Statistical Analysis of Crossover Interference Using the Chi-square Model

Hongyu Zhao, Terence P. Speed and Mary Sara McPeek

Department of Statistics

University of California at Berkeley

Berkeley, CA 94720

February 1995

Running head: Chi-square Model for Interference

Abstract

The chi-square model (also known as the gamma model with integer shape parameter) for the occurrence of crossovers along a chromosome was first proposed in the 1940's as a description of interference which was mathematically tractable, but without biological basis. Recently, the chi-square model has been reintroduced into the literature from a biological perspective. It arises as a result of certain hypothesized constraints on the resolution of randomly distributed crossover intermediates. In this paper, under the assumption of no chromatid interference, the probability for any single spore or tetrad joint recombination pattern is derived under the chi-square model. The method of maximum likelihood is then used to estimate the chi-square parameter m and genetic distances among marker loci. We discuss how to interpret the goodness-of-fit statistics appropriately when there are some recombination classes which have only a small number of observations. Finally, comparisons are made between the chi-square model and some other tractable models in the literature.

INTRODUCTION

Crossover interference has been observed in almost all organisms studied, while there is little consistent evidence of chromatid interference even within the same organism (H. ZHAO, M. S. MCPEEK, and T. P. SPEED, in press). In what follows, we assume no chromatid interference.

Information on the distribution of crossovers along a chromosome generally comes from genetic experiments in which only recombinations, not crossovers, can actually be observed. In some organisms, such as Drosophila, the results of such experiments are in the form of *single spore data*, in which the products of a single meiosis are recovered separately. Other organisms, such as yeast, yield *tetrad data*, in which all four meiotic products are recovered together. It is easy to see that there are 2^n distinct recombination patterns for single spore data involving $n+1$ markers. For tetrad data involving $n+1$ markers, $n > 1$, there are more than 3^n distinguishable tetrad patterns, but under the assumption of no chromatid interference, there are only 3^n different probabilities among these patterns, i.e. some distinct tetrad patterns have the same probability of being observed. Each different probability corresponds to one of the types $(i_1 i_2 \ldots i_n)$, where $i_j = 0, 1, 2$ corresponds to parental ditype, tetratype and nonparental ditype, respectively, between $\mathcal{A}_j$ and $\mathcal{A}_{j+1}$.

Both single spore data and tetrad data record recombination events among a set of markers. As the underlying crossovers occurring during meiosis are not directly observable from the data, any model about interference must relate the observable recombination or tetrad patterns to the underlying unobservable crossover events. Crossing-over occurs among four strands after each homologous chromosome has duplicated. A model relating crossovers to recombination should specify the distribution of crossover points along the bundle of four chromatids, and the choice of non-sister chromatids to be involved in each crossover.

The chi-square model for crossovers has a long history. M. S. McPeek and T. P. Speed (in press) briefly review the history, and fit a more general class of models, renewal processes with gamma interarrivals, which includes the class of chi-square models, to Drosophila data by maximum likelihood using a Monte Carlo method. While it has generally been of interest due to its mathematical tractability, the chi-square model has also been suggested as a plausible a biological model by Foss *et al.* (1993), motivated by observations from experiments on gene conversion. There the model is represented in the form $Cx(Co)^m$, as follows: assume that crossover intermediates (C events) are randomly distributed along the four-strand bundle, and every C event will either resolve in a crossover (Cx) or not (Co). When a C resolves as a Cx, the next m C's must resolve as Co events, and after m Co's the next C must resolve as a Cx, i.e., the C's resolve in a sequence $\cdots Cx(Co)^m Cx(Co)^m \cdots$. To make the process stationary, given a set of C events, the leftmost C has an equal chance to be one of $Cx(Co)^m$. In their paper, Foss *et al.* estimate the parameter m in $Cx(Co)^m$ from the observed ratio of Co to Cx. Here we perform a full maximum likelihood estimation procedure to estimate m and genetic distances between markers from both kinds of recombination data.

ESTIMATION UNDER THE CHI-SQUARE MODEL

Given a set of markers $\mathcal{A}_1, \ldots, \mathcal{A}_{n+1}$ along a chromosome, under the chi-square model $Cx(Co)^m$, $n+1$ parameters need to be specified, namely, m and the genetic distances between each consecutive pair of markers, $x_1, x_2, \ldots, x_n$, so that the probability of each single spore or tetrad recombination pattern can be calculated. Suppose these parameters are given, let $p = m+1$, $y_j = 2px_j$ and let $\mathbf{D_k}(y)$ be the matrix whose i,jth entry is $e^{-y}y^{pk+j-i}/(pk+j-i)!$. Then the probability of k_j crossovers between $\mathcal{A}_j$ and $\mathcal{A}_{j+1}$,

$j = 1, \ldots n$, is (see Appendix A.1):

$$\frac{1}{p}\mathbf{1}\mathbf{D_{k_1}}(y_1)\mathbf{D_{k_2}}(y_2)\ldots\mathbf{D_{k_n}}(y_n)\mathbf{1}', \qquad \text{where} \quad \mathbf{1} = (1, 1, \ldots, 1).$$

Note that when $p = 1$, the above expression reduces to the Poisson case, i.e. the no interference model of HALDANE (1919). Using the above formula, we can calculate the probability of any single spore or tetrad recombination pattern $(i_1 i_2 \ldots i_n)$. We consider the two cases separately.

For single spore data, given two consecutive markers $\mathcal{A}_j$ and $\mathcal{A}_{j+1}$, we can observe a recombination or non-recombination between them. If no crossovers occur between $\mathcal{A}_j$ and $\mathcal{A}_{j+1}$, no strand in the bundle will show any recombination between these markers. MATHER (1935) proved that under the assumption of no chromatid interference, if there are $k \geq 1$ crossovers between two markers, then the probability that these two markers recombine on any given single strand is 1/2. Recall that, for single spore data, any recombination pattern can be represented as $(i_1 i_2 \ldots i_n)$, where $i_j = 0$ or 1. Define

$$\begin{aligned} \mathbf{N_j} &= \mathbf{D_0}(y_j) + \tfrac{1}{2}\textstyle\sum_{s \geq 1} \mathbf{D_s}(y_j) \\ \mathbf{R_j} &= \tfrac{1}{2}\textstyle\sum_{s \geq 1} \mathbf{D_s}(y_j). \end{aligned}$$

Then the probability of recombination pattern $(i_1 i_2 \ldots i_n)$ is (see Appendix A.2):

$$P(i_1 i_2 \ldots i_n) = \frac{1}{p}\mathbf{1}\mathbf{M_1}\mathbf{M_2}\cdots\mathbf{M_n}\mathbf{1}',$$

where $\mathbf{M_j} = \mathbf{N_j}$ when $i_j = 0$, and $\mathbf{M_j} = \mathbf{R_j}$ when $i_j = 1$.

For tetrad data, recall that there are three different possible tetrad patterns between two markers. We let p_0, p_1 and p_2 denote the probabilities of parental ditype, tetratype and non-parental ditype respectively, between a fixed pair of markers. Given $k \geq 1$ crossovers between two loci, under the assumption of no chromatid interference, the conditional probabilities $p_0^{(k)}, p_1^{(k)}$ and $p_2^{(k)}$ of a tetrad being of parental ditype, tetratype and non-parental ditype, respectively, are given by MATHER (1935):

$$
\begin{aligned}
p_0^{(k)} &= \tfrac{1}{3}(\tfrac{1}{2} + (-\tfrac{1}{2})^k) \\
p_1^{(k)} &= \tfrac{2}{3}(1 - (-\tfrac{1}{2})^k) \\
p_2^{(k)} &= \tfrac{1}{3}(\tfrac{1}{2} + (-\tfrac{1}{2})^k).
\end{aligned}
$$

We can calculate the probability of any tetrad pattern $(i_1 i_2 \ldots i_n)$, where $i_j = 0$, 1 or 2. Define

$$
\begin{aligned}
\mathbf{P_j} &= \mathbf{D_0}(y_j) & &+ \textstyle\sum_{s\geq 2} \tfrac{1}{3}(\tfrac{1}{2} + (-\tfrac{1}{2})^k)\mathbf{D_s}(y_j) \\
\mathbf{T_j} &= & \mathbf{D_1}(y_j) &+ \textstyle\sum_{s\geq 2} \tfrac{2}{3}(1 - (-\tfrac{1}{2})^k)\mathbf{D_s}(y_j) \\
\mathbf{N_j} &= & & \textstyle\sum_{s\geq 2} \tfrac{1}{3}(\tfrac{1}{2} + (-\tfrac{1}{2})^k)\mathbf{D_s}(y_j).
\end{aligned}
$$

Then the probability of the tetrad pattern $(i_1 i_2 \ldots i_n)$ can be written as (see Appendix A.3):

$$
P(i_1 i_2 \ldots i_n) = \frac{1}{p}\mathbf{1}\mathbf{M_1}\mathbf{M_2}\cdots\mathbf{M_n}\mathbf{1}',
$$

where $\mathbf{M_j} = \mathbf{P_j}$ if $i_j = 0, \mathbf{M_j} = \mathbf{T_j}$ if $i_j = 1$, and $\mathbf{M_j} = \mathbf{N_j}$ if $i_j = 2$.

