Theor Appl Genet (1983) 66:313–321

© Springer-Verlag 1983

Estimation of recombination frequency in genetic linkage studies

E. V. Nordheim [1, 2], D. M. O'Malley [1] and R. P. Guries [1]

Department of Forestry [1], and Department of Statistics [2], University of Wisconsin, 120 Russell Laboratories, 1630 Linden Drive, Madison, WI 53706, USA

Received April 26, 1983
Communicated by A. L. Kahler

Summary. A binomial-like model is developed that may be used in genetic linkage studies when data are generated by a testcross with parental phase unknown. Four methods of estimation for the recombination frequency are compared for data from a single group and also from several groups; these methods are maximum likelihood, two Bayesian procedures, and an ad hoc technique. The Bayes estimator using a noninformative prior usually has a lower mean squared error than the other estimators and because of this it is the recommended estimator. This estimator appears particularly useful for estimation of recombination frequencies indicative of weak linkage from samples of moderate size. Interval estimates corresponding to this estimator can be obtained numerically by discretizing the posterior distribution, thereby providing researchers with a range of plausible recombination values. Data from a linkage study on pitch pine are used as an example.

Key words: Linkage – Recombination – Folded binomial – Bayesian estimation – Pitch pine

Introduction

A large literature spanning over 50 years exists in statistical genetics on detection and estimation of linkage in situations where parental information is incomplete. Prominent early work includes Fisher (1935), Haldane and Smith (1947), Morton (1955), and Smith (1959); Bailey (1961) provides a useful summary. Likelihood techniques have become the predominant methodology although Bayesian methods also find some use in practice. In recent years, research efforts have largely focused on refining likelihood procedures for linkage testing (Smith and Sturt 1976; Rao et al. 1978).

Linkage analysis in plant and animal genetics has focused on controlled breeding experiments in which parental phase is known (Tanksley and Rick 1980; Goodman et al. 1980). However, situations exist where breeding work is impractical because of long generation times and/or high cost, and estimates of recombination frequency typically are obtained from studies in which parental phase is unknown. This situation is common, for example, in forest genetics (Rudin and Ekberg 1978; Adams and Joly 1980; O'Malley and Guries 1981).

Genetic linkage studies reflecting a common mating type (test-cross, parental phase unknown) make use of the binomial-like model described by the following probability function:

$$P\{k \mid n, \theta\} = \begin{cases} \binom{n}{k}[\theta^k(1-\theta)^{n-k} + \theta^{n-k}(1-\theta)^k] & \text{for } 0 \leqq k < n/2 \\ \frac{1}{2}\binom{n}{k}[\theta^k(1-\theta)^{n-k} + \theta^{n-k}(1-\theta)^k] & \text{for } k = n/2 \end{cases} \quad (1)$$

where n is the sample size (e.g., number of gametes), k is the number of observations in the smaller class (coupling or repulsion), and θ is the recombination frequency, restricted to lie between 0 and 0.5. The genes are said to be unlinked if $\theta = 0.5$; if $\theta < 0.5$, the genes are linked with linkage considered tighter as θ becomes smaller. In (1) the term $\theta^k(1-\theta)^{n-k}$ is from the standard binomial, whereas the term $\theta^{n-k}(1-\theta)^k$ is due to the restrictions that θ lie between 0 and 0.5, and that k is the number of observations in the smaller class (but not necessarily the number of recombinants). Equation (1) can be viewed as a "folded binomial" to describe these restrictions. If θ is close to 0.5, there is a non-negligible probability that the number of recombi-

nants will exceed n/2. The restrictions noted lead to a "folding" about n/2 of the distribution for the number of recombinants so that the occurrences of k recombinants and n – k recombinants have the same probability.

Analytical study of (1) has received little recent attention. Human geneticists primarily apply likelihood methods with some attention devoted to Bayesian approaches. Forest geneticists generally have avoided formal use of (1) and have estimated θ with the binomial estimator k/n (Rudin and Ekberg 1978; Adams and Joly 1980). The purpose of this paper is to examine the estimation of θ from (1). In particular, several methods of estimation for a single group or family as well as for several groups or families are compared.

Two assumptions are made in the analysis presented here. The first assumption is that there is no linkage disequilibrium. Strong disequilibria have been noted in selfing plants (Allard 1975), and among tightly linked loci in maize (Brown and Allard 1971), but such associations seem to be uncommon in outbreeding species, including forest trees. Detection of disequilibrium often requires large sample sizes (Brown 1975), which might explain the limited number of observations on this phenomenon in natural populations. The second assumption is absence of ascertainment bias. Failure to incorporate such bias where appropriate will lead to negligible errors in estimation for moderate to large samples (eg., n > 20).

Estimation of θ from a single group

Four methods of estimation for θ in model (1) are considered. These are maximum likelihood, a Bayes procedure with a noninformative prior, a Bayes procedure with a spike of prior probability placed at $\theta = 0.5$, and an ad hoc "natural" procedure. These methods are described briefly and then compared numerically.

