Estimating Interference and Linkage Map Distance from Two-Factor Tetrad Data

Franklin W. Stahl* and Russell Lande*

*Department of Biology and Institute of Molecular Biology, University of Oregon, Eugene, Oregon 97403

Manuscript received September 6, 1994
Accepted for publication December 8, 1994

ABSTRACT

We present methods for using the model of FOSS, LANDE, STAHL, and STEINBERG to estimate interference and map distances from two-factor tetrad data. We illustrate the application of the methods with data from Neurospora and from Saccharomyces.

FOSS et al. (1993) offered a biologically based model for chiasma interference. The model is simple, mathematically tractable and flexible and has passed stringent tests of adequacy. The model successfully describes the dependence of the four-point coefficient of coincidence ($S_4$) on the map distance separating two short intervals for both Drosophila and Neurospora (FOSS et al. 1993), and it describes the distribution of exchanges along the linkage map of the Drosophila X chromosome for tetrads of ranks I, II and III, respectively (LANDE and STAHL 1993). McPeeK and SPEED (1995) have demonstrated the superior ability of the model to account for the frequencies of all genotypes arising in a Drosophila cross involving nine linked markers. Models leading to the same equations as those of FOSS et al. were presented earlier (COBB 1978; STAHL 1979; STAM 1979) but were not widely embraced, perhaps because the formalizations were neither based on a clear biological mechanism nor subjected to rigorous tests of adequacy.

The complete mathematical treatments by McPeeK and SPEED (1995) and by ZHAO et al. (1995) subsume the analysis given here. Our aim is to provide a user-friendly treatment of a particular example often encountered in tetrad analysis. We offer it for several reasons: (1) The formalisms of our model provide a more coherent description of linkage data than do other models. (2) For some fungi, the only data available for estimating interference, and hence determining the best mapping function, may be two-factor tetrad data. (3) For Saccharomyces, gene conversion intrudes upon estimates of conventional tetrad types, even for moderately long intervals. Methods for estimating interference by tetrad analysis of two-factor crosses minimizes that confusion by using the fewest possible marked loci, and those loci are well separated so that conversions of the markers are rare relative to reciprocal recombination.

Tetrad analysis of two-factor crosses provides information on both the recombination frequency ($R$) for the marked interval and the intensity of interference from the frequencies ($PD$, $T$ and $NPD$) of Parental Ditype, Tetratype, and Nonparental Ditype tetrads, respectively. $R$ is directly calculated, without assumptions, as $R = NPD + (\frac{1}{2})T$.

Interference can be sensitively detected by tetrad analysis of two-factor crosses (PAPA zIAN 1952; STRickLAND 1958; PERKINS 1962a). However, interference (and map distance) in a two-factor cross can be quantified only with the aid of a model. In the model of BARRATT et al. (1954), interference is embodied in a parameter, $k$. Their model, like that of FOSS et al., assumes the absence of chromatid interference as well as of sister exchanges that can interfere with exchanges between homologues.

The model of BARRATT et al. imagines an a priori map length that can be calculated from the observed frequency of recombinant haploid products ($R$) using the relationship $R = (\frac{1}{2})(1 - e^{-2X})$, which relates $R$ to map length, $X$ (in Morgans), in the assumed absence of interference. From this estimate of $X$, the model calculates the frequency of non-exchange tetrads as $t_0 = e^{-2X}$, where $2X$ is the mean number of a priori interhomologue exchanges per tetrad and $t_0$ is the fraction of tetrads with no exchanges in the interval calculated from that mean according to the Poisson distribution.

The model of BARRATT et al. states that the value of $t_0$ is independent of the value of the interference parameter $k$; i.e., interference is presumed to alter only the frequencies of tetrads with exchanges. The new probability for tetrads with $r$ exchanges in the marked interval is obtained by multiplying the a priori (Poisson) frequency for that rank of tetrad by $k^{(r-1)}$. This procedure alters the frequencies of the a priori tetrads, and a normalization factor, applied to tetrads with one or more exchanges, restores the sum of frequencies of all tetrads to unity.