Given a set of single spore or tetrad data, then based upon the above formulae, the likelihood of the observations, up to a constant factor, can be calculated in terms of the parameters as $\prod P(i_1 i_2 \ldots i_n)^{x_{i_1 i_2 \cdots i_n}}$, where $x_{i_1 i_2 \ldots i_n}$ is the observed frequency of single spores or tetrads with pattern $(i_1 i_2 \ldots i_n)$. The maximum likelihood estimates of the parameters are those that maximize the likelihood among all possible parameter values. The numerical method used to find the maximum likelihood estimates used in our analysis is the downhill simplex method, see Press *et al.* (1988). The standard error for each estimate is approximated using the fact that as $n \to \infty$,

$$
\sqrt{n}(\hat{\theta}_{jn} - \theta_j) \to N(0, [I(\theta)]^{-1}_{jj})
$$

where $I(\theta)$ is the Fisher information matrix.

APPLICATIONS TO VARIOUS ORGANISMS

In this section, the $Cx(Co)^m$ model is fitted to data from various organisms via the method of maximum likelihood. Data are of tetrad form except *Drosophila melanogaster* and human recombination data which are of single spore type.

Drosophila melanogaster: Many valuable recombination datasets for this organism have appeared in the literature since it was first studied by geneticists early in this century. Among these, two large, well-known datasets, namely WEINSTEIN (1935) and MORGAN, BRIDGES and SCHULTZ (1935), have drawn much attention and have frequently been used as a basis upon which to compare different models.

Seven loci that cover most of the X-chromosome of *D. melanogaster* were used in WEINSTEIN's study. A total of 28,239 offspring genotypes were determined. Among the $Cx(Co)^m$ models, $Cx(Co)^4$ fits the data best, i.e. the largest likelihood is achieved when the $Cx(Co)^4$ model is used. The estimated genetic distances and their associated standard errors are given in Table 1. The optimal m=4, estimated here by statistical analysis, is the same as that in FOSS *et al.* (1993), where they determine m from the observed proportion of gene conversions associated with crossovers.

MORGAN, BRIDGES and SCHULTZ's dataset also contains markers on the X-chromosome of Drosophila. There are 16,136 observations on nine loci in this dataset, including six of the same loci as in WEINSTEIN (1935). The $Cx(Co)^4$ model again gives the best fit to the data among the $Cx(Co)^m$ class. The results are given in Table 2. M. S. MCPEEK and T. P. SPEED (in press) fit a broader class of models, in which m is allowed to be non-integer, to the MORGAN, BRIDGES and SCHULTZ dataset, and estimated $m = 3.94$, which agrees very well with the integer value $m = 4$.

Among the loci which appear in both datasets, the genetic distances estimated from the two different datasets appear rather similar, yet the differences are large compared to the standard errors. This difference probably reflects non-homogeneity across Drosophila individuals. It is well known that recombination values are different for different individuals, and can be affected by factors such as temperature (PERKINS (1962)). But in the $Cx(Co)^m$ model, we assume that the crossover process follows the same distribution across the whole population. Thus, it is not surprising that we underestimate the variation in genetic distances.

Neurospora crassa: PERKINS (1962) contains data involving six markers on the right arm of Linkage group I in *Neurospora crassa.* Perkins' data were gathered from six different experiments. This set of data was previously analyzed by COBBS (1978) and RISCH and LANGE (1983). In his paper, Perkins observes that there are significant differences between recombination values for offspring from different parents and from the same set of parents when the temperature is varied. We estimated the interference parameter m for each experiment separately and also for the data when all six experiments are combined. In all cases, the best model is $Cx(Co)^2$, which has less crossover interference than $Cx(Co)^4$. This suggests crossover interference is weaker in Neurospora than in Drosophila. The results are given in Table 3.

There is another large multilocus Neurospora dataset in STRICKLAND (1961). He accumulated data from four experiments involving four markers at the end of the right arm of linkage group V. A total of 10,269 completely analyzable asci were recovered. We fit the $Cx(Co)^m$ model to the data from each of the four experiments separately, and the $Cx(Co)^2$ model gives the highest likelihood in all cases. The results are summarized in Table 4. As in the case of the Drosophila data, some of the estimated genetic distances are significantly different from one experiment to another.

Bole-gowda, Perkins and Strickland (1962) consider seven markers on linkage group I of *Neurospora crassa*, three on the left arm and four on the right arm. Altogether 2,920 offsprings were observed. When all markers are used in the analysis, the best model turns out to be Cx, i.e. m is estimated as zero. This estimate is inconsistent with the estimates from the data of Perkins (1962) and Strickland (1961). This discrepancy might be due to the fact that no positive interference was observed across the centromere, and in that case, the chi-square model may not be applicable to data which span the centromere.

The estimated m=2 for Neurospora is consistent with the observation of the ratio of gene conversions to crossovers as described in Foss *et al.* (1993). Moreover, from the fact that $Cx(Co)^2$ is the best $Cx(Co)^m$ model for data from both linkage groups I and V, we might suspect that the degree of interference is similar within the entire Neurospora genome, but with no interference across the centromere.

Saccharomyces cerevisiae: There are abundant two-point cross data for *Saccharomyces cerevisiae*, but, perhaps due to the high frequency of gene conversion, published multilocus tetrad data are rare. We analyze two-point cross data from a series of papers by Mortimer and Hawthorne (1960,1966,1968,1973). They were analyzed by Snow (1979) using the model proposed by Barratt *et al.* (1954). In Barratt's model, there is a parameter k which measures the degree of crossover interference, similar to the role m plays in the $Cx(Co)^m$ model. (For a full description, see Barratt *et al.* 1954.) $k = 1$ implies no interference, while $k > 1$ and $k < 1$ correspond to negative and positive interference respectively. We say that there is positive (negative) interference if the probability of double recombinations in two intervals is less (bigger) than the product of the probabilities of recombination in each interval, i.e. interference is defined through a quantity called S_3 by Foss *et al.* (1993). Snow fits Barratt *et al.*'s model to tetrad data involving thirty-four pairs of markers on twelve chromosomes in *Saccharomyces cerevisiae*. Snow's results, along with our estimated optimal m, genetic distances and associated standard errors from the $Cx(Co)^m$ model are given in Table 5.

Professor J. Haber kindly provided us with a multilocus *Saccharomyces cerevisiae* dataset involving the five markers, *met13, cyh2, trp5, cyh3* and *leu1* on chromosome VII. The markers *met13, cyh2, trp5*, and *leu1* were used in three of the fourteen experiments, and *met13, trp5, cyh3* and *leu1* were used in the other eleven experiments. We grouped the data across the experiments having the same set of markers, and fitted chi-square models. $Cx(Co)^6$ gives the best fit to the data from crosses involving *met13, cyh2, trp5, cyh3* and *leu1*, however the best model for the other group is $CxCo$. The estimated P-values are 0.48 and 0.89 respectively.

