Chapter 10

Genetic Recombination

Mary Sara McPeck

Genetic recombination and genetic linkage are dual phenomena that arise in connection with observations on the joint pattern of inheritance of two or more traits or genetic markers. For example, consider two traits of the sweet pea, *Lathyrus odoratus*, an organism studied in depth by Mendel [9]: flower color, with purple (dominant) and red (recessive) phenotypes, and form of pollen, with long (dominant) and round (recessive) phenotypes. Under the Mendelian model for flower color (recast in more current terminology), each plant carries two alleles for flower color, one inherited from each parent, where each allele can be one of two types, denoted P and p . The pair of alleles carried by a plant is known as its genotype. Plants with genotype PP or Pp have purple flowers, while plants with genotype pp have red flowers. Mendel's First Law can be interpreted as specifying that a parent plant passes on a copy of one of its two alleles to each offspring, with each parental allele having an equal chance of being copied, and with this occurring independently across offspring and across parents. Similarly, each plant carries two alleles for form of pollen, where each of these can be L or l . Plants with genotype LL or Ll have long pollen, while plants with genotype ll have round pollen. Suppose one crossed a true-breeding parental line having purple flowers and long pollen (all individuals having genotype $PPLL$) with a true-breeding parental line having red flowers and round pollen (all individuals having genotype $ppll$). Then the offspring of that cross, known as the F_1 generation, would all have genotype $PpLl$, resulting in purple flowers and long pollen. Suppose a backcross were performed, in which F_1 individuals were crossed with individuals from the $ppll$ parental line. In this example, genetic **linkage** would refer to a tendency for pairs of alleles inherited from the same parent, such as the pair PL or the pair pl , to be transmitted together during meiosis, while genetic **recombination** would refer to the event that an individual transmits a pair of alleles that were inherited from different parents, such as the pair Pl or pL . If we let $0 \leq \theta \leq .5$ denote the **recombination fraction**, which is the probability of a recombination between

M.S. McPeck

Departments of Statistics and Human Genetics, University of Chicago

e-mail: mcpeek@uchicago.edu

the genes for these two traits in a single meiosis, then in the backcross offspring, we expect individuals with genotypes $PpLl$, $ppll$, $Ppll$ and $ppLl$ to occur with relative frequencies $(1 - \theta)/2$, $(1 - \theta)/2$, $\theta/2$ and $\theta/2$, respectively.

A long-standing, important application of the ideas of linkage and recombination is to construction of genetic maps [15] and to subsequent localization of genes (or other genetic variants of interest) on those maps. The key observation is that the recombination fraction between a pair of genetic markers tends to increase with the chromosomal distance between them, with markers on different chromosomes having recombination fraction .5. Thus, by merely observing patterns of joint inheritance of traits, one can make inference about which trait genes lie on the same chromosome, chromosome, and make estimates of distances between them. The basic ideas of and mathematics behind linkage and recombination were developed early in the 20th century [10, 15, 5]. Notably, these problems attracted the interest of R. A. Fisher [3].

Starting in the early 1980s, there was a resurgence of interest in the problem of genetic map construction, spurred by the development of recombinant DNA technology, which resulted in the ability to collect genotype data on large numbers of neutral genetic markers throughout the human genome [1] as well as genomes of model organisms. It was not long after these technological breakthroughs occurred that Terry shifted much of his energy and interest into the field of statistical genetics, near the beginning of the explosion of new data and resulting need for new statistical models and methods. In human data, the map construction problem called for more sophisticated statistical analysis than that typically required in experimental organisms. In model organisms, experimental crosses can often be planned in such a way that it is feasible to simply observe the relative frequency of recombinants in any given interval and convert it to a distance using a “map function”, an analysis method that we will call the “two-point analysis.” However, in humans, crosses cannot be planned, and so any given human meiosis would typically be uninformative for some of the markers of interest. (For example, in the sweet pea example above, all meioses from an individual with genotype $Ppll$ would be uninformative for recombination between these two genes, because the recombinant and non-recombinant allele pairs are indistinguishable.) When many genetic markers are considered simultaneously in each meiosis, and many meioses from different individuals (with different patterns of informativeness) are analyzed together, substantial additional information, beyond that available from a two-point analysis, can typically be obtained by a joint analysis using a suitable statistical model for joint recombination events among a collection of genetic markers.