Although the interference parameter, $k$, has no identified biological basis, the model does allow one to relate frequencies of tetrad types with estimated map dis-
TABLE 1
Symbols employed

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>Recombination frequency</td>
</tr>
<tr>
<td>X</td>
<td>Linkage map distance (Morgans)</td>
</tr>
<tr>
<td>$S_4$</td>
<td>Four-point coefficient of coincidence</td>
</tr>
<tr>
<td>PD</td>
<td>Parental ditype tetrads</td>
</tr>
<tr>
<td>NPD</td>
<td>Nonparental ditype tetrads</td>
</tr>
<tr>
<td>T</td>
<td>Tetratype tetrads</td>
</tr>
<tr>
<td>PD</td>
<td>Frequency of PDs</td>
</tr>
<tr>
<td>NPD</td>
<td>Frequency of NPDs</td>
</tr>
<tr>
<td>$T$</td>
<td>Frequency of $T$s</td>
</tr>
<tr>
<td>$k$</td>
<td>Interference constant</td>
</tr>
<tr>
<td>$t_r$</td>
<td>Probability of tetrad of rank $r$</td>
</tr>
<tr>
<td>$m$</td>
<td>Interference constant</td>
</tr>
<tr>
<td>$C_s$</td>
<td>Conversion, visible or cryptic</td>
</tr>
<tr>
<td>$C_w$</td>
<td>$C$ with associated crossover</td>
</tr>
<tr>
<td>$C_o$</td>
<td>$C$ without associated crossover</td>
</tr>
<tr>
<td>$y$</td>
<td>Mean number of $C$s in an interval</td>
</tr>
<tr>
<td>$p_i$</td>
<td>Probability of $i$Cs</td>
</tr>
<tr>
<td>$p_j$</td>
<td>Probability of $j$Cs</td>
</tr>
</tbody>
</table>

stances. BARRATT et al. demonstrated that a given value of $k$ gave a reasonable fit both to tetrad type frequencies in two-factor crosses and to the relationship between $R$ and $X$ (the mapping function), compatible with the assumption of negligible chromatid interference. SNOW (1979) introduced the use of maximum likelihood methods for the calculation of $k$ from the frequencies of tetrad types.

The model of BARRATT et al. and the procedure of SNOW are widely used by fungal geneticists. KING and MORTIMER (1991) have offered an improved version of the model and maximum likelihood equations for its application. Both of these models, however, are count-location models, which violate the well-established rule, illustrated for Neurospora and Drosophila in FOSS and STAHL (1995) and for Saccharomyces in KING and MORTIMER (1990), that chiasma interference increases (coefficient of coincidence decreases) as the distance between the intervals examined goes to zero (McPeeK and SPEED 1995).

In the model of FOSS et al., as in that of BARRATT et al., the distribution of exchanges is arrived at by a modification of the Poisson distribution. For FOSS et al., the underlying Poisson distribution is the distribution of $C_s$, events that could lead to conversions were there markers to reveal them. $C_s$ are resolved as noncrossovers ($C_w$S) or as crossovers ($C_o$S). Interference is introduced by the assumption that a fixed number, $m$, of $C_o$S falls between neighboring $C_w$S. (Within the framework of the double chain break model for recombination [SZOSTAK et al. 1983], the assumption of a fixed number of $C_o$S is a rule governing the resolution of an intermediate that is common to $C_w$S and $C_o$S.) Thus, $m$ is the interference parameter in our model. In a simple world, $m$, as well as $k$, might be species- or chromosome-specific. Should $m$ or $k$ or any other interference parameter prove to be region- or locus-specific, the hope of establishing a theoretical relationship between $R$ and $X$ for an entire chromosome arm would be thwarted.

A $k$ value of 0.2 was found by BARRATT et al. to be suitable for predicting tetrad type frequencies for both Neurospora and Drosophila. Using different, more stringent criteria, FOSS et al. found that $m$ differs between Neurospora and Drosophila. In the first instance, $m$ is calculated from the observed fraction of $C_s$ that are $C_w$S. These $m$ values, 2 for Neurospora and 4 for Drosophila, were then seen to give optimal fits of the model to the observed dependence of the coefficient of coincidence on map distance between intervals. The value of $m = 4$ for Drosophila was further substantiated by the ability of the model so evaluated to describe the distribution of exchanges along $X$ chromosomes from tetrads of different rank (LANDE and STAHL 1993). The suitability of the $m$ values of 4 for Drosophila and 2 for Neurospora was further confirmed by the analyses of ZHAO et al. 1995.