Genetic experiments (Foss *et al.* 1993) have shown that in *Saccharomyces cerevisiae* the ratio of gene conversions to crossovers is about 2, so we might expect the model $Cx(Co)^2$ to fit best. ¿From Table 5 and results for Haber's data, we can see that unlike Drosophila and Neurospora, where the optimal m does not change from one experiment to another, in Saccharomyces, the optimal m varies for different pairs of genes even within the same chromosome. With such a small sample size (usually around 200), it may be that there is simply not sufficient information to clearly distinguish between different $Cx(Co)^m$ models. In these Saccharomyces crosses, the differences between the likelihoods under different $Cx(Co)^m$ models are usually small. For example, for the second group 2 in Haber's data, the $-\log(\text{likelihood})$'s are rather close for different m's: they are 125.6, 125.3, 125.1, 125.2 and 125.3 for m is 4, 5, 6, 7 and 8 respectively. A high rate of conversions might also create a problem here. Instead of looking for the optimal m, we could take the $Cx(Co)^2$ model as our hypothesis, and test if it is consistent with the data we have.

Schizosaccharomyces pombe: We analyze data from two different sources: those analyzed by SNOW (1979) which were from KOHLI *et al.* (1977), and those provided by Dr. P. MUNZ (personal communication). Two-point cross tetrad data on thirty pairs of markers from all three *S. pombe* chromosomes were analyzed by SNOW. Here we fit the $Cx(Co)^m$ model to all the data analyzed by SNOW. The results are given in Table 6. Overall, among the class of $Cx(Co)^m$ models, the best fitting model is the Cx model (where m=0), which is equivalent to the no-interference model of HALDANE (1919). This suggests that there is no positive crossover interference in *S. pombe*. Recall that in BARRATT's model, $k > 1$ corresponds to negative interference so the fact that the estimated k in BARRATT's model is sometimes > 1 suggests that there may be negative crossover interference in *S. pombe*. Among the $Cx(Co)^m$ models, slight negative interference may occur at large distances, but at near distances, interference cannot be negative. However, the class of $Cx(Co)^m$ models can be extended to allow for negative interference. Instead of assuming that the distance between two crossovers is a χ^2 distribution, one may assume it is a gamma distribution of which the χ^2 is a special case. It is proved in KARLIN and LIBERMAN (1983) that when the gamma shape parameter is less than 1, the model thus proposed has negative interference. However there is no explicit expression for any single spore or tetrad recombination pattern unless the shape parameter is an integer. One way to overcome this difficulty would be to use the simulation method which is described by M. S. MCPEEK and T. P. SPEED (in press).

A multilocus *S. pombe* dataset was kindly provided by Dr. P. MUNZ, who used seven markers(*ura, his, tps, h, leu, ade, lys*) in his experiment with sample size 458. As for the two-point cross data mentioned above, the Cx model, i.e. the no-interference model, fits the data best. The estimated P-value is 0.23. All the data suggest that there is no positive crossover interference in *S. pombe*, although there may be negative interference.

Aspergillus nidulans: STRICKLAND (1958) published 1,231 fully classifiable asci of *As-*

pergillus nidulans from three separate experiments. Crosses number one and three cover the same three intervals in the right arm of the *BI* chromosome, while cross two covers six intervals (the same three in the right arm, two in the left arm and the sixth spanning the centromere). The data from the first two experiments are fitted by the chi-square model. As in the case of *S. pombe*, the Cx model fits the data best. There appears to be no positive crossover interference present in this organism.

Humans: For humans, there are not yet available data of the quality and large sample size one finds for experimental organisms. We have analyzed data from McInnis *et al.* (1993) consisting of the number of crossovers inferred over a region of approximately 60 cM in each of 664 meioses. We estimate $m = 2$ overall, but when males and females are considered separately, we estimate $m = 1$ for females and $m = 4$ for males. The results when male and female are assumed to have the same interference parameter m are given in Table 7.

THE INTERPRETATION OF GOODNESS-OF-FIT STATISTICS

As multiple recombination tends to be rare, there are many possible multilocus recombination events which each occur only a small number of times in the datasets considered here. As a result, the asymptotic χ^2 distribution of a test statistic such as Pearson's Chi-square statistic or the likelihood ratio statistic may not be a good approximation to the actual distribution. There are two ways to get around with this difficulty: by using a Monte Carlo method to approximate the distribution, or by grouping some classes with small expected counts together to form a bigger class. Suppose the sample size is N, the model is of the form $f(\theta_1, \ldots, \theta_k)$, and the test statistic for the sample is T. A Monte Carlo approximation can be carried out as follows: (1) Estimate parameters $\hat{\theta}_1, \ldots, \hat{\theta}_k$ in the model by the method of maximum likelihood, (2) generate N observations according to the distribution $f(\hat{\theta}_1, \ldots, \hat{\theta}_k)$, and calculate the test statistic t_i for this sample; (3) repeat step (2) M times, then the P-value of T can be estimated as the proportion of times when t_i is bigger than T.

We did a simulation study to explore which test statistic should be used and whether grouping small cells gives a more powerful test when both the hypothesis and alternative are of the $Cx(Co)^m$ form. The simulation is carried out as follows: suppose there are 5 equally spaced markers on the chromosome, with genetic distance 10cM between each consecutive pair of markers. The null hypothesis is that the crossover process follows the $Cx(Co)^m$ model; the alternative is that crossover process follows the $Cx(Co)^k$ model, $k \neq m$. A sample of size 500 is generated from the alternative model. The hypothesis model is used to fit the simulated data, and the P-value of the test statistic is calculated by the Monte Carlo approximation as described above. Two thousand such samples are generated and the P-values are calculated. For a level-α test, we reject the null hypothesis if the calculated P-value is less than α. The power of the test thus can be approximated by the percentage of times that the null hypothesis is rejected. For each dataset, several test statistics from the so-called *power divergence family* (Read and Cressie 1988) are used. The test statistics in this family have the form

$$\frac{2}{\lambda(\lambda+1)}\sum_{i=1}^{n} X_i[(\frac{X_i}{E_i})^{\lambda} - 1].$$

This family includes several well-known test statistics. For example, $\lambda = 1$ gives Pearson's Chi-square statistic, and $\lambda = 0$ gives the log-likelihood ratio test statistic. For each sample, each test is applied on both ungrouped and grouped classes. Grouping is done in such a way that all offspring types with more than three recombinations are put together to form a larger class while the other types are kept separate. Two pairs of null and alternative hypotheses are considered. In both cases, the null hypothesis is set to be the model $Cx(Co)^2$. The alternative is $Cx(Co)^3$ in the first pair and $CxCo$ in the second pair. The results are summarized in Table 8 and Table 9.

¿From Table 8 and Table 9, we can see that on average, tests based on grouped data are more powerful than those based on the original classes, and the power varies more among different test statistics when the ungrouped classes are used. For tests based on grouped classes, there are no big differences between the power of different test statistics. Actually, when the data are grouped, χ^2 is a good approximation to the real distribution of all test statistics, so they are very similar to each other. We think the reason is that large classes are more informative than small classes. Tests of goodness of fit should give more weight to these larger classes.

COMPARISON WITH GOLDGAR & FAIN'S MODEL

A thorough comparison of different models is made by M. S. McPeek and T. P. Speed (in press). In this section, we focus on the comparison of the $Cx(Co)^m$ model with a model proposed by Goldgar and Fain (1988). In that model, which is similar to the count-location model (Karlin and Liberman (1979) and Risch and Lange (1979)), Goldgar and Fain (1988) assume that the number of crossovers follows a distribution which have to be estimated from the data. Their model differs from the count-location model in two respects: (1) given the number of crossovers, their locations are not independent, but they follow a specified joint distribution in which some parameters have to be estimated; (2) instead of putting the distribution on the four-strand bundle, these distributions are put on the single meiotic product, i.e. it is a two-strand model. In fact, it is not possible to construct a four-strand no chromatid interference (NCI) model which is consistent with their model on a single meiotic product, see D. Goldstein, H. Zhao and T. P. Speed (unpublished manuscript). In their paper, Goldgar and Fain show that their model fits data much better than the count-location model and several two-strand models based on map functions. Estimates of genetic distances among markers from Weinstein's data and Morgan's data by their model (Goldgar and Fain 1988), as well as the estimates based on the $Cx(Co)^m$ model are given in Table 10 and Table 11.