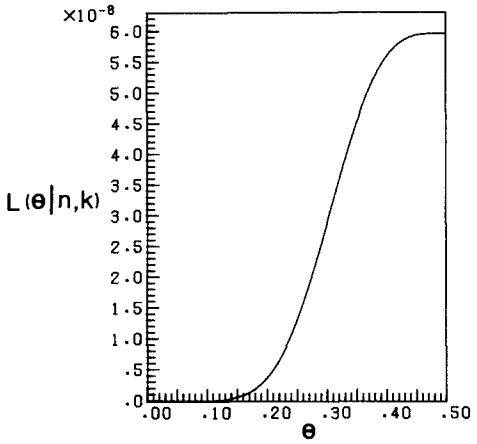

Fig. 1. $L(\theta \mid n, k)$ versus θ for $n = 25$, $k = 10$

Denote the likelihood by $L(\theta \mid n, k)$ which is expressed as $L(\theta \mid n, k) \propto \theta^k (1-\theta)^{n-k} + \theta^{n-k}(1-\theta)^k$. The maximum likelihood estimator, $\hat{\theta}_1$, is that value of θ which maximizes $L(\theta \mid n, k)$; in general, this value must be found numerically. The maximum value of the likelihood can occur at $\theta = 0.5$ even if k is less than n/2. When $L(\theta \mid n, k)$ is plotted versus θ for $n = 25$ and $k = 10$, the peak occurs at $\theta = 0.5$ (Fig. 1); with $n = 25$ the peak also occurs at $\theta = 0.5$ if $k = 11$ or $k = 12$.

Bayesian estimation makes use of $L(\theta \mid n, k)$ in a different manner. $L(\theta \mid n, k)$ is modified by a prior probability on θ and results in a posterior probability for θ. The Bayes estimator is the mean (average value) for θ from the posterior distribution.

The first Bayes estimator, $\hat{\theta}_2$, is the posterior mean assuming a (non-informative) uniform prior distribution (Jeffreys 1961). This estimator is given by:

$$\hat{\theta}_2 = \int_0^{0.5} \theta\, L(\theta \mid n, k)\, d\theta \Big/ \int_0^{0.5} L(\theta \mid n, k)\, d\theta$$

$$= \frac{k+1}{n+2} I_{0.5}(k+2, n-k+1) + \frac{n-k+1}{n+2} I_{0.5}(n-k+2, k+1)$$

where $I_x(a, b)$ is the incomplete beta function as defined by eqn. (26.5.1) in Abramowitz and Stegun (1972). Because $x = 0.5$ and a, b are integer-valued, $\hat{\theta}_2$ can be computed exactly by use of the relation between the incomplete beta and binomial distributions (Abramowitz and Stegun 1972 – 26.5.24).

Bayesian methods were advocated by Smith (1959) primarily for use in testing $H_0: \theta = 0.5$. It has been traditional to use a prior distribution with a spike of probability, β, at $\theta = 0.5$ and uniform density on $0 \leqq \theta < 0.5$ such that the total probability on $0 \leqq \theta < 0.5$ is $1-\beta$. In large part the spike reflects the high probability that a pair of loci are unlinked due to their occurrence on distinct chromosomes. The value of β depends on the number of chromosome pairs in the organism under study. The Bayes estimator, $\hat{\theta}_3$, is the posterior mean assuming this spiked prior. This estimator is written

$$\hat{\theta}_3 = \left[\beta\, 0.5^n + 2(1-\beta) \int_0^{0.5} \theta\, L(\theta \mid n, k)\, d\theta\right] \Big/ \left[\beta\, 0.5^{n-1} + 2(1-\beta) \int_0^{0.5} L(\theta \mid n, k)\, d\theta\right].$$

The fourth estimator can be viewed as a "natural" estimator for a model with the binomial-like structure of (1). This ad hoc estimator is written $\hat{\theta}_4 = k/n$. Use of $\hat{\theta}_4$ has been made in some studies of genetic linkage in conifers (Rudin and Ekberg 1978). Adams and Joly (1980) make use of this estimator but apply it only if a

prior test for detection of linkage leads to a conclusion that linkage is present.

In order to compare these four estimators, the bias and mean squared error (MSE) were computed for 50 values of θ ranging from 0.01 to 0.50 in increments of 0.01 with n = 10, 25, and 100. (The MSE of an estimator is the square of the bias of the estimator plus the variance of the estimator. The MSE serves as the principal criterion for comparing estimators.) The values of n considered are representative of sample sizes of interest in practice, particularly to forest geneticists.

Equation (1) was used to determine the probability of each possible value of k for each combination of θ and n. For each k, estimates for θ were calculated for each estimation method. For each method, the estimates were averaged, weighting each estimate by the probability of observing the value of k leading to the estimate, and the bias and MSE were computed. The value of β, the prior probability that $\theta = 0.5$, used for $\hat{\theta}_3$ was 0.94; this value was chosen as typical for many coniferous tree species but the nature of conclusions is not strongly dependent on the specific selected value. The results for bias and MSE for $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_4$ are displayed in Fig. 2 with MSE results comparing $\hat{\theta}_2$ and $\hat{\theta}_3$ shown in Fig. 3. The bias results for $\hat{\theta}_3$ are not displayed as little useful information would be added.

In the case of the usual binomial problem, the estimator k/n is unbiased. However, for the folded binomial, the "natural" estimator $\hat{\theta}_4$ is unbiased only for those values of θ where the folding has no real effect; i.e., those values of θ for which $\theta^{n-k}(1-\theta)^k$ is much smaller than $\theta^k(1-\theta)^{n-k}$. There is an increase in the negative bias of $\hat{\theta}_4$ as folding becomes more important. Also, the value of θ when folding begins to have effect increases as n becomes larger.

The maximum likelihood estimator $\hat{\theta}_1$ has bias similar to that of $\hat{\theta}_4$ for small θ. The bias becomes positive for intermediate θ and becomes negative for θ close to 0.5. With intermediate values of θ, there is a non-negligible probability that the estimator value is 0.5 (Fig. 1); this causes positive bias. The bias for $\hat{\theta}_1$ is negative for θ close to 0.5 due to the boundary at $\theta = 0.5$. The groupings small θ, intermediate θ, and θ close to 0.5 change with changing n, according to the effect of folding.