RESULTS

Our symbolism is, as far as possible, that of FOSS et al. (1993) and LANDE and STAHL (1993) and is summarized in Table 1.
Estimating Interference

As in Foss et al., \( y = 2 ( m + 1 ) X \) is the mean number of C events per X Morgans, and \( p_i = y e^{-j/i} \) is the probability of \( i \) C events per X Morgans. Then, defining \( P_j \) as the probability of \( j \) C events per X Morgans,

\[
P_0 = \sum_{i=0}^{n} \frac{m+1-i}{m+1} p_i,
\]

and

\[
P_j = \sum_{i=-m}^{\infty} \frac{m+1-i}{m+1} p_j(m+1)+i \text{ for } j \geq 1,
\]

where \(|i|\) is the absolute value of \( i \).

In tetrads with \( j \) C events, the frequencies \( T \), \( NPD \) and \( PD \) are as follows (Mather 1935):

- For \( j = 0 \): \( T = 0 \), \( NPD = 0 \), \( PD = 1 \).
- For \( j = 1 \): \( T = 1 \), \( NPD = 0 \), \( PD = 0 \).
- For \( j \geq 2 \): \( T = \frac{j}{2} \left[ 1 - \left( \frac{1}{2j} \right) \right] \),
  \( NPD = PD = \left( \frac{j}{2} \right) \left( 1 - T \right) \).

Among all tetrads

\[
T = P_1 + \sum_{j=2}^{\infty} \frac{j}{2} \left[ 1 - \left( \frac{1}{2j} \right) \right] P_j,
\]

\[
NPD = \sum_{j=2}^{\infty} \left\{ \frac{1}{2} + \left( -\frac{1}{2j} \right) \right\} P_j,
\]

\[
PD = P_0 + NPD.
\]

Figure 1 depicts \( NPD \) and \( T \) vs. \( X \) for various values of \( m \) from 0 to 20. Note that \( T + NPD + PD = 1 \), and that for \( X \gg \frac{1}{2} \), \( T \to \frac{y}{2} \), \( NPD \to \frac{y}{2} \) and \( PD \to \frac{y}{2} \).

**Selecting a value of \( m \) from two-factor data:** In Figure 2, \( NPD \) is graphed vs \( T \) for various values of \( m \). Since the curves do not cross each other (except possibly for values of \( X > 1 \)), each pair of \( NPD, T \) values is determined by a unique value for \( m \). To estimate \( m \) from experimental data, plot the observed \( NPD, T \) value on (a xerographic copy of) Figure 2. If the point falls conspicuously closer to one curve than to any other, take that as your estimate of \( m \). Armed with that estimate, one can select a mapping function from the list below (Foss et al. 1993):

\[
R = \frac{1}{2} \left( 1 - e^{-2x} \right) \text{ for } m = 0,
\]

\[
R = \frac{1}{2} \left[ 1 - (1 + 2X) e^{-4X} \right] \text{ for } m = 1,
\]

\[
R = \frac{1}{2} \left[ 1 - (1 + 4X + 6X^2) e^{-6X} \right] \text{ for } m = 2,
\]

\[
R = \frac{1}{2} \left[ 1 - \left( 1 + 6X + 16X^2 + \frac{64}{3} X^3 \right) e^{-8X} \right] \text{ for } m = 3,
\]

\[
R = \frac{1}{2} \left[ 1 - e^{-2} \sum_{i=0}^{\infty} \frac{1}{i!} \left( 1 - \frac{i}{m+1} \right) \right] ;
\]

\[
y = 2 ( m + 1 ) X \text{ for } m = 4.
\]

**EXAMPLES**

**Neurospora:** Perkins (1962b, Table 7) observed the following numbers of tetrat types for the \( cr-or \) interval: 123 PD, 60 NPD and 418 T. The fraction of Ts is 0.696; the fraction of NPDs is 0.100. This point falls closest to the \( NPD vs. T \) curve for \( m = 2 \), in agreement with the value for \( m \) estimated by Foss et al. for Neurospora both from the \( C_r/C \) ratio and the dependence of \( S_K \) on linkage map distance. The nonelliptical 95% confidence region (Appendix) includes \( m \) values of 1 to 3.
(Figure 3). With \( m = 2 \) and with \( R = 0.448 \pm 0.011 \), determined from the frequencies of tetrade types, we estimate map distance \( X \) with the appropriate mapping function (by numerical approximation) to be 69.1 \( \pm \) 5.1 cM with an approximate 95% confidence interval of (60.5–82.0). This is (not significantly) larger than the map distance, 64.7, computed from the conventional 50 \( (T + 6NPD) \) (PERKINS 1949), which assumes a maximum of two exchanges.