Besides genetic distances, the parameters used by GOLDGAR and FAIN are: d_i, i=0, 1, 2 and 3, the probabilities of 0, 1, 2 and 3 crossovers respectively; k, which measures the degree of interference, and x_0, the genetic distance between the centromere and the marker closest to it. So when $n+1$ markers are involved in the experiment, a total of $n+5$ parameters have to be estimated. On the other hand, for the $Cx(Co)^m$ model, $n+1$ parameters are used, including n genetic distances between each pair of consecutive markers and the parameter m which measures interference. Thus, in general, four fewer parameters are needed for the $Cx(Co)^m$ model than for GOLDGAR and FAIN's model. For some organisms, it is reasonable to assume there are no more than three crossovers. For example, among 28,239 offspring in MORGAN's *Drosophila melanogaster* data, only two offspring showed recombination in four intervals at the same time, and no such individual was recorded in WEINSTEIN's data. On the other hand, for those organisms which have a large number of crossovers during meiosis, e.g. *S. pombe*, probabilities of 4, 5 or even more crossovers on the four-strand bundle must be estimated when GOLDGAR and FAIN's model is used. One should also specify the joint distribution of these crossover locations. In this case, the model loses its simplicity and credibility when many joint distributions must be assumed based on empirical observations.

A good model should both fit the data and be biologically reasonable. Recall that crossovers occur among four chromatid strands, so under the assumption of no chromatid interference, we can relate the probabilities of crossover patterns on the four-strand bundle and those on a single strand. Under GOLDGAR and FAIN's model, when the probabilities of crossover patterns on a single strand are specified, some crossover patterns on the four strand bundle will have negative probabilities under the assumption of no chromatid interference. Thus, the model they describe is incompatible with the assumption of no chromatid interference. We tried a variation of GOLDGAR and FAIN's model in which we put the distribution on the four-strand bundle instead of on a single strand. Under the assumption of no chromatid interference, we derived the probability of each recombination pattern for single spore data. This slightly modified version of GOLDGAR and FAIN's model fits the Drosophila data as well as the original one.

Coincidence curves: The traditional measure of interference is coincidence (MULLER 1916; STURTEVANT 1915), which is expressed as a ratio. The numerator is the chance of simultaneous recombination across both of two disjoint intervals on the chromosome. The denominator is the product of the marginal probabilities of recombination across the intervals.

$$S = \frac{r_{11}}{(r_{10} + r_{11})(r_{01} + r_{11})} ,$$

where S is the coincidence and r_{ij} is the chance of i recombinations across the first interval and j recombinations across the second interval. The coincidence curve for a model is a plot of the coincidence against the genetic distance between two intervals, where the widths of the two intervals are taken to be infinitesimal. (FOSS *et al.* call this quantity S_4.) FOSS *et al.* (1993) compare the coincidence curves (S) for the $Cx(Co)^m$ model with empirical coincidence curves estimated from data. They find that the theoretical curves are very close to the empirical ones. Similarly we draw the S curves based on the modified version of GOLDGAR and FAIN's model (Fig 1). In GOLDGAR and FAIN's model, S is not only a function of the genetic distance between the two regions under study, but also depends on how far these regions are from the centromere, so the S curve cannot be uniquely drawn on the graph. Instead, for a given genetic distance between two regions, S can vary according to the distance to the centromere. It is clear from the graph that GOLDGAR and FAIN's model predicts that S will be greater than 1.5 when the distance between the two regions is bigger than 60cM, no matter where they are located on the chromosome. But this prediction is not consistent with the empirical results in which S is always smaller than 1.2.

DISCUSSION

Based on the derived probabilities for each single spore or tetrad recombination pattern, we use the method of maximum likelihood to fit the $Cx(Co)^m$ model to a variety of organisms. The estimated m's based on statistical analyses of *Drosophila melanogaster* and *Neurospora crassa* data agree with those given by FOSS *et al.* (1993), where m was estimated by the ratio of gene conversions to crossovers in genetic experiments. In humans, we are currently able to provide only a preliminary estimate of m, although we hope that more extensive human datasets will soon become available. While some amount of positive interference is shown in the above organisms, it is not present in two other organisms we analyzed, *Schizosaccharomyces pombe* and *Aspergillus nidulans*.

The m estimated from different experiments using genes on different chromosomes within the same organism turn out to be rather similar. This implies that as a measure of interference, m does not change across different chromosomes, thus, the degree of interference might be determined by factors specific to each organism.

We have discussed how to interpret the goodness-of-fit test statistic appropriately after fitting a model to a multilocus dataset. In multilocus data, since many single spore or tetrad recombination patterns have rather small expected numbers of observations, even in an experiment of moderate size, the χ^2 approximation to goodness-of-fit statistics often fails. Two ways of avoiding this difficulty are proposed: 1) simulating the distribution of the test statistic by a Monte Carlo method, or 2) grouping small classes into a larger class. Our simulation study shows that the tests based on grouped classes usually have larger power than the tests based on ungrouped classes.

Among the four-strand models considered here, the four-stranded version of the model of GOLDGAR and FAIN (1988) gave the smallest likelihood ratio statistics (see Table 10 and Table 11), but considering that this model has four more parameters than the chi-square model, the difference in likelihood ratio statistics is not impressive. The count-location model has two parameters more than the chi-square model, yet it performed worse. In comparing the $Cx(Co)^m$ model with the count-location model and GOLDGAR and FAIN's model, we consider that a good model should not only yield a small test statistic, but should also be parsimonious (i.e. have few parameters), be biologically reasonable and be generalizable to many organisms. In these respects the $Cx(Co)^m$ model seems superior. It has a biological basis and is computationally tractable. Due to its simple structure, it can be applied to a broad range of organisms.

Although the $Cx(Co)^m$ model proposed in this paper applies well to data of different organisms, and gives some insight into the underlying crossover process, there is a lot room left for improvement. First of all, the parameter m need not be restricted to be an integer, but when m is not an integer, there are no explicit expressions for the probabilities of single spore or tetrad recombination patterns. In this case, M. S. MCPEEK and T. P. SPEED (in press) use a simulation method to estimate the parameters. Second, we might suspect that the amount of interference varies in different regions within the same chromosome. A local m rather than a global m might be fitted in the model. Finally, for some organisms with a high proportion of conversion data observed, we need to develop a model to include both gene conversions and crossovers. A good model of this kind should help us understand more about the crossing-over process.

The no-interference model is widely used in human genome mapping. Although it has been shown by SPEED, MCPEEK and EVANS (1992) that the no-interference model is asymptotically robust for gene ordering, we do lose some efficiency in ordering and in excluding a test locus when there is interference in the underlying crossover process. D. GOLDSTEIN, H. ZHAO and T. P. SPEED (unpublished manuscript) study the loss in efficiency using the no-interference model when the actual crossover process follows the

$Cx(Co)^m$ model. They find that the number of gametes required for these tasks is 10% - 50% smaller for the $Cx(Co)^m$ model than for the no interference model, depending on the degree of interference and the distances between the markers.

ACKNOWLEDGEMENT

We would like to thank Professor James Haber and Dr. Peter Munz for kindly supplying the data. This work was supported by National Science Foundation grant DMS-9113527.

LITERATURE CITED

BARRATT, R. W., D. NEWMEYER, D. D. PERKINS, AND L. GARNJOBST, 1954 Map construction in *Neurospora crassa*. Advan. Genet. **6:** 1-93.

BOLE-GOWDA, B. N., D. D. PERKINS, and W. N. STRICKLAND, 1962 Crossing-over and interference in the centromere region of linkage group I of Neurospora. Genetics **47:** 1243-1252.

COBBS, G., 1978 Renewal process approach to the theory of genetic linkage: case of no chromatid interference. Genetics **89:** 563-581.

FOSS, E., R. LANDE, F. W. STAHL, and C. M. STEINBERG, 1993 Chiasma interference as a function of genetic distance. Genetics **133:** 681-691.

GOLDGAR, D. E., and P. R. FAIN, 1988 Models of multilocus recombination: nonrandomness in chiasma number and crossover positions. Am. J. Hum. Genet. **43:** 38-43.

HALDANE, J. B. S., 1919 The combination of linkage values, and the calculation of distances between the loci of linked factors. J. Genetics **8:** 299-309.

HAWTHORNE, D. C., and R. K. MORTIMER, 1960 Chromosome mapping in Saccharomyces: centromere-linked genes. Genetics **45:** 1085-1110.

HAWTHORNE, D. C., and R. K. MORTIMER, 1968 Genetic mapping of nonsense suppressors in yeast. Genetics **60:** 735-742.

KARLIN, S., and U. LIBERMAN, 1979 A natural class of multilocus recombination processes and related measure of crossover interference. Adv. Appl. Prob. **11:** 479-501.