The Bayes estimator with noninformative prior, $\hat{\theta}_2$, tends to shrink extreme estimates towards the middle, thus resulting in positive bias for small θ and negative bias for large θ. The Bayes estimator with spiked prior, $\hat{\theta}_3$, has large positive bias (not shown) for most values of θ below 0.5 due to the spike. The maximum values of this bias are 0.23 (at $\theta = 0.14$), 0.14 (at $\theta = 0.28$), and 0.073 (at $\theta = 0.39$) for n = 10, 25, and 100, respectively.

The MSE values for estimators $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_4$ manifest surprisingly variable behavior (Fig. 2 b). As noted above, k/n is an unbiased estimator in the case of the standard binomial problem. Thus, the MSE of this estimator is equal to its variance and is given by $\theta(1-\theta)/n$. For those values of θ with inconsequential folding, $\hat{\theta}_4$ ("natural") follows the same pattern. The MSE for $\hat{\theta}_4$ has smaller value as folding becomes important, due to substantially decreased estimator variability. The MSE increases sharply due to the bias as θ approaches 0.5, and equals $\theta(1-\theta)/n$ when $\theta = 0.5$. For small θ, $\hat{\theta}_1$ (maximum likelihood) has MSE similar to that of $\hat{\theta}_4$. The MSE increases due to the positive bias as folding becomes important. The MSE for $\hat{\theta}_1$ is below the peak value when θ is close to 0.5, because the probability that $\hat{\theta}_1$ equals 0.5 is substantial.

The Bayes estimator with noninformative prior, $\hat{\theta}_2$, has a larger MSE than that for $\hat{\theta}_1$ and $\hat{\theta}_4$ at very small θ and again for θ close to 0.5 due to substantial bias. However, for a broad range of intermediate θ, the MSE for $\hat{\theta}_2$ is considerably less than that for the others, due largely to the shrinkage effect noted above. Estimator $\hat{\theta}_3$ (Bayes with spike prior) has large MSE values corresponding to the large biases resulting from the spike. For larger n, the MSE values for $\hat{\theta}_2$ and $\hat{\theta}_3$ are identical for small θ; however, as θ becomes large enough so that the prior dominates the likelihood, the MSE for $\hat{\theta}_3$ becomes considerably larger. The MSE for $\hat{\theta}_3$ is very small for θ very close to 0.5, due to the spike (Fig. 3).

An interval estimate for θ can be obtained utilizing highest posterior density regions as discussed in Box and Tiao (1973). This "Bayes confidence interval" corresponds well with estimator $\hat{\theta}_2$. Such interval estimates are obtained numerically by discretizing the posterior distribution of θ.

Estimation of θ for several groups

Consider several (T) independent groups with common θ, each distributed according to (1). The likelihood is written

$$L(\theta) \equiv L(\theta \mid n_i, k_i, i = 1, \ldots, T) \propto \prod_{i=1}^{T} [\theta^{k_i}(1-\theta)^{n_i-k_i} + \theta^{n_i-k_i}(1-\theta)^{k_i}] \qquad (2)$$

with $0 \leqq \theta \leqq 0.5$; k_i integer-valued so that $0 \leqq k_i \leqq n_i/2$.

The same four methods of estimation are considered. The maximum likelihood estimator $\hat{\theta}_1$ is that value of θ which maximizes $L(\theta)$. The Bayes estimator with noninformative prior is written

$$\hat{\theta}_2 = \int_0^{0.5} \theta L(\theta)\, d\theta \Big/ \int_0^{0.5} L(\theta)\, d\theta .$$

