Abstract

Estimation of linkage between pairs of gene loci is an important problem in genetics. The strength of linkage between pairs of gene loci is measured with the recombination value \( r \) and is called the genetic map length. The recombination value is the probability that a gamete of an individual has resulted from a crossing over between loci. When each of two gene loci has two genes (A,a and B,b), four gametes are possible (AB, Ab, aB, ab). The calculated number of individuals in each genetic class from the phenotypes of selfed double heterozygotes may be used to estimate the recombination value. Computer programs are available that give maximum likelihood estimates of \( r \). However, homogeneity tests of \( r \) values are not easily performed, nor are these programs conducive to processing large data sets. This paper demonstrates how the SAS™ CATMOD procedure may be used to estimate and test hypotheses about \( r \) values between pairs of gene loci when the phenotypes of offspring from selfed heterozygotes are available. We concluded that in certain situations, the CATMOD procedure is a more effective way of estimating and testing hypotheses about gene linkage than are currently available computer programs.

Introduction

An important problem in genetics is the study of linkage between pairs of gene loci. Linkage may be used to identify important gene markers and to aid geneticists with identifying more productive plant and animal genotypes. In addition, linkage studies may be used to establish the relative distance between two gene loci on a chromosome and thus are useful in chromosome mapping.

The strength of linkage between two gene loci is measured with the recombination value \( r \). The recombination value is the probability that the gamete of an individual has resulted from a crossing over between loci. For example, consider two loci A and B with genes A, a and B, b. A double heterozygote AB/ab that received gametes AB and ab from its two parents is expected to produce four types of gametes in the following proportions:

\[
\begin{array}{cccc}
AB & Ab & aB & ab \\
\frac{1-r}{2} & \frac{r}{2} & \frac{r}{2} & \frac{1-r}{2}
\end{array}
\]

When \( r = 1/2 \), the two loci are not linked and all four types are transmitted in equal proportions. If they are perfectly linked, then \( r = 0 \) and only gametes of types AB and ab are produced.

Recombination values can be estimated in several different ways but are most commonly estimated using maximum likelihood methods (Fisher and Balmukand, 1928; Allard 1956; Weir 1990). Computer programs are available that estimate \( r \) based on the type of linkage experiment (Suiter et al, 1983); however, these programs are not conducive to (1) processing large data sets, or (2) testing more complicated hypotheses regarding recombination values. It would be quite useful to geneticists to estimate and test a wide variety of hypotheses about recombination values using a single computer run. The purpose of this paper is (1) to demonstrate how the SAS™ CATMOD procedure may be used to estimate and test hypotheses about recombination rates between pairs of loci from offspring of selfed double heterozygotes, and (2) to compare estimates and tests from the CATMOD procedure with estimates and tests based on maximum likelihood.

Estimation

Estimating and testing hypotheses of recombination values between pairs of gene loci from offspring of selfed double heterozygotes in repulsion phase (Ab/aB) may be estimated from the...
observed numbers of individuals in each of four classes and their expectations (Fisher and Balmukand, 1928):

<table>
<thead>
<tr>
<th>Phenotypic Class</th>
<th>AB</th>
<th>Ab</th>
<th>aB</th>
<th>ab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>a</td>
<td>b</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>Expectation</td>
<td>(\frac{1}{4}(2+r^2))</td>
<td>(\frac{1}{8}(1-r^2))</td>
<td>(\frac{1}{8}(1-r^2))</td>
<td>(\frac{1}{4}r^2)</td>
</tr>
</tbody>
</table>

Coupling phase (AB/ab) expectations may be found by replacing \(r\) in the above expectations with \(r-1\).

**Maximum likelihood:**

The likelihood \(L(r)\) is proportional to: \((2+r^2)(1-r^2)^2(r^2)^d\). The derivative of the log likelihood is:

\[
\frac{d\ln L(r)}{dr} = a \cdot \frac{2r}{(2+r^2)} + (b+c) - \frac{2r}{(1-r^2)} + \frac{d^2}{r}
\]

This derivative is set equal to zero and solved for the maximum likelihood estimate \(\hat{r}\). The variance of \(\hat{r}\) is (Allard, 1956; Weir, 1990):

\[
\text{Var}(r) = \frac{(2+r^2)(1-r^2)}{2n(1+2r^2)}
\]

Linkage I (Suiter, et al. 1983) was used to compute the maximum likelihood estimates of \(r\).