KARLIN, S., and U. LIBERMAN, 1983 Measuring interference in the chiasma renewal formation process. Adv. Appl. Prob. **15:** 471-487.

KOHLI, J., H. HOTTINGER, P. MUNZ, A. STRAUSS, and P. THURIAUX, 1977 Genetic mapping in *Schizosaccharomyces pombe* by mitotic and meiotic analysis and induced haploidization. Genetics **87:** 471-489.

MATHER, K., 1935 Reduction and equational separation of the chromosomes in bivalents and multivalents. J. Genet. **30:** 53-78.

McInnis, M. G., A. Chakravarti, J. Blaschak, M. B. Petersen, V. Sharma, D. Avramopoulos, J-L. Blouin, U. Konig, C. Brahe, T. C. Matise, A. C. Warren, C. C. Talbot Jr, C. Van Broeckhoven, M. Litt, and S. E. Antonarakis, 1993 A linkage map of human chromosome 21: 43 PCR markers at average intervals of 2.5 cM. Genomics **16:** 562-571.

Morgan, T. H., C. B. Bridges, and J. Schultz, 1935 Report of investigations on the constitution of the germinal material in relation to heredity. Carnegie Institute of Washington **34:** 284-291.

Mortimer, R. K., and D. C. Hawthorne, 1966 Genetic Mapping in Saccharomyces. Genetics **53:** 165-173.

Mortimer, R. K., and D. C. Hawthorne, 1973 Genetic mapping in Saccharomyces IV. Mapping of temperature-sensitive genes and use of disomic strains in localizing genes. Genetics **74:** 33-54.

Muller, H. J., 1916 The mechanism of crossing-over. The American Naturalist **50:** 193-434.

Perkins, D. D., 1962 Crossing-over and interference in a multiply marked chromosome arm of Neurospora. Genetics **47:** 1253-1274.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, 1988 *Numerical Recipes in C.* Cambridge University Press, Cambridge.

Read, R. C., and N. A. C. Cressie, 1988 *Goodness-of-fit Statistics for Discrete Multivariate Data.* Springer-Verlag, New York.

Risch, N., and K. Lange, 1979 An alternative model of recombination and interference. Annals of Human Genetics **43:** 61-7 0.

Risch, N., and K. Lange, 1983 Statistical analysis of multilocus recombination. Biometrics **39:** 949-963.

Snow, R., 1979 Maximum likelihood estimation of linkage and interference from tetrad data. Genetics **92:** 231-245.

SPEED, T. P., M. S. MCPEEK, and S. N. EVANS, 1992 Robustness of the no-interference model for ordering genetic markers. Proc. Natl. Acad. Sci. USA **89:** 3103-3106.

STRICKLAND, W. N., 1958 An analysis of interference in *Aspergillus nidulans*. Proc. Roy. Soc., B **149:** 82-101.

STRICKLAND, W. N., 1961 Tetrad analysis of short chromosome regions of *Neurospora crassa*. Genetics **46:** 1125-1141.

STURTEVANT, A. H., 1915 The behavior of the chromosomes as studied through linkage. Zeit. f. ind. Abst. u. Vererb. **13:** 234-287.

WEINSTEIN, A., 1935 The theory of multiple strand crossing over. Genetics **21:** 155-199.

APPENDIX

A.1 Lemma: Under the $Cx(Co)^m$ model, the probability of k_j crossovers between $\mathcal{A}_j$ and $\mathcal{A}_{j+1}$, $j = 1, \ldots, n$ is

$$\frac{1}{p}\mathbf{1}\mathbf{D}_{\mathbf{k_1}}(y_1)\mathbf{D}_{\mathbf{k_2}}(y_2)\ldots\mathbf{D}_{\mathbf{k_n}}(y_n)\mathbf{1}'.$$

where $p = m + 1$, $y_j = 2px_j$,and $\mathbf{D}_\mathbf{k}(y)$ has i,jth entry $e^{-y}y^{pk+j-i}/(pk + j - i)!$.

Proof: We start with the simplest case when there are only two markers. Since the C events are randomly distributed on the four-strand bundle, and the number of C's follows the Poisson distribution with parameter, say y, the chance of s C's is $e^{-y}y^s/s!$. $1/p$ of these C's will resolve as crossover event, and under the assumption of no chromatid interference, each strand has chance 1/2 of being involved in each crossover. So on average, each strand has $s/2p$ crossovers, given s C events. Recall that the genetic distance is defined to be the expected number of crossovers on a single strand, so the genetic distance x and the Poisson parameter x are related by $x = y/2p$, i.e. $y = 2px$.

In the following discussion, suppose markers $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$ are laid out from left to right, and the C events occur also from left to right. The $Cx(Co)^m$ model assumes that the crossover intermediate (C) events resolve in sequence like $CxCoCo\ldots CoCxCo\ldots$ and the process is stationary, so the first C event to the right of $\mathcal{A}_1$ has an equal chance of resolving as any of the $m+1$ elements of $Cx(Co)^m$. The occurrence of k crossovers between $\mathcal{A}_1$ and $\mathcal{A}_2$ might be the result of p^2 possible situations, depending on the number of Co's before the first Cx to right of $\mathcal{A}_1$ and the number of Co's between $\mathcal{A}_2$ and the nearest Cx left to it. The number can vary from 0 to $p-1$. Therefore, the chance of k_1 Cx's between $\mathcal{A}_1$ and $\mathcal{A}_2$ can be computed as:

$$\frac{e^{-y_1}}{p}\sum_{i=1}^{p}\sum_{j=0}^{p-1}\frac{y_1^{pk_1-p+i+j}}{(pk_1-p+i+j)!}$$

The case $i=1$ corresponds to the situation where the leftmost C between $\mathcal{A}_1$ and $\mathcal{A}_2$ is a Cx, and the rightmost C could be either one Cx (pk_1-p+1 C's altogether between $\mathcal{A}_1$ and $\mathcal{A}_2$), or the first Co after a Cx (pk_1-p+2 C's), or the second Co after a Cx, etc. $i-1$ corresponds to number of Co's between $\mathcal{A}_1$ and the first Cx.

We can write the sum in a matrix product form:

$$\frac{1}{p}(1\ \ 1\ \ \ldots\ \ 1)\mathbf{D}_{\mathbf{k_1}}(y_1)\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}.$$

Each element in the first column of the matrix corresponds to the last C event between $\mathcal{A}_1$ and $\mathcal{A}_2$ being a Cx, the second column corresponds to the last C being the first Co after the k_1th Cx, the jth column to the jth C after the k_1th Cx. Therefore, the sum of the jth ($j>$0) column multiplied by $1/p$ is the probability that there are k_1 crossovers between $\mathcal{A}_1$ and $\mathcal{A}_2$ and the last C event is the $(j-1)$th Co after the k_1th Cx. Therefore if we define

$$(p_{k_1}^1 p_{k_1}^2 \ldots p_{k_1}^p) = \frac{1}{p}(1\ \ 1\ \ \ldots 1)\mathbf{D}_{\mathbf{k_1}}(y_1),$$

then $p_{k_1}^j$ is the probability that the last C between $\mathcal{A}_1$ and $\mathcal{A}_2$ is the (j-1)th Co after the k_1th Cx with the exception that $p_{k_1}^0$ is the probability of the last C being the k_1th Cx.

Now we consider the case for three markers $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$. Given that the first C to the right of $\mathcal{A}_2$ is the lth Co after a Cx, the probability of k_2 crossovers between $\mathcal{A}_2$ and $\mathcal{A}_3$ is

$$e^{-y_2} \sum_{i=1}^{p} \frac{y_2^{pk_2+l-i}}{(pk_2+l-i)!}.$$

The chance that there are l Co's between $\mathcal{A}_2$ and the first Cx after $\mathcal{A}_2$ is the same as the chance that the last C between $\mathcal{A}_1$ and $\mathcal{A}_2$ is the (p-l-1)th C after a Cx which is $p_{k_1}^{p-l-1}$. Therefore the chance of k_1 crossovers between $\mathcal{A}_1$, $\mathcal{A}_2$, and k_2 crossovers between $\mathcal{A}_2$ and $\mathcal{A}_3$ is

$$e^{-y_2} \sum_{i=1}^{p} p_{k_1}^{p+1-i} \sum_{j=0}^{p-1} \frac{{y_2}^{pk_2-p+i+j}}{(pk_2-p+i+j)!}$$

.