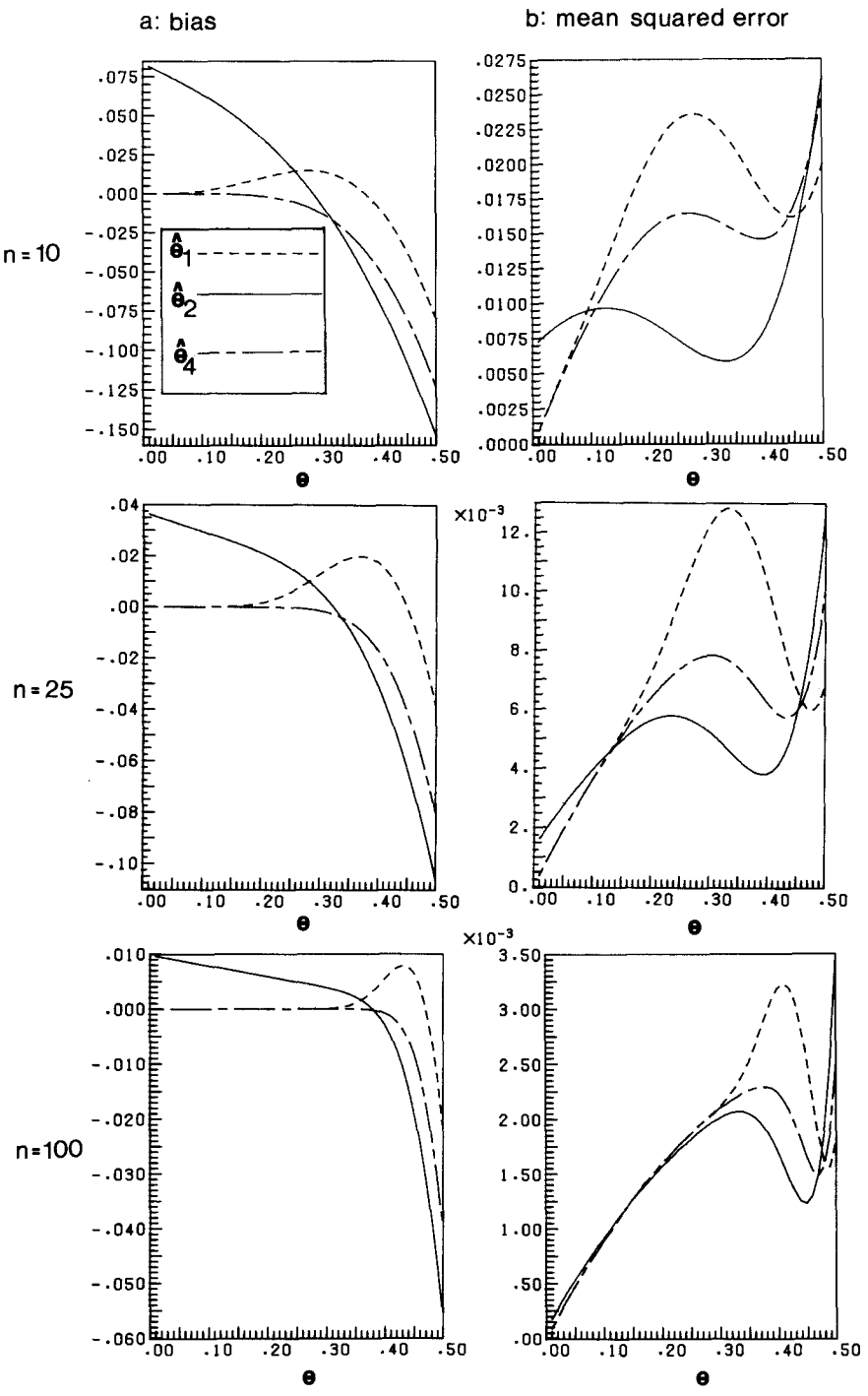

Fig. 2. Bias and mean squared error versus θ for $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_4$; single group case

This can be expressed as the ratio of sums of incomplete beta functions where these latter functions can be evaluated exactly, as indicated previously. The Bayes estimator with spike prior is written

$$\hat{\theta}_3 = \left[\beta\, 0.5^{N-T+1} + 2(1-\beta)\int_0^{0.5} \theta\, L(\theta)\, d\theta\right] \Big/ \left[\beta\, 0.5^{N-T} + 2(1-\beta)\int_0^{0.5} L(\theta)\, d\theta\right]$$

where $N = \sum_{i=1}^{T} n_i$. The "natural" estimator is written $\hat{\theta}_4 = \sum_{i=1}^{T} k_i \Big/ \sum_{i=1}^{T} n_i$. Note that $\hat{\theta}_4$ is equivalent to the estimator obtained for a single family with $K = \sum_{i=1}^{T} k_i$ and $N = \sum_{i=1}^{T} n_i$.

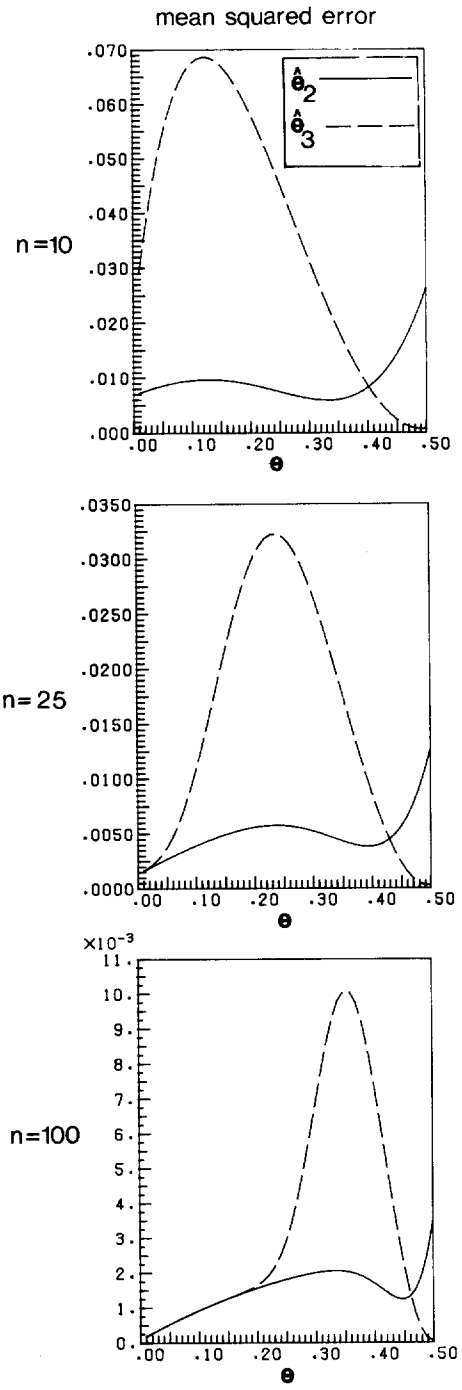

Fig. 3. Mean squared error versus θ for $\hat{\theta}_2$, $\hat{\theta}_3$; single group case

To compare estimators, the bias and MSE were computed for the 4-group situation (T=4). The balanced sample case with $n_i=10$ and $n_i=25$ was considered. The bias and MSE for $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_4$ are plotted versus θ (Fig. 4). Many of the same patterns noted for the single family situation are evident here. However, for the "natural" estimator $\hat{\theta}_4$, the bias becomes negative more rapidly than for other estimators due to the inefficient manner with which $\hat{\theta}_4$ incorporates folding. Similarly the MSE is relatively large for θ close to 0.5 due to this bias. The difficulties with this estimator are compounded when there are more groups. The results for $\hat{\theta}_3$ (Bayes with spike prior, not shown) compare with those for $\hat{\theta}_2$ in the same way as in the single group case.