Thus, the statistical challenges of genetic mapping in humans naturally led to consideration of probability models for the crossover process that causes the observation of recombination. In humans and other diploid eukaryotes, crossing over takes place during a phase of meiosis in which the two parental versions of a given chromosome have each been duplicated, and all four resulting strands or chromatids are lined up together, forming a tight bundle. Crossovers can be modeled as points located along this bundle, with each crossover involving exactly two of the four chromatids. It is assumed that the two chromatids involved in any particular

crossover are nonsister chromatids, that is, the two chromatids cannot be the two identical copies of one of the parent's versions of the chromosome. After crossing over has occurred, the four resulting chromatids are each mosaics of the original parental chromosomes. Keeping in mind this framework, one can consider two key aspects of the model: (1) the distribution of crossover points along the bundle of four chromatids and (2) the choice of nonsister pair of chromatids involved in each crossover. Perhaps the simplest useful model is the no-interference model of Haldane [5], which models aspect (1) by assuming that the crossover points form a homogeneous Poisson process and models aspect (2) by assuming that each nonsister pair is equally likely to be chosen for each crossover, independently across crossovers. **Interference** refers to deviation from Haldane's model. Interference, in the form of local inhibition of crossover points on a resulting single chromatid, was readily apparent in early *Drosophila* data [16, 11]. It is convenient to refer to failure of assumption (1) of Haldane's model as **crossover interference** and failure of assumption (2) of Haldane's model as **chromatid interference**.

Under the assumption of no chromatid interference (NCI), Speed et al. [14] derive a set of constraints, on the multilocus recombination probabilities, that are necessary and sufficient to ensure the existence of a counting process model for the distribution of crossover points along the bundle of four chromatids. They apply these constraints to prove a consistency result for the maximum likelihood estimate of the map order of a finite number of genetic markers along a chromosome. Specifically, they show that, under the assumption of NCI, in the case of complete data, i.e. when all meioses are informative for all markers, if maximum likelihood estimation is performed assuming the Haldane model, then the MLE will converge almost surely to the true map order, even when the true crossover point process is not Poisson (it can be any counting process).

The idea that the assumption of NCI imposes constraints on multilocus recombination probabilities is developed further in Zhao et al. [18], in which the main goal is assessment of the empirical evidence for chromatid interference. This paper extends the constraints from single spore data (such as that from humans and *Drosophila*) to tetrad data (from organisms such *Neurospora crassa*, *Saccharomyces cerevisiae* and *Aspergillus nidulans*) in which data on all 4 chromatid strands are available for each meiosis, providing much more information about strand choice and, hence, allowing a more powerful test of the NCI assumption. An efficient iterative algorithm for maximum likelihood estimation under the constraints is developed, and a likelihood ratio test is proposed to assess whether there is evidence that the constraints are not satisfied by the multinomial model assumed to generate the data. An empirical bootstrap approach is used to assess significance. Some of the experiments did provide evidence for chromatid interference, but overall there was no consistent pattern. The extent and type of chromatid interference seemed to vary across organisms and across experiments. Because the loci considered in these experiments are functional genes, as opposed to neutral markers, it is possible that differential viability may play a role in the results as well. In single-spore data, in particular, the constraints imposed by NCI are rather weak, and the available data do not provide

much power to contradict them. Therefore, it seemed reasonable to assume NCI and focus attention on models for the crossover process.

Because the Haldane no-interference model was so clearly contradicted by most of the available, relevant data, Terry was somewhat concerned about relying on it for map inference. If a more flexible, yet still parsimonious and tractable, model could be developed and shown to fit the data better, Terry reasoned, it could be useful for a range of applications in genetic inference. This problem is addressed by McPeck and Speed [8], in which a range of point process models, involving one or two additional parameters, are fit to *Drosophila* data by maximum likelihood. Goodness of fit of the models is assessed, and the pattern of interference generated by each model is compared to that in data. The most promising model that emerges from this study, the **gamma model**, is a stationary gamma renewal process on four strands, combined with the assumption of NCI to generate a thinned process. In addition to fitting the data better and providing a pattern of interference that mimics that in data, the gamma model is also parsimonious and, when an integer shape parameter is used, results in efficient computational methods. This promising model is further developed in Zhao et al. [19], in which the gamma model with integer shape parameter is referred to as the **chi-square model** because it results in a stationary renewal process having chi-square interarrivals (with even degrees of freedom) for the process on a single strand. The model is fit to datasets from a number of different organisms, with different datasets from the same organisms having similar estimated shape parameter. The results of the analyses suggest that it may be reasonable to use an organism-specific shape parameter to model interference.

In a closely-related line of research, Terry and colleagues sought to connect probability modeling of the crossover process with the initially mysterious-seeming map functions commonly used in two-point analysis. A **map function** is used to convert probability of recombination across an interval to genetic distance of the interval, where **genetic distance** is defined as the expected number of crossovers per strand per meiosis. A difficulty in application of map functions to multilocus analyses is that when there are more than three markers, the multilocus recombination probabilities cannot be uniquely determined from the map function [3]. Earlier work [4, 13, 7] had proposed to solve this identifiability problem by constraining the probability of an odd number of crossovers across a union of disjoint intervals to depend only on the total length of these intervals. However, this is not a biologically plausible assumption, and, as shown by Evans et al. [2], assuming NCI, the class of count-location models [6, 12] is the only class of models having map functions that satisfy this constraint. Zhao and Speed [17] remove this biologically implausible constraint, and instead solve the general problem of developing stationary renewal process models that can generate specific map functions. They show that in most cases of previously-proposed map functions, one can construct a stationary renewal process that generates the map function. Furthermore, they show that this stationary renewal process can typically be approximated quite well by the gamma or chi-square model. The useful practical consequence of this is that two-point analyses using a particular map function can easily be extended to more informative multipoint analyses, an approach that is particularly valuable in the presence of missing data.