Rewriting the above relation in matrix form, we get,

$$(p_{k_1}^1 \;\; p_{k_1}^2 \;\; \ldots p_{k_1}^p)\mathbf{D_{k_2}}(y_2)\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Recall that $(p_{k_1}^1 \;\; p_{k_1}^2 \;\; \ldots p_{k_1}^p) = (11\ldots1)\mathbf{D_{k_1}}(y_1)$, thus the probability of k_1 crossovers between $\mathcal{A}_1$ and $\mathcal{A}_2$, k_2 crossovers between $\mathcal{A}_2$ and $\mathcal{A}_3$ is

$$\mathbf{1}\mathbf{D_{k_1}}(y_1)\mathbf{D_{k_2}}(y_2)\mathbf{1}'.$$

The general result involving n intervals can be proved by the same method.

A.2 Theorem 1: Define

$$\begin{aligned} \mathbf{N_j} &= \mathbf{D_0}(y_j) + \tfrac{1}{2}\textstyle\sum_{s\geq 1}\mathbf{D_s}(y_j) \\ \mathbf{R_j} &= \tfrac{1}{2}\textstyle\sum_{s\geq 1}\mathbf{D_s}(y_j), \end{aligned}$$

then the probability of recombination pattern $(i_1 i_2 \ldots i_n)$ is

$$P(i_1 i_2 \ldots i_n) = \frac{1}{p}\mathbf{1}\mathbf{M_1}\mathbf{M_2}\cdots\mathbf{M_n}\mathbf{1}'$$

where $\mathbf{M_j} = \mathbf{N_j}$ when $i_j = 0$, and $\mathbf{M_j} = \mathbf{R_j}$ when $i_j = 1$.

Proof: It is well known that given $k \geq 1$ crossovers between two markers, the chance of a recombination on a single strand is $1/2$, and there can be no recombination if no crossovers occur. We can write $p_0^{(k)}$ for the probability of no recombination and $p_1^{(k)}$ for the probability of recombination given k crossovers occurring. So $p_0^{(0)} = 1$, $p_1^{(0)} = 0$ and $p_1^{(k)} = p_0^{(k)} = 1/2$ when $k \geq 1$. Write $P_{(i_1 i_2 \ldots i_n)}^{(k_1,k_2,\ldots,k_n)}$ for the probability of observing recombination pattern $(i_1 i_2 \ldots i_n)$ when there are k_j crossovers in the jth interval, j=1,2,..., n. Then we have

$$
\begin{aligned}
& P(i_1 i_2 \ldots i_n) \\
= \; & \textstyle\sum_{k_1,k_2,\ldots,k_n} P_{(i_1 i_2 \ldots i_n)}^{(k_1,k_2,\ldots,k_n)} \\
= \; & \textstyle\sum_{k_1}\sum_{k_2}\cdots\sum_{k_n} p_{i_1}^{(k_1)} p_{i_2}^{(k_2)} \cdots p_{i_n}^{(k_n)} \frac{1}{p}\mathbf{1}\mathbf{D_{k_1}}(y_1)\mathbf{D_{k_2}}(y_2)\ldots\mathbf{D_{k_n}}(y_n)\mathbf{1}' \\
= \; & \textstyle\frac{1}{p}\mathbf{1}(\sum_{k_1}\sum_{k_2}\cdots\sum_{k_n} p_{i_1}^{(k_1)} p_{i_2}^{(k_2)} \cdots p_{i_n}^{(k_n)} \mathbf{D_{k_1}}(y_1)\mathbf{D_{k_2}}(y_2)\ldots\mathbf{D_{k_n}}(y_n))\mathbf{1}' \\
= \; & \textstyle\frac{1}{p}\mathbf{1}(\sum_{k_1} p_{i_1}^{(k_1)}\mathbf{D_{k_1}}(y_1)\sum_{k_2} p_{i_2}^{(k_2)}\mathbf{D_{k_2}}(y_2)\ldots\sum_{k_n} p_{i_n}^{(k_n)}\mathbf{D_{k_n}}(y_n))\mathbf{1}' \\
= \; & \textstyle\frac{1}{p}\mathbf{1}\mathbf{M_1}\mathbf{M_2}\cdots\mathbf{M_n}\mathbf{1}'.
\end{aligned}
$$

A.3 Theorem 2: Define

$$
\begin{aligned}
\mathbf{P_j} &= \mathbf{D_0}(y_j) && + \textstyle\sum_{s\geq 2}\frac{1}{3}(\frac{1}{2}+(-\frac{1}{2})^k)\mathbf{D_s}(y_j) \\
\mathbf{T_j} &= \mathbf{D_1}(y_j) && + \textstyle\sum_{s\geq 2}\frac{2}{3}(1-(-\frac{1}{2})^k)\mathbf{D_s}(y_j) \\
\mathbf{N_j} &= && \textstyle\sum_{s\geq 2}\frac{1}{3}(\frac{1}{2}+(-\frac{1}{2})^k)\mathbf{D_s}(y_j).
\end{aligned}
$$

Then the probability of tetrad pattern $(i_1 i_2 \ldots i_n)$ can be written as

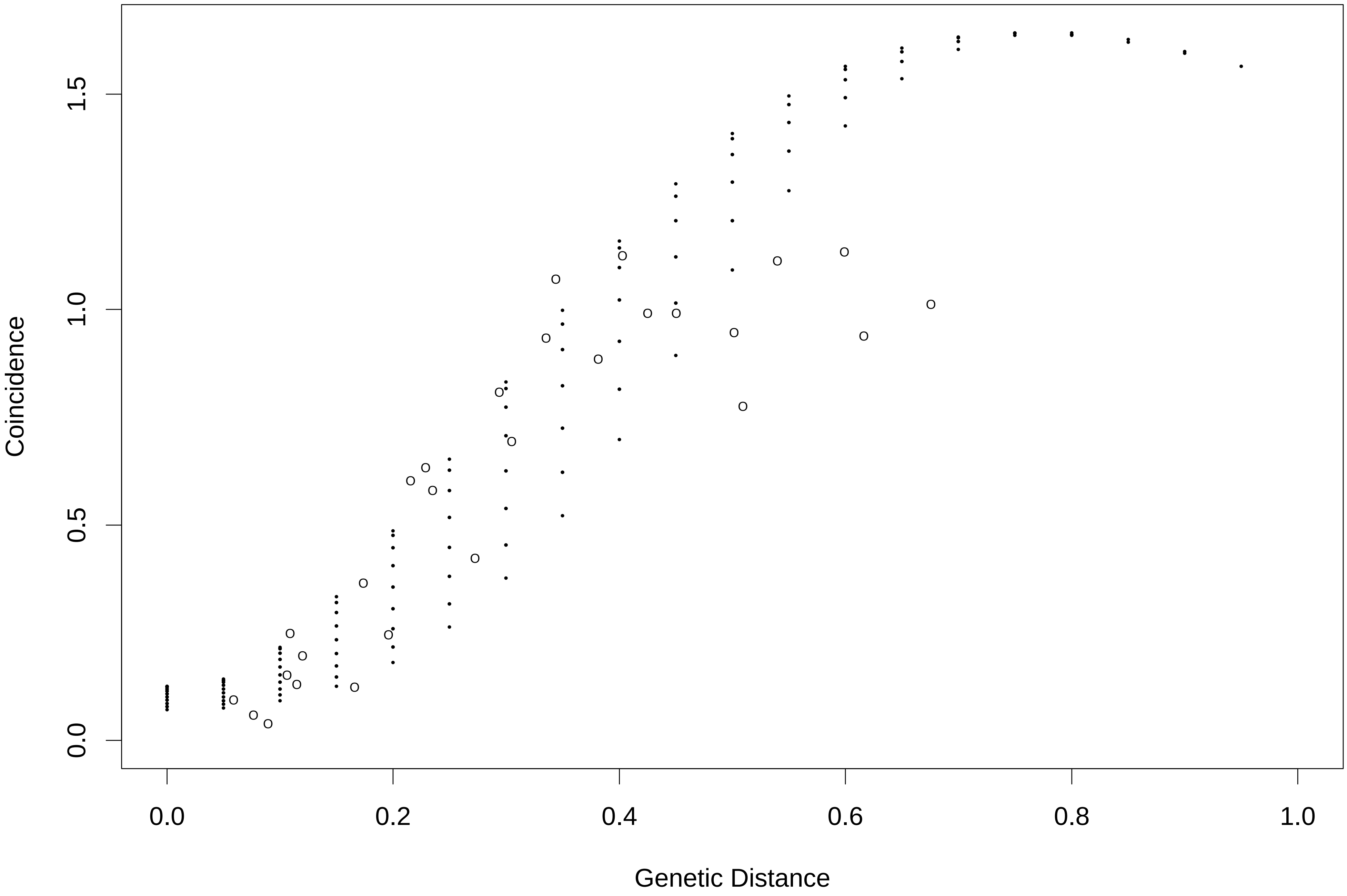
$$P(i_1 i_2 \ldots i_n) = \frac{1}{p}\mathbf{1}\mathbf{M_1}\mathbf{M_2}\cdots\mathbf{M_n}\mathbf{1}'$$

where $\mathbf{M_j} = \mathbf{P_j}$ if $i_j = 0, \mathbf{M_j} = \mathbf{T_j}$ if $i_j = 1$, and $\mathbf{M_j} = \mathbf{N_j}$ if $i_j = 2$.