Comparing the 4-group $n_i=25$ case to the single group $n=100$ situation, note that the bias and MSE values for $\hat{\theta}_1$ and $\hat{\theta}_2$ are similar. The magnitudes are somewhat greater in the 4-group case for θ close to 0.5. These two estimators appear to incorporate the effects of folding quite efficiently.

Example from *Pinus rigida* (Mill.)

To demonstrate the applicability of these procedures, consider data from a genetic linkage study of pitch pine (*Pinus rigida* Mill.). In this study linkage between certain enzyme loci was examined using haploid megagametophytes from open-pollinated seed of single-tree collections. Three pairs of loci for which double heterozygotes were available are considered (Table 1); we arbitrarily designate by A or a and B or b the alternative allelic forms for each locus. By simultane-

Table 1. Data from pitch pine linkage study

Enzyme loci	Accession no.	No. of megagametophytes per class					
		AB	Ab	aB	ab	n	k
ACO : AAT-1	CC438	9	7	10	8	34	17
	CC389	14	7	11	8	40	18
	HE54	9	12	12	7	40	16
	BRD131	7	8	10	14	39	18
	MX95	12	16	9	8	45	20
G6PD : PGM-1	WPA22	5	17	10	9	41	14
	WPA7	14	7	10	13	44	17
	EP49	11	15	13	6	45	17
	CC430	10	8	19	7	44	17
PGI-1 : 6PGD-2	CC388	9	33	33	15	90	24
	EP148	14	7	7	12	40	14

E.C. nomenclature for these enzymes:

Enzyme Name	Abbreviation	E.C. Designation
aconitase	ACO	3.1.3.2
aspartate aminotransferase	AAT-1	2.6.1.1
glucose-6-phosphate dehydrogenase	G6PD	1.1.1.49
phosphoglucomutase	PGM-1	2.7.5.1
phosphoglucose isomerase	PGI-1	5.3.1.9
6-phosphogluconic dehydrogenase	6PGD-2	1.1.1.44

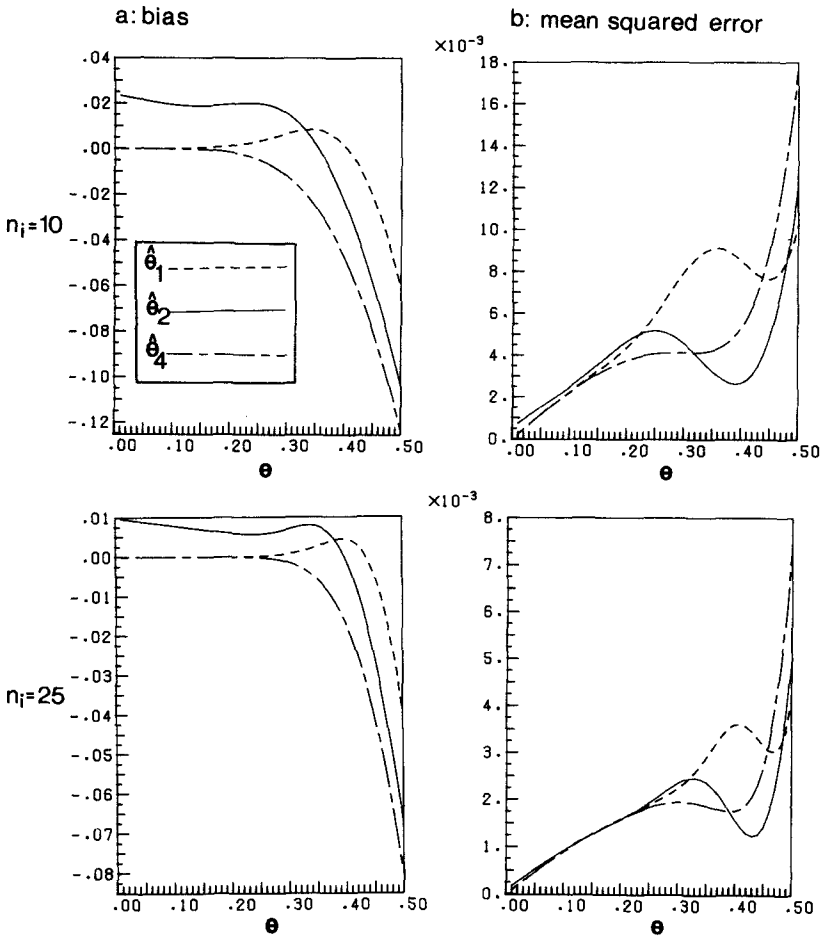

Fig. 4. Bias and mean squared error versus θ for $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_4$; four group case

ously staining for two enzymes on slices from the same gel (Guries et al. 1978), data can be obtained on the number of gametes that fall into each of the four categories AB, Ab, aB, and ab. The numbers k and n are given by

$$k = \min(m_{AB} + m_{ab}, m_{Ab} + m_{aB})$$

$$n = m_{AB} + m_{Ab} + m_{aB} + m_{ab}$$

where, for example, m_{AB} is the number of gametes in category AB. (Note: the notation used for enzyme loci follows from Guries et al. 1978).