References

- [1] D. Botstein, R. L. White, M. Skolnick, and R. W. Davis. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.*, 32:314–331, 1980.
- [2] S. N. Evans, M. S. McPeck, and T. P. Speed. A characterisation of crossover models that possess map functions. *Theor. Popul. Biol.*, 43:80–90, 1993.
- [3] R. A. Fisher. The theory of linkage in polysomic inheritance. *Phil. Trans. Roy. Soc. B*, 233:55–87, 1947.
- [4] H. Geiringer. On genetic map functions. *Ann. Math. Statist.*, 142:1369–1377, 1944.
- [5] J. B. S. Haldane. The combination of linkage values, and the calculation of distances between the loci of linked factors. *J. Genet.*, 8:299–309, 1919.
- [6] S. Karlin and U. Liberman. A natural class of multilocus recombination processes and related measures of crossover interference. *Adv. Appl. Prob.*, 11: 479–501, 1979.
- [7] U. Liberman and S. Karlin. Theoretical models of genetic map functions. *Theor. Popul. Biol.*, 25:331–346, 1984.
- [8] M. S. McPeck and T. P. Speed. Modeling interference in genetic recombination. *Genetics*, 139:1031–1044, 1995.
- [9] G. Mendel. Versuche über pflanzenhybriden. *Verh. Naturf. Ver. Brünn*, 4:3–44, 1866.
- [10] T. H. Morgan. Random segregation versus coupling in Mendelian inheritance. *Science*, 34:384, 1911.
- [11] H. J. Muller. The mechanism of crossing-over. *Am. Nat.*, 50:193–221, 284–305, 350–366, 421–434, 1916.
- [12] N. Risch and K. Lange. An alternative model of recombination and interference. *Ann. Hum. Genet.*, 43:61–70, 1979.
- [13] F. W. Schnell. Some general formulations of linkage effects in inbreeding. *Genetics*, 46:947–957, 1961.
- [14] T. P. Speed, M. S. McPeck, and S. N. Evans. Robustness of the no-interference model for ordering genetic markers. *Proc. Natl. Acad. Sci. USA*, 89:3103–3106, 1992.
- [15] A. H. Sturtevant. The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *J. Exp. Zool.*, 14:43–59, 1913.
- [16] A. H. Sturtevant. The behavior of the chromosomes as studied through linkage. *Zeit. f. ind. Abst. u. Vererb.*, 13:234–287, 1915.
- [17] H. Zhao and T. P. Speed. On genetic map functions. *Genetics*, 142:1369–1377, 1996.
- [18] H. Zhao, M. S. McPeck, and T. P. Speed. Statistical analysis of chromatid interference. *Genetics*, 139:1057–1065, 1995.
- [19] H. Zhao, T. P. Speed, and M. S. McPeck. Statistical analysis of crossover interference using the chi-square model. *Genetics*, 139:1045–1056, 1995.

Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 3103–3106, April 1992
Genetics

Robustness of the no-interference model for ordering genetic markers

(multilocus recombination/chromosome mapping/no chromatid interference)

T. P. SPEED, M. S. MCPEEK, AND S. N. EVANS

Department of Statistics, University of California, Berkeley, CA 94720

Communicated by Peter J. Bickel, December 30, 1991

ABSTRACT Under the assumption of no chromatid interference, we derive constraints on the probabilities of the different recombination patterns among $m + 1$ genetic loci. An application of these constraints is a proof that the ordering of the loci that maximizes the likelihood under the assumption of no interference is, in fact, a consistent estimator of the true order even when there is interference.

Genetic mapping involves ordering a set of markers on a chromosome and finding distances between them. One way this may be done is through analysis of data on meiotic recombination between the markers. Meiotic recombination is believed to be the result of crossing-over between nonsister chromatids during the pachytene phase of meiosis. If a particular chromatid passed on in meiosis was involved in an odd number of crossovers between two loci, a *recombination* is said to have taken place between the two loci (see Fig. 1).