**Moment estimator based on the cross product ratio:**

(Fisher and Balmukand 1928, Weir 1990). Equating the cross product ratio \(\alpha\) with each term replaced with its expected value gives the following equation:

\[
\alpha = \frac{a \cdot d}{b \cdot c} = \frac{r^2(2+r^2)}{(1-r^2)^2}
\]

This equation results in the following formula for \(r^2:\)

\[
r^2 = \frac{-(1+\alpha) + \sqrt{1+3\alpha}}{1-\alpha}
\]

Note that \(r^2\) is a function of \(\alpha\) which is a function of the counts. Thus it is possible to use Grizzle, Starmer and Koch’s (1969) generalized least squares (GLS) approach in the CATMOD procedure to estimate \(r\). However, in this approach, the variance of \(r\) is only approximate since \(r\) is a nonlinear function of the observed counts.

**Testing hypotheses**

If \(r\) is estimated for more than one group of data, it may be important to test hypotheses about these values.

**Maximum Likelihood:** Following Allard (1956), when \(r\) is homogeneous for \(N\) sets of data, deviations from zero of the logarithm of maximum likelihoods for each group of data multiplied by the variance is distributed chi-square with \(N-1\) degrees of freedom. That is:

\[
X^2 = \sum_{i=1}^{N} D_i(r)^2 V_i(r) - \chi^2(N-1)
\]

where

\[
D_i(r) = \frac{d\ln L(r)}{dr} \text{ for the } i^{th} \text{ data set}
\]

\[
V_i(r) = \text{ is the variance of } r \text{ for the } i^{th} \text{ data set as defined above.}
\]

**Moment method.** Various hypotheses regarding \(r\) may be tested using the Grizzle-Starmer-Koch generalized least squares (GLS) approach in the CATMOD procedure. Since \(r^2\) is a function of the observed counts, the RESPONSE statement in the CATMOD procedure may be written to specify \(r\). When repulsive phase (Ab/aB) heterozygotes are selfed, the following RESPONSE statement should be used:

\[
\text{EXP} * \text{ .5 LOG} \\
\text{EXP} * \text{ 1-1 LOG} -1 \text{ 0 0 0 1} \\
\text{EXP} 1 \text{ 0 0 0 0 5 0 0 0 0 1 LOG} + 1, 1, 1 * 1 \text{ 0 0 0 0 3 0}, \\
0 0 1 * 1, 1, 1 \\
\text{EXP} * 1 \text{ -1-1 LOG ;}
\]

When coupling phase (AB/ab) heterozygotes are selfed, the following RESPONSE statement should be used:

\[
+ 1 * -1 \text{ EXP} * \text{ .5 LOG} \\
\text{EXP} * 1 \text{ -1 LOG} -1 \text{ 1 0 0 0} \\
\text{EXP} 1 \text{ 0 0 0 0 5 0 0 0 0 1 LOG} * 1 \text{ 0 0 0 0 1 0 0 0 0 3 0, 0 0 1 -1 * 1, 1, 1} \\
\text{EXP} * 1 \text{ -1-1 LOG ;}
\]

**Results**

To evaluate the usefulness of the GLS method in analyzing recombination values, estimates and tests based on both maximum likelihood and the GLS method were compared using several different data sets.

The first data set was used to estimate linkage in maize between a sugary factor and a base leaf color factor (Fisher and Balmukand, 1928). Repulsion phase (Ab/aB) heterozygotes, which were selfed, produced 3839 seedlings that were cross-classified into the following classes:

<table>
<thead>
<tr>
<th>AB</th>
<th>Ab</th>
<th>aB</th>
<th>ab</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>906</td>
<td>904</td>
<td>32</td>
</tr>
</tbody>
</table>
Estimates of r and standard errors are given in Table 1. Maximum likelihood (ML) and GLS method estimates were identical, but standard errors differ slightly due to the approximate nature of variances using the GLS method.

Table 1. Maximum likelihood (ML) and generalized least squares (GLS) estimates of recombination value (r) and their standard errors using data from Fisher and Balmukand (1928).

<table>
<thead>
<tr>
<th>r</th>
<th>STDERR</th>
<th>r</th>
<th>STDERR</th>
</tr>
</thead>
<tbody>
<tr>
<td>.1888</td>
<td>.0154</td>
<td>.1888</td>
<td>.0159</td>
</tr>
</tbody>
</table>

A second data set on lima beans was given by Allard (1956) where the genes involved dominant (A) vs. indeterminate growth habit (a) and dark red (B) vs. red seed coat (b). The first three experiments, whose parental heterozygotes were in the coupling phase (AB/ab), were used here. Estimates of r and chi-square ($\chi^2$) for testing the hypothesis of equal recombination across experiments were given in Table 2. ML and GLS estimates were nearly identical and test statistics differed only slightly.