The four estimation methods were applied to the data from each individual tree and to the grouped data for each pair of loci. Estimates of θ together with 95% "Bayes confidence intervals" and the posterior probability that $\theta = 0.5$ assuming the spike prior are shown in Table 2. All values are accurate to three decimal places. The interval estimates for the grouped data are obtained in the same manner as those for the single groups.

The strong effect of the spike prior causes $\hat{\theta}_3$ to behave in a qualitatively different fashion from the other estimators. However, for several individual trees (CC438, CC389, BRD131, MX95, and CC388), and the combined data for the first and third gene pairs (ACO : AAT-1 and PGI-1 : 6PGD-2, respectively), $\hat{\theta}_3$ is similar to the other estimators. For the individual tree (exempting HE54) and grouped data from ACO : AAT-1, all estimates indicate that little or no linkage exists. For tree CC388 and the grouped data from PGI-1 : 6PGD-2, the likelihood dominates the spike prior and all estimates are comparable. In this regard note the low posterior probability that $\theta = 0.5$.

Estimators $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_4$ show more uniformity when θ is well below 0.5 than when θ is close to 0.5. For tree CC438, where $k = n/2$, $\hat{\theta}_2$ (Bayes with non-informative prior) equals 0.434 demonstrating the shrinkage due to the prior. It is interesting to compare the differences between $\hat{\theta}_1$ (maximum likelihood) and $\hat{\theta}_2$ for trees WPA22, EP148, EP49, WPA7 (CC430), and HE54. Whereas both estimators increase monotonically, $\hat{\theta}_1$ increases faster; $\hat{\theta}_1$ changes from 0.342 to 0.410 whereas $\hat{\theta}_2$ goes from 0.348 to 0.397. This behavior demonstrates the differing responses of these estimators to folding. With the grouped data the

Table 2. Estimates of θ from four methods

	$\hat{\theta}_1$ (maximum likelihood)	$\hat{\theta}_2$ (Bayes-non-informative prior)	$\hat{\theta}_3$ (Bayes-spike prior)	$\hat{\theta}_4$ ("natural")	95% "Bayes confidence interval"	Posterior probability that $\theta=0.5$ (given $\beta=0.94$)
CC438	0.500	0.434	0.499	0.500	[0.339, 0.500]	0.987
CC389	0.500	0.427	0.499	0.450	[0.327, 0.500]	0.985
HE54	0.410	0.397	0.497	0.400	[0.284, 0.500]	0.973
BRD131	0.500	0.431	0.499	0.462	[0.334, 0.500]	0.986
MX95	0.500	0.427	0.499	0.444	[0.329, 0.500]	0.985
ACO : AAT-1	0.500	0.462	0.500	0.449	[0.408, 0.500]	0.992
WPA22	0.342	0.348	0.487	0.341	[0.227, 0.492]	0.913
WPA7	0.389	0.387	0.496	0.386	[0.276, 0.500]	0.965
EP49	0.379	0.380	0.495	0.378	[0.270, 0.500]	0.958
CC430	0.389	0.387	0.496	0.386	[0.276, 0.500]	0.965
G6PD : PGM-1	0.375	0.380	0.479	0.374	[0.301, 0.461]	0.821
CC388	0.267	0.272	0.273	0.267	[0.183, 0.363]	0.005
EP148	0.350	0.355	0.490	0.350	[0.240, 0.500]	0.931
PGI-1 : 6PGD-2	0.292	0.296	0.296	0.292	[0.219, 0.374]	0.003

similarity of estimate values ($\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_4$) for G6PD : PGM-1 and particularly PGI-1 : 6PGD-2 is not surprising given the relatively large sample sizes.

Discussion

Both likelihood and Bayesian methods (assuming a spike prior) have been used for detection of linkage (Morton 1955; Smith 1959). A lively debate among adherents of both points of view has continued to the present. A procedure utilizing the spike (i.e., using $\hat{\theta}_3$) does not seem desirable for estimation of recombination frequency due to the substantial effect the prior can have on bias and mean squared error (MSE). The spike prior is motivated by the knowledge that a high proportion, β, of all pairs of loci will be unlinked largely because the probabilities are high that they fall on different chromosomes. However, when estimating the recombination frequency for a particular gene pair, it appears quite undesirable to impose the influence of such a powerful prior on the estimation procedure. The strength of this effect is demonstrated by the single tree and grouped data results for G6PD : PGM-1 (Table 2).

Traditionally, estimation of θ has involved maximum likelihood (i.e., $\hat{\theta}_1$) as the recommended procedure. We advocate a role for $\hat{\theta}_2$, the Bayesian estimator with noninformative prior. This estimator, as noted in previous sections, has substantially smaller MSE for a wide range of intermediate θ values. Additionally, this estimator has associated with it a readily computed interval estimator, the highest posterior density region. The results of our estimator comparisons (Figs. 2 and 4) indicate that $\hat{\theta}_2$ has a relatively large bias for θ near 0 and also near 0.5. However, the general superiority of $\hat{\theta}_2$ (ie., smaller MSE) is due to a substantially reduced variance that more than compensates for the increased bias. The major difficulty with $\hat{\theta}_1$ (maximum likelihood) is that this estimator yields rather high values when k is near n/2 and equals 0.5 for a range of k. This behavior leads to a larger estimator variance, and hence MSE, particularly for intermediate θ. The estimator $\hat{\theta}_2$ avoids this difficulty.