It is important to keep in mind that crossing-over takes place in the four-stranded state, when each chromosome has duplicated to form two sister chromatids. In that case, the two aspects relevant to recombination are (i) the distribution of crossovers along the chromosome and (ii) which pair of nonsister chromatids is involved in each crossover. Two simplifying assumptions are often made. The first is that the locations of different crossovers are independent and identically distributed (i.i.d.) along the chromosome. The second is that each pair of nonsister chromatids is equally likely to be involved in a crossover, independent of which were involved in other crossovers. If the occurrence of a crossover influences the probability of another's occurring nearby, in violation of the first assumption, it is termed *crossover position interference*. Following Whitehouse (1), we prefer this over the traditional term *chiasma interference*. If the second assumption is violated, it is termed *chromatid interference*. There is a considerable body of data demonstrating both kinds of interference (1), but on the whole the extent of chromatid interference seems slight. On the other hand, position interference can be substantial and can take different forms. In what follows, we will permit an arbitrary crossover location point process, but we will assume no chromatid interference.

A General Model with No Chromatid Interference

Assuming no chromatid interference, we model the occurrence of crossovers along a chromosome as a realization of a point process, with the points corresponding to the locations of the crossovers. That is, we associate the chromosome with the interval $[0, 1]$ and require (i) a distribution for n = the total number of points in the interval and (ii) for each $n \geq 1$, the joint distribution of the positions of the points,

FIG. 1. Three-strand double crossover.

given that their total number is n . Furthermore, we require that the point process be *simple*: i.e., with probability 1, no two points shall occupy the same location. A simple point process on $\mathfrak{R}$ or a subset of $\mathfrak{R}$ is known as a *counting process*.

Define the *avoidance function* or *zero function* Z of the process by $Z(A) = \Pr(\text{no points in } A)$, for each measurable set A . It is well known that the distribution of a simple point process on a complete separable metric space is determined by the values of the avoidance function on the Borel sets (2). We shall find that the avoidance function of the crossing-over process is closely related to the recombination probabilities.

Following is a well-known derivation (see, e.g., ref. 3) of Mather's Formula (4), which expresses the chance of recombination in an interval in terms of the avoidance function, in the case of no chromatid interference. If we assume that there is no chromatid interference, then each crossover is equally likely to involve any of the four possible nonsister pairs of chromatids, independent of which pairs are involved in other crossovers. In that case, if there are $n > 0$ crossovers between loci at locations A and B , $0 \leq A < B \leq 1$, then any given chromatid has probability

$$\binom{n}{i} \times \frac{1}{2^n}$$

of being involved in exactly i of them, for $0 \leq i \leq n$. In a given meiotic product (sperm or oocyte), a recombination between A and B will have occurred if the chromatid passed on in the meiosis has been involved in an odd number of crossovers between A and B . Thus, the chance of a recombination given that $n > 0$ crossovers have occurred is

$$\frac{1}{2^n} \times \sum_{i=0}^{n-1} \binom{n-1}{i} \left(\frac{n}{2i+1} \right) = \frac{1}{2},$$

so the chance of a recombination is $\frac{1}{2} \times \Pr\{n > 0\} = \frac{1}{2} \times (1 - Z([A, B]))$.

More generally, consider the recombination pattern among $m + 1$ ordered loci. Let A_j denote the interval between loci j and $j + 1$. Let $p_x, x = (x_1, \dots, x_m) \in \{0, 1\}^m$, denote the probability of the event of x_j recombinations in $A_j, j = 1, \dots, m$. Let Z_x denote $Z(\cup\{A_j; x_j = 1\})$, the probability that the point process avoids all of the intervals A_j with $x_j = 1$. In the genetics literature, p_x is known as the *crossover distribution*,

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Genetics: Speed *et al.*

and Z_x is known as the *linkage value* associated with x . The following relationship between Z_x and p_x is well known:

$$Z_x = \sum_y (-1)^{y \cdot x} p_y$$

and

$$p_x = \frac{1}{2^m} \times \sum_y (-1)^{y \cdot x} Z_y,$$

where the sum is over all $y = (y_1, \dots, y_m) \in \{0, 1\}^m$ and $y \cdot x = \sum_{j=1}^m y_j x_j$. That is, (Z_x) is the 2^m Fourier transform of (p_x) , and (p_x) is the inverse 2^m Fourier transform of (Z_x) (5-7).

This relationship between the distribution of recombinations among $m + 1$ loci and the avoidance function can be exploited to get necessary and sufficient conditions for the distribution of recombinations among $m + 1$ loci to be compatible with at least one underlying crossover point process. It is easy to get necessary conditions on the p_x values as follows: define q_x , $x = (x_1, \dots, x_m) \in \{0, 1\}^m$ to be the probability of the event of no crossovers in each of the intervals A_j with $x_j = 0$ and at least one crossover in each of the intervals A_j with $x_j = 1$; $j = 1, \dots, m$. Note that (q_x) and (Z_x) refer to the distribution of crossovers along the four-stranded chromosome, while (p_x) refers to the recombination distribution on a single strand. Then, clearly

$$Z_x = \sum_{y \leq x'} q_y,$$

where $y \leq x'$ denotes $y_j \leq 1 - x_j$, $j = 1, \dots, m$. Plugging into the formula for p_x in terms of Z_x , we get