Table 2. Maximum likelihood (ML) and generalized least squares (GLS) estimates of recombination values (r) and chi-square statistics using data from Allard (1956).

<table>
<thead>
<tr>
<th>Experiment</th>
<th>ML</th>
<th>GLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.3961</td>
<td>.3954</td>
</tr>
<tr>
<td>2</td>
<td>.4207</td>
<td>.4205</td>
</tr>
<tr>
<td>3</td>
<td>.3657</td>
<td>.3654</td>
</tr>
<tr>
<td>Chi-Square ($\chi^2$)</td>
<td>2.53</td>
<td>2.62</td>
</tr>
</tbody>
</table>

A final data set is given in a paper by Tulsieram et al. (1992). In this study, individuals from $F_2$ families from two maize populations were grown in two locations (greenhouse and field), and linkage between loci (Pgm1 and Adh1) was estimated to assess the effects of population and location on recombination rates. In each population/location combination, each individual was classified into one of the four phenotypic groups (AB, Ab, aB, ab). $F_2$ parental heterozygotes were in the coupling phase (AB/ab).

Table 3 gives ML and GLS estimates for each population/location. Table 4 gives chi-square statistics for tests of homogeneity of r values. As before, ML and GLS estimates are nearly identical and chi-square statistics differ somewhat. The GLS method using the CATMOD procedure has the advantage of being able to test population x environment interaction. Testing interaction was not possible with the ML method based on Suiter et al. (1983).

In all three studies, the GLS approach gave r estimates that were nearly identical to the maximum likelihood values. However, the GLS approach appeared to be slightly less efficient than the ML method. Based on these data, it would appear that this loss of power was only slight.

Table 3. Maximum likelihood (ML) and generalized least squares (GLS) estimates of recombination value (r) using data from Tulsieram et al (1992).

<table>
<thead>
<tr>
<th>Population/Location</th>
<th>r values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>Location</td>
</tr>
<tr>
<td>ML</td>
<td>GLS</td>
</tr>
<tr>
<td>1</td>
<td>.2755</td>
</tr>
<tr>
<td>1</td>
<td>.2036</td>
</tr>
<tr>
<td>2</td>
<td>.1917</td>
</tr>
<tr>
<td>2</td>
<td>.1755</td>
</tr>
</tbody>
</table>

Table 4. Chi-square statistics for homogeneity tests using maximum likelihood (ML) and generalized least squares (GLS).

<table>
<thead>
<tr>
<th>Homogeneity test</th>
<th>Chi-square values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population/Location</td>
<td>ML</td>
</tr>
<tr>
<td>Combination</td>
<td>15.52</td>
</tr>
<tr>
<td>Population</td>
<td>8.10</td>
</tr>
<tr>
<td>Location</td>
<td>9.81</td>
</tr>
<tr>
<td>Population x Location</td>
<td>1.40</td>
</tr>
</tbody>
</table>

Conclusions

Estimation and testing hypotheses about gene linkage is an important problem in genetics. In situations where the offspring of selfed double heterozygotes are available, the SAS™ CATMOD procedure can be used to estimate recombination values. In this situation, use of the CATMOD procedure has advantages over other computer programs (Suiter, et al, 1983) since it (1) may be
easily used to test a broad range of different hypotheses which are not easily tested in other programs and (2) it can be used to efficiently process large data sets containing many different loci and many different sets of individuals. However, the CATMOD procedure appears to be somewhat limited due to a slight loss of power relative to tests based on maximum likelihood.

Contact Author

Kent M. Eskridge
Department of Biometry
University of Nebraska-Lincoln
Lincoln, NE 68583-0712
(402) 472-2903

References


Appendix

1. Program for data set 1. Data are from Fisher and Balmukand, (1928).

```plaintext
OPTIONS LS=72;
DATA A; INPUT PHENO$ WT @@;
CARDS;
   A 1997 B 906 C 904 D 32
PROC PRINT;
PROC CATMOD; WEIGHT WT;
RESPONSE /* REPULSION PHASE RESPONSE STATEMENT */
   EXP * .5 LOG
   EXP * 1 -1 LOG -1 1 0, 0 0 1 EXP 1 0 0, 0 .5 0, 0 0 1 LOG
   + 1, 1, 1 * 1 0 0, 0 3 0 , 0 0 -1 * 1, 1, 1
   EXP * 1 -1 -1 1 LOG ;
MODEL PHENO = ;
```