Some mild cautionary comments should be offered on the use of $\hat{\theta}_2$. First, it appears important to avoid performing a test for $H_0: \theta=0.5$ on the basis of inclusion of 0.5 in a 95% "Bayes confidence interval". In making conclusions about the existence of linkage, it appears appropriate to take into account the large prior probability that loci are unlinked due to the number of chromosomes. However, if a 99.9% interval were computed, the corresponding test would provide similar results to the lod (log odds) score, or z score procedure, based on the Morton (1955) likelihood methods. Second, the one region for which $\hat{\theta}_2$ has undesirable behavior (i.e., relatively large MSE) is for (true) θ very close to 0.5; the shrinking effect of the prior keeps the estimate values below 0.5. Note again the results for tree CC438 (Table 2). However, this concern appears quite minor, for if the given interval estimate includes $\theta=0.5$, then even a testing procedure using the interval cannot rule out the hypothesis of unlinked loci. In such

a case there appears little point in providing an estimate of θ other than to conclude that the loci are unlinked. For (true) θ close to 0.5, detection of linkage and good estimation will require large sample sizes regardless of specific methodology.

Although likelihood and Bayesian procedures have received predominant attention, forest geneticists have made use of estimators like $\hat{\theta}_4$ (Rudin and Ekberg 1978). This method works well when the numbers of gametes in the two observed categories (coupling and repulsion) are quite disparate and it appears safe to identify the category with the smaller number as recombinant. However, there are problems with the procedure when the categories have comparable numbers of gametes and the certainty of correct identification of the recombinant category diminishes. These difficulties are enhanced when there are several groups as indicated by the relatively large MSE values for large θ.

As noted earlier, most linkage work in plant and animal genetics has assumed known parental phase. Biochemical methods have made it possible to survey genetic variation at many loci, but the characterization and mapping of such genes can entail considerable experimental effort. Detailed genetic maps have been constructed using allozyme data for several well-studied plants and animals (McMillin and Scandalios 1981; Treat-Clemons and Doane 1982). Most organisms for which such maps exist have relatively short life cycles, a low chromosome number, and/or other features which facilitate controlled breeding experiments from which data for such maps are derived. However, due to high cost or long generation times, breeding for such purposes is impractical in organisms such as forest trees and linkage testing is limited to gene combinations revealed in surveys of natural populations or seed orchards (Guries et al. 1978; Conkle 1981). This latter application requires estimation of recombination frequency assuming unknown phase; thus it is useful to determine to what extent estimator precision differs from the phase known case.

The testcross model reduces to a standard binomial with usual estimator k/n when parental phase is known. Neglecting those exceedingly rare situation when the true recombination frequency, θ, can be larger than 0.5 (pseudolinkage, Wright et al. 1980), an estimation procedure requiring the estimator to lie within the interval [0, 0.5] is obtained by the replacement of k/n by 0.5 whenever k/n > 0.5. This estimator, defined as $\hat{\theta}_0$, can be written $\hat{\theta}_0 = \text{minimum}\,(k/n, 0.5)$.

Comparisons of efficiency of different models for estimating recombination frequency are often made using the concept of information (Allard 1956). Such an approach does not appear feasible in the current situation due to violation of boundary conditions at $\theta = 0.5$ (Rao 1973). A comparison based on estimator mean squared error, with calculations performed in the same manner as before, was employed. Using $\hat{\theta}_2$ (Bayes with noninformative prior) as the best estimator for the phase unknown situation, this comparison of MSE values was made for the same single group and multigroup cases as before. (In the multi-group case define $\hat{\theta}_0$ by $\hat{\theta}_0 = \text{minimum}\left(\sum_{i=1}^{T} k_i / \sum_{i=1}^{T} n_i, 0.5\right)$. The single group results with $n = 25$ and the four group results with $n_i = 10$ for each group typify the overall pattern (Fig. 5). For reference, the standard binomial MSE, $\theta(1-\theta)/n$, $\left(\theta(1-\theta)/\sum_{i=1}^{T} n_i \text{ for the multigroup case}\right)$ is also displayed and denoted as $\hat{\theta}_b$.

There is little difference between the estimators for small θ because there is no uncertainty about the recombinants in the phase unknown situation; the small difference is due to shrinkage effects of $\hat{\theta}_2$. The MSE for $\hat{\theta}_2$ is smaller than that for $\hat{\theta}_0$ for intermediate θ, because the shrinkage properties dominate the effects of known phase; for θ close to 0.5, $\hat{\theta}_0$ achieves smaller MSE values than $\hat{\theta}_2$.

Overall there appears no loss in terms of MSE in estimation of θ utilizing the phase unknown model. Virtually nothing is gained by knowledge of parental gene arrangement. The results of this comparison might differ somewhat if an alternative estimator for $\hat{\theta}_0$ (Bayesian counterpart) were chosen. However, the general conclusion remains valid.

The principal area for application of $\hat{\theta}_2$ is likely to be survey work as in the pitch pine allozyme study. In

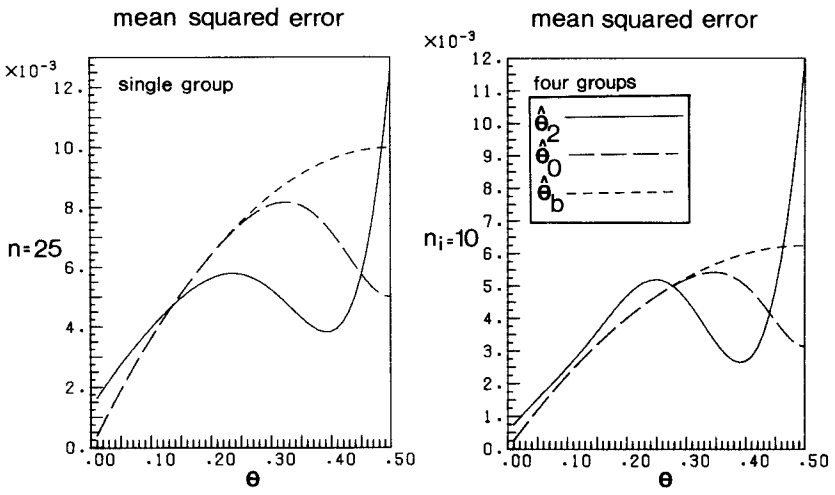

Fig. 5. Mean squared error versus θ for $\hat{\theta}_2, \hat{\theta}_0, \hat{\theta}_b$ (standard binomial)