Proof: Notice that given $k \geq 1$ crossovers between two markers, the probabilities of parental ditype, tetratype and nonparental ditype are $\frac{1}{3}(\frac{1}{2}+(-\frac{1}{2})^k)$, $\frac{2}{3}(1-(-\frac{1}{2})^k)$ and $\frac{1}{3}(\frac{1}{2}+(-\frac{1}{2})^k)$ respectively. Using the same method as Theorem 1, the conclusion follows.

FIGURE LEGEND

Figure 1: Comparison between predicted and observed S values for GOLDGAR and FAIN's model. The dots above any given genetic distance represent the range of possible predicted S values, which vary according to location on the chromosome. 'O' denotes the observed S value.



Coincidence
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
Genetic Distance

TABLE 1: Data of WEINSTEIN(1935): observed and expected counts under the $Cx(Co)^4$ model

Legend: $Cx(Co)^4$ gives the best fit among $Cx(Co)^m$ models. The estimated genetic distances and their standard errors between these markers are: 7.13(0.14), 9.55(0.17), 8.28(0.16), 14.75(0.20), 11.45(0.18) and 11.47(0.19) cM. In the table, 0 stands for no recombination, 1 recombination. The estimated P-value is less than 0.001.

sc-e	*e-cv*	*cv-ct*	*ct-v*	*v-g*	*g-f*	Expected	Observed
0	0	0	0	0	0	12934	12776
1	0	0	0	0	0	1266	1407
0	1	0	0	0	0	1909	2018
0	0	1	0	0	0	1831	1976
0	0	0	1	0	0	3420	3378
0	0	0	0	1	0	2454	2356
0	0	0	0	0	1	2119	2067
1	1	0	0	0	0	5	9
1	0	1	0	0	0	34	16
1	0	0	1	0	0	205	142
1	0	0	0	1	0	240	198
1	0	0	0	0	1	226	206
0	1	1	0	0	0	8	11
0	1	0	1	0	0	146	136
0	1	0	0	1	0	280	261
0	1	0	0	0	1	327	318

(continued)

sc-e	e-cv	cv-ct	ct-v	v-g	g-f	Expected	Observed
0	0	1	1	0	0	33	42
0	0	1	0	1	0	150	148
0	0	1	0	0	1	258	212
0	0	0	1	1	0	65	123
0	0	0	1	0	1	252	315
0	0	0	0	1	1	30	59
1	1	1	0	0	0	0	3
1	1	0	1	0	0	0	1
1	1	0	0	0	1	1	2
1	0	1	0	1	0	3	3
1	0	1	0	0	1	5	3
1	0	0	1	1	0	3	10
1	0	0	1	0	1	14	15
1	0	0	0	1	1	3	1
0	1	1	1	0	0	0	1
0	1	0	1	1	0	2	2
0	1	0	1	0	1	9	10
0	1	0	0	1	1	3	1
0	0	1	1	0	1	2	5
0	0	1	0	1	1	1	5
0	0	0	1	1	1	0	1
1	1	1	1	0	0	0	1
1	1	1	0	0	1	0	1

TABLE 2: Data of Morgan *et al.*(1935): estimated genetic distances with standard errors under the $Cx(Co)^4$ model

Legend: $Cx(Co)^4$ model gives the best fit among $Cx(Xo)^m$ models. The estimated P-value is less than 0.001.

Interval	*sc-e*	*e-cv*	*cv-ct*	*ct-v*	*v-s*	*s-f*	*f-ca*	*ca-b*
Distance (cM)	5.13	9.83	7.49	13.29	8.42	15.55	7.47	4.42
SE	0.17	0.23	0.20	0.26	0.21	0.28	0.20	0.16

TABLE 3: Data of PERKINS(1962): observed and expected counts under the $Cx(Co)^2$ model

Legend: $Cx(Co)^2$ gives the best fit among $Cx(Co)^m$ models. The estimated genetic distances with standard errors between these markers are: 10.66(0.59), 22.78(0.82), 11.81(0.60), 8.57(0.55) and 21.69(0.80) cM. In the table, 0 stands for parental ditype, 1 tetratype and 2 nonparental ditype. The estimated P-value is 0.69.

cr-th	*th-ni*	*ni-au*	*au-ni*	*ni-os*	Expected	Observed
0	0	0	0	0	106	103
1	0	0	0	0	57	65
0	1	0	0	0	189	201
0	0	1	0	0	109	108
0	0	0	1	0	74	52
0	0	0	0	1	141	126
1	1	0	0	0	26	19
1	0	1	0	0	32	24
1	0	0	1	0	27	32
1	0	0	0	1	61	79
0	2	0	0	0	3	5
0	1	1	0	0	41	36
0	1	0	1	0	50	40
0	1	0	0	1	152	188
0	0	2	0	0	0	1
0	0	1	1	0	8	15

(continued)

cr-th	*th-ni*	*ni-au*	*au-ni*	*ni-os*	Expected	Observed
0	0	1	0	1	52	47
0	0	0	1	1	19	22
1	1	1	0	0	4	2
1	1	0	1	0	6	8
1	1	0	0	1	20	14
1	0	1	1	0	2	5
1	0	1	0	1	15	10
1	0	0	1	1	7	6
0	1	1	1	0	2	7
0	1	1	0	1	19	18
0	1	0	1	1	13	11
0	0	1	1	1	2	2
1	1	1	1	0	0	1
1	1	1	0	1	2	1
1	0	1	1	1	0	1
0	1	1	1	1	1	3

TABLE 4: Data of STRICKLAND(1961): estimated genetic distances with standard errors and estimated P-values

Legend: $Cx(Co)^2$ is the best model among $Cx(Co)^m$ models for all four experiments.

	Experiment 1	Experiment 2	Experiment 3	Experiment4
hist1-inos	4.33(0.33)	7.07(0.40)	6.50(0.35)	5.40(0.29)
inos-bis	4.97(0.36)	5.27(0.35)	6.02(0.36)	6.23(0.32)
bis-pab2	10.1(0.47)	9.18(0.45)	10.7(0.45)	10.6(0.40)
P-value	0.50	0.02	0.02	0.13

TABLE 5: Data of *Saccharomyces cerevisiae* analyzed in SNOW (1979)

Legend: Chro: the chromosome on which each gene pair are located, m: estimated m in the $Cx(Co)^m$ model, n: sample size, x_{χ^2}: estimated genetic distance from the $Cx(Co)^m$ model, sd: standard error for the estimated genetic distance, p: estimated P-value, x_s: estimated genetic distance in SNOW (1979), k: estimated k in BARRATT *at al*'s model in SNOW(1979).