$$p_x = \sum_{y \leq x'} \frac{1}{2^{m-1}} q_y.$$

Alternatively, the relation between the p s and the q s may be proved from Lemma 1 of ref. 8, by induction. Inverting, we have

$$q_x = \sum_{y \geq x'} (-1)^{(y-x') \cdot 1} Z_y$$

and

$$q_x = 2^{x \cdot 1} \times \sum_{y \geq x'} (-1)^{(y-x') \cdot 1} p_y.$$

Now it is clearly necessary that we have

$$(i) \quad \sum_x q_x = 1,$$

and

$$(ii) \quad \text{for all } x, q_x \geq 0.$$

Expressed in terms of the p_x values, these conditions are

$$(i) \quad \sum_x p_x = 1,$$

and

$$(ii) \quad \text{for all } x, 0 \leq \sum_{y \geq x'} (-1)^{(y-x') \cdot 1} p_y.$$

We further note that these conditions can be expressed in terms of the Z_x values as

$$(i) \quad Z(\emptyset) = 1,$$

and

Proc. Natl. Acad. Sci. USA 89 (1992)

$$(ii) \quad \text{for all } x, 0 \leq \sum_{y \geq x} (-1)^{(y-x) \cdot 1} Z_y.$$

If we assume that any given pattern of crossovers is possible, that is, that each of the parameters q_x is nonzero, then the weak inequalities satisfied by the parameters (p_x) and (Z_x) become strict. Of course, the constraints in terms of the p_x values are of greatest interest, since the p_x values correspond to the observed data.

In fact, these conditions on the p_x values are also sufficient, under the assumption of no chromatid interference, for the existence of an underlying point process of crossovers that would be compatible with the p_x values. To show this, we need only construct, given any set of q_x values with $q_x \geq 0$ and $\sum_x q_x = 1$, a point process on $[0, 1]$ that is compatible with them. This is easily done by fixing one point in each interval A_j and allowing crossovers only at those points. The pattern of crossovers among those m points is then chosen to be x with probability q_x , where for all j ,

$$x_j = \begin{cases} 1 & \text{if there is a crossover in } A_j \\ 0 & \text{otherwise} \end{cases}$$

Note that our constraints are stronger than those required by Karlin and Liberman (5) for their class of "natural" recombination distributions. They require

$$(i) \quad Z(\emptyset) = 1,$$

and

$$(ii) \quad \text{for all } x, Z_x \geq 0.$$

An Application of the Constraints: Robustness of the Poisson Model

PROPOSITION. Suppose we have recombination data on $m + 1$ loci whose true order is unknown. Assume that there is no chromatid interference, but crossover location interference of an arbitrary form may be present. That is, the positions of crossovers and the occurrence of recombinations are given by a model of the type described in the preceding section. Assume further that we are in the nondegenerate situation in which each of the parameters q_x is nonzero. Suppose that our data are from n meioses and for each meiosis we can observe whether or not a recombination occurred between each of the $\frac{1}{2}m(m + 1)$ pairs of loci. Suppose that for each possible order of the loci we fit the data by maximum likelihood under the assumption of no crossover-location interference. Then with probability 1 for n sufficiently large, the maximized likelihood will be largest for the true order.

If we arbitrarily choose an ordering of the loci $f = (f_1, \dots, f_{m+1})$, where f is a permutation of $(1, \dots, m + 1)$ and fit the data by maximum likelihood, assuming no interference, we will be fitting m parameters $(\theta_{f_1, f_2}, \dots, \theta_{f_m, f_{m+1}})$, where $\theta_{f_j, f_{j+1}}$ is the chance of a recombination between loci f_j and f_{j+1} . The proof (see Appendix) lies in showing that for any nonidentity permutation f , with probability 1 for n sufficiently large, the collection $\{\theta_{f_1, f_2}, \dots, \theta_{f_m, f_{m+1}}\}$ dominates $\{\theta_{1, 2}, \dots, \theta_{m, m+1}\}$. That is, the two sets can be put into a one-to-one correspondence such that each element of the first set is larger than or equal to the corresponding element of the second set, with at least one strict inequality. This is because, with probability 1 for n sufficiently large, the constraints on the p_i values imply constraints on the data. Then it follows that the likelihood will be maximized by the true order.

Genetics: Speed *et al.*

Proc. Natl. Acad. Sci. USA 89 (1992)

Discussion

As we try to increase the resolution of genetic maps, we shall find that interference plays a greater role. Although estimation of order of loci under the assumption of no crossover position interference is consistent, still the number of data points n may need to be very large in practice. It is likely that with a reasonable model for interference one could use the data to estimate order more efficiently. We note that even in the absence of a specific interference model one could compare maximized likelihoods under different orders, where the likelihood is in terms of $2^m p_x$ values subject to the constraints, rather than in terms of $m \theta$ values.