such studies modest sample sizes (as typified by Table 1) are available for a large number of pairs of loci. Analysis of data such as the combined data on G6PD : PGM-1 can benefit most from use of this estimator; it is for intermediate values of θ where the reduction in MSE for $\hat{\theta}_2$ is the largest (Fig. 4). Although the evidence that the loci are linked cannot be viewed as conclusive – note for instance that the posterior probability assuming a spike prior for $\theta = 0.5$ is 0.821 – a stronger case for linkage can clearly be made for this pair of loci than for ACO : AAT-1. The fact that this posterior probability value of 0.821 is below the prior value of 0.94 indicates that the likelihood curve is shifted away from $\theta = 0.5$. The upper limit of 0.461 for the interval estimate for G6PD : PGM-1 should be viewed cautiously as implied above. However, the lower limit of 0.301 can serve as a conservative lower bound for the recombination frequency.

There appears no singularly superior methodology for analysis of linkage data with unknown phase. However, with weak linkage and modest sample sizes, the estimator $\hat{\theta}_2$ has a smaller MSE than other estimators. Although we would recommend against exclusive reliance on this estimator, we feel that it could be useful in many studies. We would also like to provide a note of caution about the "natural" estimator $\hat{\theta}_4$. This estimator performs relatively poorly (high MSE) in the several-group case for θ values indicative of weak linkage and hence should probably not be the estimator of choice.

Acknowledgements. This project was supported in part by McIntire-Stennis Project No. 142-C385 and the College of Agricultural and Life Sciences, University of Wisconsin. The authors would like to thank C. Denniston, J. J. Rutledge, an associate editor and two anonymous referees for helpful comments.

References

Abramowitz M, Stegun IA (1972) Handbook of mathematical functions. Dover, New York

Adams WT, Joly RJ (1980) Linkage relationships among twelve allozyme loci in loblolly pine. J Hered 71:199–202

Allard RW (1975) The mating system and microevolution. Genetics 79s: 115–126

Allard RW (1956) Formulas and tables to facilitate the calculation of recombination values in heredity. Hilgardia 24:235–278

Bailey NTJ (1961) Introduction to the mathematical theory of genetic linkage. Clarendon, Oxford

Box GEP, Tiao GC (1973) Bayesian inference in statistical analysis. Addison-Wesley, Reading, Mass

Brown AHD (1975) Sample sizes to detect linkage disequilibrium between two or three loci. Theor Popul Biol 8:184–201

Brown AHD, Allard RW (1971) Effect of reciprocal recurrent selection on isozyme polymorphisms in maize (*Zea mays* L.). Crop Sci 11:888–893

Conkle MT (1981) Isozyme variation and linkage in six conifer species. In: Conkle MT (ed) Isozymes of North American forest trees and forest insects. USDA Forest Ser Gen Tech Rpt PSW-48: 11–17

Fisher RA (1935) The detection of linkage with 'dominant' abnormalities. Ann Eugen 6: 187–201

Goodman MM, Stuber CW, Newton K, Weissinger HH (1980) Linkage relationships of 19 enzyme loci in maize. Genetics 96:697–710

Guries RP, Friedman ST, Ledig FT (1978) A megagametophyte analysis of genetic linkage in pitch pine (*Pinus rigida* Mill.). Heredity 40:309–314

Haldane JBS, Smith CAB (1947) A new estimate of the linkage between the genes for haemophilia and colour-blindness in man. Ann Eugen 14: 10–31

Jeffreys H (1961) Theory of probability. Clarendon Press, Oxford

McMillin DE, Scandalios JG (1981) Chromosome location of genes coding for biochemical markers in *Zea mays*. Isozyme Bull 14: 13–18

Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:277–318

O'Malley DM, Guries RP (1981) Detecting linkage from conifer genotypic survey data. Isozyme Bull 14:50

Rao CR (1973) Linear statistical inference and its application. Wiley, New York

Rao DC, Koats BJB, Morton NE, Yee S, Lew R (1978) Variability of human linkage data. Am J Hum Genet 30:516–529

Rudin D, Ekberg I (1978) Linkage studies in *Pinus sylvestris* L. using macrogametophyte allozymes. Silvae Genet 27: 1–12

Smith CAB (1959) Some comments on the statistical methods used in linkage investigations. Am J Hum Genet 11: 289–304

Smith CAB, Sturt E (1976) The peak of the likelihood curve in linkage testing. Ann Hum Genet 39:423–426

Tanksley SD, Rick CM (1980) Isozymic linkage map of the tomato: applications in genetics and breeding. Theor Appl Genet 57: 161–170

Treat-Clemons LG, Doane WW (1982) Biochemical loci of the "fruit fly" (*Drosophila melanogaster*). Isozyme Bull 15:7–24

Wright JE, May B, Stoneking M, Lee GM (1980) Pseudolinkage of the duplicate loci for supernatent aspartate aminotransferase in brook trout (*Salvelinus pontinalis*). J Hered 71:223–228