2. Program for data set 2. Data are the first three data sets from Allard (1956).

```plaintext
OPTIONS LS=72;
DATA A; INPUT EXPT PHENO$ WT @@; CARDS;
   1 A 200 1 B 57 1 C 49 1 D 30
   2 A 842 2 B 234 2 C 255 2 D 126
   3 A 274 3 B 71 3 C 64 3 D 45
PROC PRINT;
PROC CATMOD; WEIGHT WT;
RESPONSE /* COUPLING PHASE RESPONSE STATEMENT */
   + 1 * -1
   EXP * .5 LOG
   EXP * 1 -1 LOG 1 -1 0, 0 0 1 EXP 1 0 0, 0 .5 0, 0 0 1 LOG
   * 1 0 0, 0 1 0, 0 0 -1 + 1, 1, 1 * 1 0 0, 0 3 0 , 0 0 -1 * 1, 1, 1
   EXP * 1 -1 -1 1 LOG ;
MODEL PHENO = EXPT ;
```

DATA B; SET A; R = 4058; *OVERALL ML ESTIMATE;
R = R -1; *USE R-1 FOR COUPLING PHASE - SEE ALLARD (1956);
SCORE = (PHENO = 'A')*WT*(2*R/(2+R**2)) +
   (PHENO = 'B')*WT*(-2*R/(1-R**2)) +
   (PHENO = 'C')*WT*(2*R/(1-R**2)) +
   (PHENO = 'D')*WT*(2/R);
INFO = WT*(2*(1+2*R**2))/((2+R**2)*(1-R**2));
DATA C; SET B ; BY EXPT; IF FIRST.EXPT THEN DO;
   SCORET = 0; INFOT = 0; END;
   SCORET + SCORE; INFOT + INFO;
DATA D; SET C; BY EXPT; IF LAST.EXPT; NN = 1;
   S2 = SCORET**2; RATIO = S2/INFOT; CHISQ + RATIO;
DATA E; SET D; BY NN; IF LAST.NN ;
PROC PRINT; VAR EXPT CHISQ;

1046

```sas
DATA B; INPUT POP LOC PHENO$ WT @@; TRT=10*POP+LOC; CARDS;
  1 1 A 351 1 B 60 1 C 92 1 D 95
  1 2 A 351 1 B 63 1 C 59 1 D 127
  2 1 A 232 2 B 29 2 C 48 2 D 88
  2 2 A 264 2 B 27 2 C 38 2 D 71
PROC CATMOD; WEIGHT WT;
RESPONSE  /* COUPLING PHASE RESPONSE STATEMENT */
  + 1 * -1
  EXP * .5 LOG
  EXP * -1 * LOG 1 -1 0, 0 0 1 EXP 1 0 0, 0 .5 0, 0 0 1 LOG
  + 1 0 1 0, 0 1 1 * 1 0 0, 0 3 0, 0 0 -1 * 1, 1, 1
  EXP * -1 * 1 1 LOG ;
MODEL PHENO = TRT; * USE FOR CHI-SQUARE FOR 4 GROUPS;
*MODEL PHENO = POP; * USE FOR CHI-SQUARE FOR POP ;
*MODEL PHENO = LOC; * USE FOR CHI-SQUARE FOR LOC ;
*MODEL PHENO = POP LOC POP*LOC; * USE FOR CHI-SQUARE FOR POP*LOC;

DATA C; SET B; R=.2162; /* OVERALL ML R ESTIMATE;
  */ USE OVERALL ML R ESTIMATE FOR CHI-SQUARE AMONG 4 GROUPS
  USE POP R'S FOR CHI-SQUARE FOR POPULATION EFFECTS
  USE LOC R'S FOR CHI-SQUARE FOR LOCATION EFFECTS */
  / IF POP=1 THEN R=.2377 ; IF POP=2 THEN R=.1840 ;
  IF LOC=1 THEN R=.2397 ; IF LOC=2 THEN R=.1931 ; /
R=R-1; *USE R-1 FOR COUPLING PHASE - SEE ALLARD (1956);
SCORE = (PHENO='A')*WT*(2*R/(2+R**2)) +
  (PHENO='B')*WT*(-2*R/(1-R**2)) +
  (PHENO='C')*WT*(-2*R/(1-R**2)) +
  (PHENO='D')*WT*(2/R) ;
INFO = WT*(2*(1+2*R**2))/(2*(1+R**2)*R**2);
DATA D; SET C ; BY TRT; IF LAST.TRT THEN DO;
  SCORET=0; INFOT=0; END;
SCORET+SCORE; INFOT+INFO;
DATA E;SET D;BY TRT; IF LAST.TRT; NN=1;
S2=SCORET**2; RATIO = S2/INFOT; CHISQ+RATIO;
DATA F; SET E ; BY NN; IF LAST.NN;
PROC PRINT; VAR TRT CHISQ;
```

1047