Another area in which the constraints may be useful is in determining when a map function has an underlying crossover point process [an extension to Liberman and Karlin (6)]. Liberman and Karlin define a map function M (which converts expected number of crossovers to chance of recombination) to be "multilocus feasible" if it satisfies

$$(i) \quad M(0) = 0$$

and

$$(ii) \quad \text{for all } x, 0 \leq \sum_{y \in [0,1]^m} (-1)^{yx} (1 - 2M(\gamma_y)),$$

where γ_y is the map distance of $\cup_{j:y_j=1} A_j$. We could replace ii with the more stringent condition

$$(iii) \quad \text{for all } x, 0 \leq \sum_{y \in x} (-1)^{(y-x)1} (1 - 2M(\gamma_y)).$$

The connection between map functions and crossover point processes has been determined (unpublished work).

Appendix

Proof of Proposition: As before, let A_j denote the interval between loci j and $j+1$ in the true order, $j = 1, \dots, m$, and let $p_x, x = (x_1, \dots, x_m) \in \{0, 1\}^m$ denote as before. Assume that the constraints of the general model described above hold with strict inequalities.

The data consist of $r_x^n, x = (x_1, \dots, x_m) \in \{0, 1\}^m$, where r_x^n denotes the number of meioses in which x_j recombinations occurred in $A_j, j = 1, \dots, m$. Note that we can still calculate the set of numbers $\{r_x^n\}$ without knowing the true order. We could do the calculation assuming an arbitrary order, and the resulting set of counts $\{r_x^n\}$ would be the same but for an (unknown) permutation of indices. Assuming that the recombination patterns in different meioses are i.i.d., (r^n) is distributed as Multinomial(n, p), where r^n is the vector of r_x^n values and p is the vector of p_x values.

If we arbitrarily choose an ordering of the loci $f = (f_1, \dots, f_{m+1})$, where f is a permutation of $(1, \dots, m+1)$ and fit the data by maximum likelihood, assuming no interference, we will be fitting m parameters $(\theta_{f_1 f_2}, \dots, \theta_{f_m f_{m+1}})$, where $\theta_{f_j f_{j+1}}$ is the chance of a recombination between loci f_j and f_{j+1} . Suppose we maximize the likelihood under the true order and under any other order f and compare the maximized likelihoods. In the case in which we can observe any recombination among the loci, the maximum likelihood estimates of the θ s are very simple. For the true order, we have $\hat{\theta}_j = \min(\frac{1}{2}, n^{-1} \sum_{x: x_j=1} r_x^n)$, where $\hat{\theta}_j$ is the maximum likelihood estimate of the chance of a recombination in the interval $A_j, j = 1, \dots, m$. For the order f , we have $\hat{\theta}_{f_j f_{j+1}} = \min(\frac{1}{2}, n^{-1} \sum_{x \in I} r_x^n)$ where I is the set whose members x all satisfy

$$\sum_{k: A_k \text{ lies between } f_j \text{ and } f_{j+1}} x_k \text{ is odd.}$$

For all $x \in \{0, 1\}^m$ set

$$c_x^n = 2^{x-1} \sum_{y \geq x} (-1)^{(y-x)1} r_y^n.$$

and consider the strict constraints $c_x^n > 0, x \in \{0, 1\}^m$. By the law of large numbers, the probability that these constraints are satisfied is 1 for n sufficiently large because of the constraints on p . Assume that these constraints on r^n hold. In terms of the c_x^n values, we have $\hat{\theta}_j = \frac{1}{2} n^{-1} \sum_{x: x_j=1} c_x^n$ and $\hat{\theta}_{f_j f_{j+1}} = \frac{1}{2} n^{-1} \sum_{x \in I} c_x^n$, where $I = \{x: x_k = 1 \text{ for at least one } A_k \text{ lying between } f_j \text{ and } f_{j+1}\}$. Note that we have $\hat{\theta}_j$ and $\hat{\theta}_{f_j f_{j+1}}$ both $\leq \frac{1}{2}$ under the constraints. From this representation, we can see that if A_k lies between f_j and f_{j+1} , then $\hat{\theta}_k \leq \hat{\theta}_{f_j f_{j+1}}$, since $c_x^n > 0$ for all x .

Now we will match each of the $m \hat{\theta}_{f_j f_{j+1}}$ values in one-to-one correspondence with a $\hat{\theta}_k$ that is smaller than or equal to it. This will prove the proposition, for the function $g(x) = x \log(x) + (1-x) \log(1-x)$ is decreasing in $0 \leq x \leq \frac{1}{2}$. Thus, if L_f is the maximized likelihood under order f , and L_{true} is the maximized likelihood under the true order, we will have

$$L_f = n \sum_{j=1}^m g(\hat{\theta}_{f_j f_{j+1}}) \leq n \sum_{j=1}^m g(\hat{\theta}_j) = L_{true}.$$