Chro	Gene pair	m	n	x_{χ^2}	se	p	x_s	k
2	*cyh1-gal1*	1	146	19.9	2.6	0.98	23.0	0.337
	gal1-lys2	2	383	52.7	3.5	0.71	79.5	0.488
	lys2-try1	3	335	34.8	1.9	0.77	50.0	0.245
	try1-his7	5	127	42.2	3.5	0.90	70.5	0.194
	SUP45-lys2	2	104	25.7	3.1	0.89	32.7	0.258
	SUP45-tyr1	0	105	17.4	3.4	0.64	17.4	1.817
3	*his4-mat1*	2	278	39.8	2.8	0.77	56.3	0.294
	his4-leu2	1	521	17.4	1.2	0.85	19.8	0.370
	leu2-mat1	1	481	35.8	2.2	0.73	44.6	0.429
	mat1-thr4	1	434	21.3	1.6	0.90	24.8	0.404
	thr4-MAL2	2	286	29.2	2.2	0.82	38.2	0.294
4	*SUP35-aro1*	5	101	46.1	7.5	0.99	80.9	0.204
	trp1-cdc2	1	205	18.2	2.1	0.92	20.8	0.284
5	*his1-trp2*	2	215	24.5	2.2	0.79	30.8	0.277
	ura3-hom3	5	206	34.7	2.3	0.99	53.2	0.115

(continued)

Chro	Gene pair	m	n	x_{χ^2}	sd	p	x_s	k
6	*SUP11-his2*	1	105	22.0	3.2	0.97	25.6	0.386
7	*trp5-ade6*	0	106	82.1	4.9	0.62	83.3	1.333
	ade5-tyr3	1	166	73.3	10.	0.87	101.	0.764
	tyr3-lys5	0	162	8.0	1.7	0.89	0.80	2.355
	cyh2-trp5	3	160	44.8	3.9	0.88	70.5	0.265
	leu1-ade6	2	507	34.3	1.8	0.89	46.7	0.271
	MAL1-ade3	1	138	52.0	6.4	0.78	68.5	0.495
8	*pet1-CUP1*	3	232	46.7	3.4	0.80	74.7	0.271
	thr1-CUP1	3	486	24.4	1.3	0.96	31.8	0.112
	CUP1-pet1	3	240	36.0	2.6	0.98	52.4	0.210
9	*his6-lys1*	1	411	47.1	3.3	0.55	61.1	0.461
10	*SUP4-SUP7*	2	179	53.1	5.4	0.78	80.1	0.498
11	*met14-met1*	1	109	46.2	6.1	0.96	59.7	0.545
	met1-MAL4	1	133	28.0	3.4	0.70	33.5	0.633
15	*ser1-ade2*	4	210	27.4	2.0	0.98	37.5	0.098
	ade2-cyh4	1	236	33.9	3.0	0.78	41.9	0.397
	pet17-ade2	2	232	48.8	3.9	0.79	72.8	0.367
17	*met2-pha2*	2	146	36.8	3.5	0.75	50.6	0.386
	pet2-pha2	1	151	50.7	5.9	0.93	66.4	0.551

TABLE 6: Data of *Schizosaccharomyces pombe* analyzed in Snow (1979)

Legend: Symbols in this table are the same as those of TABLE 5.

Chro	Gene pair	m	n	x_{χ^2}	se	p	x_s	k
1	*cyh1-cdc1*	1	142	58.0	7.5	0.75	77.7	0.545
	cdc1-leu2	1	124	51.7	7.0	0.73	67.7	0.698
	his1-leu2	0	102	11.5	2.7	0.22	11.5	4.691
	sup3-aro3	0	170	75.1	10.	0.56	76.2	1.350
	ura2-ade2	0	364	35.8	3.1	0.75	35.7	0.888
	ade2-ade4	0	290	122.	20.	0.72	121.	0.869
	lys3-ura1	0	692	19.9	1.4	0.17	19.9	1.771
	ura1-lys5	0	131	38.4	5.5	0.63	32.8	0.716
	pro1-ade3	0	100	49.6	8.1	0.92	49.7	1.087
	ade3-pro2	0	589	96.9	8.5	0.94	96.7	0.985

(continued)

Chro	Gene pair	m	n	x_{χ^2}	sd	p	x_s	k
2	*ade7-ura5*	1	556	12.3	1.0	0.94	13.4	0.229
	ade7-his3	0	392	69.5	6.1	0.72	69.8	1.117
	glu1-his3	0	212	66.8	8.0	0.81	66.5	0.911
	his3-mat1	1	728	80.1	6.0	0.49	111.	0.837
	tsl24-mat1	0	451	96.2	9.8	0.48	97.6	1.261
	leu1-his5	0	371	28.2	2.5	0.91	28.2	0.951
	his5-leu3	0	100	62.1	11.	0.94	61.8	0.921
	ade1-his4	0	498	42.5	3.0	0.48	42.7	1.249
	his4-trp1	0	128	109.	23.	0.75	111.	1.227
	ade8-arg4	0	185	31.3	4.0	0.87	31.2	0.894
3	*ade10-fur1*	0	142	17.6	3.0	0.70	17.5	0.581
	ade10-ade6	0	202	31.8	3.8	0.98	31.7	0.974
	fur1-sin2	0	206	25.4	3.1	0.90	25.3	0.892
	fur1-min5	0	199	22.4	2.9	0.85	22.4	1.177
	ade6-min5	0	337	4.5	0.8	0.05	4.4	7.693
	tsl5-arg1	5	48	5.16	7.3	0.98	42.3	0.302
	arg1-ade5	1	157	6.14	7.8	0.99	82.7	0.645
	arg1-aro4	0	67	56.3	12.	0.34	58.0	2.523
	trp3-aro4	0	126	19.7	3.4	0.90	19.6	1.151
	ade5-wee1	0	67	69.1	15.	0.85	68.5	0.850

TABLE 7: Data of McInn *et al.*(1993): observed and expected counts under the $Cx(Co)^2$ model

Legend: When m is assumed to be the same for both male and female, $Cx(Co)^2$ gives the best fit. Exp: expected counts, Obs: observed counts.

Crossovers	Exp(Male)	Obs(Male)	Exp(Female)	Obs(Female)
0	206	196	123	130
1	115	131	102	93
2	11	4	20	20
3	0	1	1	3

TABLE 8: Power comparison between grouped and ungrouped tests (1)

Legend: Simulated data are generated from the $Cx(Co)^3$ model, then tested against the hypothesis that the data are from the $Cx(Co)^2$ model. Different λ's correspond to different test statistics in power divergence family. α is the significance level at which the test is carried out.

λ	-1/2	0	1/2	1	5/2
grouped test					
α=0.01	0.020	0.010	0.001	0.001	0.003
α=0.05	0.099	0.055	0.021	0.014	0.014
α=0.10	0.200	0.121	0.060	0.039	0.041
ungrouped test					
α=0.01	0.045	0.042	0.037	0.031	0.019
α=0.05	0.144	0.139	0.127	0.124	0.107
α=0.10	0.223	0.221	0.217	0.213	0.196

TABLE 9: Power comparison between grouped and ungrouped tests (2)

Legend: Simulated data are generated from the $CxCo$ model, then tested against the hypothesis that the data are from the $Cx(Co)^2$ model.

λ	-1/2	0	1/2	1	5/2
grouped test					
α=0.01	0.009	0.086	0.123	0.077	0.048
α=0.05	0.069	0.213	0.278	0.235	0.017
α=0.10	0.139	0.321	0.407	0.393	0.308
ungrouped test					
α=0.01	0.099	0.111	0.115	0.115	0.125
α=0.05	0.250	0.254	0.251	0.256	0.271
α=0.10	0.359	0.363	0.366	0.372	0.379

TABLE 10: Comparison with Goldgar and Fain's model (Weinstein's Data)

Legend: Estimated genetic distances based on Goldgar and Fain's model and the $Cx(Co)^4$ model, and likelihood ratio statistics (LR).

Interval	*sc-ec*	*ec-cv*	*cv-ct*	*ct-v*	*v-g*	*g-f*	LR
G & F	7.2	9.9	8.6	15.0	11.4	11.4	132.4
$Cx(Co)^4$	7.1	9.6	8.3	14.8	11.5	11.5	219.1

Table 11: Comparison with Goldgar and Fain's model (Morgan's Data)

Legend: Estimated genetic distances based on Goldgar and Fain's model and the $Cx(Co)^4$ model, and likelihood ratio statistics (LR).

Interval	*sc-ec*	*ec-cv*	*cv-ct*	*ct-v*	*v-s*	*s-f*	*f-ca*	*ca-b*	LR
G & F	5.2	10.1	7.8	13.5	8.3	15.7	7.3	4.5	159.5
$Cx(Co)^4$	5.1	9.8	7.5	13.3	8.4	15.6	7.5	4.4	174.8