Bole-Gowda et al. (1962, Table 1) noted 1238 PDs, 76 NPDs and 1614 Ts for the hist-2 to aur interval. The fraction of Ts is 0.551 and of NPDs is 0.026. These fractions determine a point about midway between the curves for \( m = 1 \) and \( m = 2 \), with a nonelliptical confidence region that includes the \( m = 1 \) curve and almost includes that for \( m = 2 \) (Figure 3).

Saccharomyces: Foss et al. (1993) failed to deal with yeast data because of the confusion resulting from gene conversions. As described above, this confusion is largely avoided in the present analysis because only two markers are involved, and they are far apart so that conversion is rare relative to reciprocal recombination. Fogel et al. (1981) reported data on the \( C_t/C \) ratio for several linked yeast markers. When they corrected their data for incidental exchanges (assuming no interference between a \( C_t \) and nearby \( C_t \)s), the mean \( C_t/C \) value was 0.37 with a range from 0.18 to 0.66 in five estimates. Assuming these data are like the Neurospora data in general features, a correction within the framework of the model of Foss et al. would drop the \( C_t/C \) ratio \( \sim 10\% \), to a mean of 0.33 with range from 0.16 to 0.60 (PERKINS et al. 1993). This value for \( C_t/C \) implies \( m = 2 \), with appreciable uncertainty. These \( C_t/C \) ratios have about the same mean value as do the values from Neurospora (mean = 0.30 \( \pm \) 0.01) (PERKINS et al. 1993). However, the Neurospora data have a narrower range, 0.18 to 0.38 in nine estimates.

Mortimer and Hawthorne (1966, Table 2) reported one interval for which 23 NPDs were scored. The PDs and Ts were 101 and 259, respectively. These values imply \( m = 2 \), with the confidence region compatible with \( m = 1 \) or 3 (Figure 4).

King and Mortimer (1991, Table 4) presented tetrade data for several lengths \( X > 0.5 \) intervals, reconstructed on the assumption of no chromatid interference. The \( NPD vs. T \) values for these intervals are shown, with their 95% confidence areas, in Figure 4.

There is considerable heterogeneity in the \( m \) values among the yeast data sets in Figure 4. Four of the data sets are compatible with \( m = 2 \), but two data sets indicate larger values of \( m \), one in the range of 4 to 5 and the other in the range of 5–9.

**DISCUSSION**

Although the biological basis underlying the model of Foss et al. (1993) has not been supported (Foss and Stahl 1995), the utility of the equations resulting from the model has been well demonstrated (Foss et al. 1993; Lande and Stahl 1993; Zhao et al. 1995). At the very least, therefore, the model will be useful as the formal framework for the quantification of interference and for the construction of linkage maps.
The value \( m = 2 \) obtained from the frequencies of Neurospora tetrad types is the same as that obtained from the observation that about \( 1/3 \) of conversions are accompanied by crossovers and from the fit of the model evaluated at \( m = 2 \) to the data for \( S_t \) vs. \( X \) (Foss et al. 1993).

The model of Foss et al. is an extension of a proposal by Mortimer and Fogel (1974) that Cs are Poisson-distributed but that the two subclasses, \( C_a \) and \( C_b \), alternate with each other. Their proposal was based on observations made in Saccharomyces. However, the variability in the estimates of \( m \) manifest in Figure 4 suggest that the application of the model to yeast will not be as straightforward as it appears to be for Neurospora (and see Foss and Stahl 1995). Zhao et al. (1995) report similar high variability in estimates of \( m \) from other sets of Saccharomyces data.

We note that for \( R < \) about 0.2 our estimate of \( X \) will be nearly the same as that of Perkins (1962b). For \( 0.4 < R < 0.5 \), our estimate of \( X \) will be substantially larger than that of Perkins, but, unless the sample size is very large \((n = \text{thousands})\), the estimates are not likely to be significantly different. Nevertheless, because our method is based on equations that provide a good description of interference, estimates of \( X \) derived by our method will generally be more accurate than those obtained by previous methods of analyzing two-point tetrad data.

Our text benefited from discussions with Eric Foss, Jeff King, Mary Sara McPeek, David Perkins, Terry Speed and Honolulu Zhao. This work was supported by National Institutes of Health (NIH) grant GM-33677 and National Science Foundation grant DMB8905310 to F.W.S. and by NIH grant GM-27120 to R.L. F.W.S. is American Cancer Society Research Professor of Molecular Genetics.