To see that the $\{\hat{\theta}_{f_j f_{j+1}}\}$ can be matched with the $\{\hat{\theta}_k\}$ in such a way that the $\hat{\theta}_k$ corresponding to $\hat{\theta}_{f_j f_{j+1}}$ is no larger than it, we use P. Hall's matching theorem (see p. 401 of ref. 9). We associate each interval (f_j, f_{j+1}) in the ordering f with the set of intervals A_k in the true ordering that lie between f_j and f_{j+1} . As noted above, this ensures that $\hat{\theta}_{f_j f_{j+1}}$ is greater than or equal to $\hat{\theta}_k$ for any k of this kind. The condition of Hall's theorem that must be checked is that any set $\{(f_j, f_{j+1}): j \in J\}$ of $|J|$ distinct intervals in the order f must contain at least $|J|$ distinct intervals in the original order. Then there is a matching with the property stated and the proof is complete.

To show that the condition holds, argue by induction on the number of loci. The condition clearly holds for two loci. Suppose it holds for M loci, and consider the case of $M+1$ loci. Let $\{(f_j, f_{j+1}): j \in J\}$ be any set of $|J|$ distinct intervals in some order f . Let $i = \min\{j \in J\}$, so f_i is an endpoint of exactly one interval in the set. Now consider the M loci $\{1, 2, \dots, M+1\} \setminus \{f_i\}$ and the $|J|-1$ intervals $\{(f_j, f_{j+1}): j \in J \setminus \{i\}\}$. By the induction hypothesis, these intervals cover at least $|J|-1$ distinct intervals in the original order of $\{1, 2, \dots, M+1\} \setminus \{f_i\}$. Note that if the $\{(f_j, f_{j+1}): j \in J \setminus \{i\}\}$ cover exactly k distinct intervals in the original order of $\{1, 2, \dots, M+1\} \setminus \{f_i\}$, then they must cover either k or $k+1$ distinct intervals in the original order of $\{1, 2, \dots, M+1\}$. If at least $|J|$ intervals of the original order of $\{1, \dots, M+1\}$ are covered by $\{(f_j, f_{j+1}): j \in J \setminus \{i\}\}$, then we are done, so assume without loss of generality that exactly $|J|-1$ distinct intervals of the original orders of both $\{1, \dots, M+1\}$ and $\{1, \dots, M+1\} \setminus \{f_i\}$ are covered. That is, the same number of intervals are covered in the original order whether locus f_i is included or not. Thus, if we say f_i is between loci $f_i - 1$ and $f_i + 1$ in the original order, then neither $(f_i - 1, f_i)$ nor $(f_i, f_i + 1)$ could be covered. Otherwise, $(f_i - 1, f_i + 1)$ would have to be covered by an element of $\{(f_j, f_{j+1}): j \in J \setminus \{i\}\}$, but then adding in locus f_i would add in one more interval covered, which contradicts our assumption. Since the interval (f_i, f_{i+1}) must contain at least one of $(f_i - 1, f_i)$ and $(f_i, f_i + 1)$, then $\{(f_j, f_{j+1}): j \in J\}$ covers at least $|J|$ distinct intervals in the original order of the $M+1$ loci. The argument is similar if f_i is the first or last locus in the original order.

T.P.S. was supported in part by National Science Foundation Grant DMS-880237. M.S.M. was supported in part by a National Science Foundation predoctoral research fellowship. S.N.E. was supported in part by National Science Foundation Grant DMS-9015708.

Genetics: Speed *et al.*

1. Whitehouse, H. L. K. (1973) *Towards an Understanding of the Mechanism of Heredity* (St. Martin's, New York), 3rd Ed., pp. 111–112.
2. Daley, D. J. & Vere-Jones, D. (1988) *An Introduction to the Theory of Point Processes* (Springer, New York), pp. 216–218.
3. Karlin, S. & Liberman, U. (1983) *Adv. Appl. Probab.* **15**, 471–487.
4. Mather, K. (1938) *Biol. Rev.* **13**, 252–292.

Proc. Natl. Acad. Sci. USA 89 (1992)

5. Karlin, S. & Liberman, U. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 6332–6336.
6. Liberman, U. & Karlin, S. (1984) *Theor. Popul. Biol.* **25**, 331–346.
7. Risch, N. & Lange, K. (1983) *Biometrics* **39**, 949–963.
8. Lange, K. & Risch, N. (1977) *J. Math. Biol.* **5**, 55–59.
9. Aigner, M. (1979) *Combinatorial Theory* (Springer, New York), pp. 152, 401.

Modeling Interference in Genetic Recombination

Mary Sara McPeck and Terence P. Speed

Department of Statistics, University of California, Berkeley, California 94720

Manuscript received December 17, 1993

Accepted for publication October 25, 1994

ABSTRACT

In analyzing genetic linkage data it is common to assume that the locations of crossovers along a chromosome follow a Poisson process, whereas it has long been known that this assumption does not fit the data. In many organisms it appears that the presence of a crossover inhibits the formation of another nearby, a phenomenon known as "interference." We discuss several point process models for recombination that incorporate position interference but assume no chromatid interference. Using stochastic simulation, we are able to fit the models to a multilocus *Drosophila* dataset by the method of maximum likelihood. We find that some biologically inspired point process models incorporating one or two additional parameters provide a dramatically better fit to the data than the usual "no-interference" Poisson model.

