clearly not the case for situations C and D, probably due to the effect of undetected higher order interactions, *i.e.*, these estimates were biased by components of multigenic epistasis. In a dispersion cross, however, since these components are functions of the coefficient of dispersion, they had no effect. Thus estimates of m, $[h]$ and $[l]$ from these crosses, whatever the type of epistasis involved, can be considered reliable. Since the degree of gene dispersion is irrelevant in the inbred lines experiments, it is not surprising that estimates of m from this design were not significantly different from the values of $m+[i]$ from the basic generations arising from dispersion crosses.

Owing to the considerable multigenic component involved, other anomalies concerning first degree statistics from association crosses have occurred. In situation C, a six parameter model was fitted to the data of the basic generations. For an association cross showing complementary interactions, all components would be expected to be of the same sign. This was not the case since $[l]$ alone was negative, a pattern which does not fit the classical interpretations. In *Nicotiana rustica*, association crosses of material known to exhibit complementary epistasis also gave similarly aberrant values for $[l]$ (Pooni and Jinks, 1981).

Estimates of components of second degree statistics from the basic generations are expected to be biased by epistasis (digenic or otherwise). Since they are so highly correlated with each other, these biases will be even more serious, to the extent that certain estimates of D are negative in our study! Pooni (1976) has indicated that for a trait showing duplicate epistasis the estimate of D from the basic generations arising from an association cross should be significantly smaller than the dispersion cross estimate. This was the case for situation D in our simulation.

In a triple test cross using dispersed parents, whatever the type of epistasis involved, estimates of D from the three comparisons were

homogeneous. Thus they could not be used as indicators of the nature of interaction involved. It is clear that the estimates of D from the triple test cross and basic generations did not compare favourably with the "true" value of $D+I$ obtained from the variance between inbred families. In the present study, however, the most reliable estimates do come from the triple test cross.

6. Conclusions

It is clear that in the absence of epistasis (A) our model simulates the natural situation in all aspects. Upon the inclusion of epistasis using the sigmoid function (C and D) in the majority of cases, the classical expectations are fulfilled. There is evidence in this study for considerable multigenic interactions. This, together with the occurrence of anomalies previously encountered in *N. rustica*, would seem to indicate that our sigmoid model is a valid one when predicting the effects of epistasis on the multiple mating designs cited.

Model fitting using the algebraic method has extended beyond the digenic situation. Jinks and Perkins (1969) modelled and detected trigenic interactions and generalised multilocus models have also been proposed (Jinks, 1979). However, the use of a computer simulation utilising our function is an inherently more flexible system to predict the effects of epistasis. It has been shown (Sturley, 1982) that by adjusting the "epistatic range", data specifying a purely digenic interactive model can be obtained. Here estimates of m, $[h]$ and $[l]$ from association and dispersion crosses, were equivalent to a situation that would be expected in the presence of purely digenic interactions.

The facility also exists to introduce linkage and genotype-environmental interactions into our model. Hence, it is clear that all types and combinations of genetical control of a character can be simulated. In this way the efficiency of various experimental designs in detecting these factors can be determined and ultimately the potential of a breeding program could be ascertained. For example the nature and degree of interaction present has been shown to have little practical effect when predicting the properties of inbred lines derived by single seed descent, despite the biases introduced into the prediction method (Sturley, 1982). This type of simulation could also be used to optimise the size and structure of experiments to detect the genetical control of a character. This could provide a considerable saving both in manpower and experimental space.

Acknowledgements. We would like to express our gratitude to Professor J. L. Jinks for helful discussions and the provision of data. One of us (S.L.S.) also wishes to acknowledge receipt of an SERC Advanced Course Studentship which enabled this work to be carried out. The computing facilities were made available by the University of Birmingham Computer Centre.

7. References

CAVALLI, L. L. 1952. An analysis of linkage in quantitative inheritance. In Reeve, E. C. R. and Waddington, C. H. (eds.) *Quantitative Inheritance*, H.M.S.O., London, pp. 135–144.

COMSTOCK, R. E. AND ROBINSON, H. F. 1952. Estimation of average dominance of genes. In Gowen, J. W. (ed.) *Heterosis*, Iowa State College Press, Ames, Iowa, pp. 494–516.

JINKS, J. L. 1979. The biometrical approach to quantitative variation. In Thompson, J. N. and Thoday, J. M. *Quantitative Genetic Variation*, Academic Press, London, 81–109.

JINKS, J. L. AND JONES, R. M. 1958. Estimation of the components of heterosis. *Genetics, 43*, 223–234.

JINKS, J. L. AND PERKINS, J. M. 1969. The detection of linked epistatic genes for a metrical trait. *Heredity, 24*, 465–475.

JINKS, J. L. AND PERKINS, J. M. 1970. A general method for the detection of additive, dominance and epistatic components of variation: III. F_2 and backcross populations. *Heredity, 25*, 419–429.

JINKS, J. L., PERKINS, J. M. AND POONI, H. S. 1973. The incidence of epistasis in normal and extreme environments. *Heredity, 31*, (2) 263–269.

JINKS, J. L. AND POONI, H. S. 1976. Predicting the properties of recombinant inbred lines derived by single seed descent. *Heredity, 36*, 253–266.

KEARSEY, M. J. AND JINKS, J. L. 1968. A general method of detecting additive, dominance and epistatic variation for metrical traits. I. Theory. *Heredity, 23*, 403–409.

MATHER, K. M. AND JINKS, J. L. 1982. *Biometrical Genetics*, 2nd Edn, Chapman and Hall.

POONI, H. S. 1976. Exploitation of Inbreds in *Nicotiana rustica*. Ph.D. Thesis, Department of Genetics, University of Birmingham.

POONI, H. S. AND JINKS, J. L. 1976. The efficiency and optimal size of triple test cross designs for detecting epistatic variation. *Heredity, 36*, 215–227.

POONI, H. S. AND JINKS, J. L. 1979. Sources and biases of the predictors of the properties of recombinant inbreds produced by single seed descent. *Heredity, 42*, 41–48.

POONI, H. S. AND JINKS, J. L. 1981. The true nature of non-allelic interaction in *Nicotiana rustica* revealed by association crosses. *Heredity, 47*, 253–258.

POONI, H. S., JINKS, J. L. AND CORNISH, M. A., 1977. The causes and consequences of non-normality in predicting properties of recombinant inbred lines. *Heredity, 38*, 329–338.

STURLEY, S. L. 1982. The consequences of epistasis in predicting the properties of recombinant inbred lines derived by single seed descent, as revealed by a computer simulation. M.Sc. Thesis, Department of Genetics, University of Birmingham.