LITERATURE CITED


Communicating editor: P. J. Pukila

APPENDIX: CONFIDENCE INTERVALS FOR ESTIMATES OF \( m \) AND \( X \)

Russell Lande

Department of Biology, University of Oregon, Eugene, Oregon 97403

In a sample of \( n \) tetrads, the observed frequencies, \( \hat{T} \), \( \hat{NPD} \) and \( \hat{PD} \) are assumed to follow a trinomial distribution. Since the observed frequencies sum to 1, all of the information in the sample is contained in any two of the observed frequencies, e.g., \( \hat{NPD} \) and \( \hat{T} \). These have sampling variances and covariance:

\[
\text{Var}[\hat{NPD}] = \frac{\hat{NPD}(1-\hat{NPD})}{n},
\]

\[
\text{Var}[\hat{T}] = \frac{T(1-T)}{n},
\]

\[
\text{Cov}[\hat{NPD}, \hat{T}] = -\frac{T \cdot \hat{NPD}}{n},
\]

from which it can be shown that

\[
\text{Var}[\hat{R}] = \frac{[R(1-R) - T/4]}{n}.
\]

If the observed number of nonparental diotypes, \( n \cdot NPD \) is large, on the order of \( \geq 100 \), a 95% confidence ellipse can be drawn around the observed data point on the \( NPD \) vs. \( T \) graph using (1) and (2). Any
of the lines for different values of $m$ that intersect the ellipse give values of $m$ that are compatible with the data.

A more accurate method of determining a 95% confidence region (valid when $n$ is large) is based on the logit transformation of all frequencies (Hosmer and Lemeshow 1989; McCullagh and Nelder 1989),

$$f_1 = \ln \left( \frac{T}{1 - T} \right), \quad f_2 = \ln \left( \frac{NPD}{1 - NPD} \right),$$

which have approximate variances and covariance

$$\text{Var} [f_1] \approx \frac{1}{nT(1 - T)},$$
$$\text{Var} [f_2] \approx \frac{1}{nNPD (1 - NPD)},$$
$$\text{Cov} [f_1, f_2] \approx -\frac{1}{n} \left( \frac{1}{1 - T} + \frac{1}{1 - NPD} \right).$$

Let the column vector $(\hat{f}_1, \hat{f}_2)'$ (where prime indicates matrix transposition) have variance-covariance matrix $C$ with elements in (4) and (5). Because $C$ is symmetric, it can be represented by the spectral decomposition

$$C = M' \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} M,$$

where $\lambda_1$ and $\lambda_2$ are the (real positive) eigenvalues of $C$, and $M$ is a unitary, orthogonal matrix with columns that are the eigenvectors of $C$ normalized to unit length.

The 100 $(1 - \alpha)$% confidence region is described by the parametric relationship

$$T(\tau) = 1 / (1 - e^{-h^{(\tau)}}),$$
$$NPD(\tau) = 1 / (1 - e^{-h^{(\tau)}}),$$

with

$$\begin{pmatrix} f_1 (\tau) \\ f_2 (\tau) \end{pmatrix} = \sqrt{-2 \ln (1 - \alpha)} \begin{pmatrix} \lambda_1^{1/2} & 0 \\ 0 & \lambda_2^{1/2} \end{pmatrix} \times \begin{pmatrix} \cos (\tau) \\ \sin (\tau) \end{pmatrix} + \begin{pmatrix} \hat{f}_1 \\ \hat{f}_2 \end{pmatrix},$$

for $0 \leq \tau \leq 2\pi$.

An estimated recombination rate, $\hat{R} = NPD + \hat{T}/2$, has sampling variance given by Equation 3. For large samples, the 95% confidence interval around the estimate can be approximated by $\pm 2$ SDs around the estimate.

An estimate of map distance, $\hat{X}$, corresponding to an estimated recombination rate can be obtained using the mapping function with the closest value of $m$, either numerically from the formulas or graphically from Figure 5. The 95% confidence interval around $\hat{X}$ can be approximated either by using the appropriate mapping function to transform the endpoints of the corresponding confidence interval for $\hat{R}$, or, somewhat less accurately, by using $\pm 2$ SDs around the estimate. The sampling variance of $\hat{X}$ in large samples is approximately

$$\text{Var} [\hat{X}] = \frac{\text{Var} [\hat{R}]}{(dR/dX)^2},$$

where

$$\frac{dR}{dX} = e^{-y} \sum_{i=0}^{m} \frac{y^i}{i!} \quad \text{with} \quad y = 2 (m + 1) X.$$