THE phenomenon of interference in genetic recombination was noticed very early this century by *Drosophila melanogaster* geneticists in THOMAS HUNT MORGAN's lab (STURTEVANT 1915; MULLER 1916). They found that simultaneous recombination in two or three nearby chromosomal intervals occurred much less often than would be expected under independence, and that the effect appeared to decrease with distance. At present the biological nature of genetic interference is still not well understood, nor has it been adequately modeled mathematically. Virtually all multilocus linkage analyses use the assumption of no interference. Although this assumption may give consistent results for locus ordering (SPEED *et al.* 1992) and estimation of recombination fractions in some cases, it clearly does not fit the data. It is natural to ask if such an analysis could be improved by the use of a reasonable interference model. What is required is a biologically plausible point process model for crossovers along a chromosome, which should be fit to data. The *Drosophila* dataset of MORGAN *et al.* (1935) is ideal for this purpose. Previous attempts to fit a crossover point process to the MORGAN *et al.* (1935) *Drosophila* dataset (COBBS 1978; RISCH and LANGE 1983; PASCOE and MORTON 1987; GOLDBERG and FAIN 1988; FOSS *et al.* 1993) have been severely limited by the difficulty of calculating the likelihood of the data under all but the simplest models. Using a Monte Carlo method, we are able to fit a wide range of point process models to the data.

BACKGROUND

In diploid eukaryotes, crossing over takes place during the pachytene phase of meiosis, when the two ho-

mologous versions of any particular chromosome have each been duplicated and all four resulting strands or chromatids are lined up together, forming a very tight bundle. We model crossovers as being points located along this bundle, and each crossover involves exactly two of the four chromatids. We assume that the two chromatids involved in any particular crossover are nonsister chromatids, that is, the two chromatids cannot be the two copies of a single homologous chromosome. After crossing over has occurred, the four resulting chromatids are mixtures of the original parental types. In *Drosophila*, for each meiosis, we will have information on only one of these four resulting chromatids. If that chromatid was involved in an odd number of crossovers between two loci, a recombination is said to have taken place between the two loci.

It is important to keep in mind that crossing over takes place among four chromatids. In that case the two aspects relevant to recombination are the distribution of crossovers along the bundle of four chromatids and which pair of nonsister chromatids is involved in each crossover. The concept of interference is usually divided into two parts, corresponding to these two aspects. First, we say that there is position or chiasma interference if the number and location of crossovers in a given region are not independent of the numbers and locations of crossovers in disjoint regions. Second, we say there is chromatid interference if it is not the case that each pair of nonsister chromatids is equally likely to be involved in a crossover, independent of which were involved in other crossovers. There is little consistent evidence of chromatid interference (ZHAO *et al.* 1995b).

We assume no chromatid interference and try to find a better-fitting point process model than the Poisson for

Corresponding author: Mary Sara McPeck, Department of Mathematics, University of Southern California, Los Angeles, CA 90089-1113. E-mail: mcpeck@galton.uchicago.edu

M. S. McPeck and T. P. Speed

the occurrence of crossovers along the chromosome. In the case of no chromatid interference, the point process on the single chromatid can be obtained from the point process on the bundle of four by independently thinning (*i.e.*, deleting) each point with probability $1/2$. This is because the given chromatid has chance $1/2$ of being involved in a particular crossover, independent of involvement in any others. As a result of assuming no chromatid interference, the chance of recombination across an interval increases monotonically as the interval is enlarged, with an upper bound of $1/2$.

The traditional measure of interference is the coincidence (STURTEVANT 1915; MULLER 1916), 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.

$$C = \frac{r_{ij}}{(r_{i0} + r_{i1})(r_{0j} + r_{1j})},$$

where C is the coincidence and r_{ij} is the chance of i recombinations across the first interval and j recombinations across the second interval. If there were no position interference, the coincidence would equal one. Observed coincidences tend to be near zero for small, closely-linked intervals, increasing to one for more distant intervals.

The recombination point processes considered here are stationary in terms of some distance metric, although this distance metric is generally not equivalent to physical distance. This point will be discussed further below. For each of the point process models considered here, there is more than one such choice of a distance metric that will make the process stationary; these metrics are the same up to a multiplicative constant. The genetic distance associated with a chromosome interval is defined to be the expected number of crossovers occurring on a single chromatid within that interval. The metrics in which the models discussed here are stationary are all constant multiples of genetic distance. We shall choose one for notational convenience and call it the stationary metric. Among other things stationarity of the model means that coincidence will be a function of the distances across each of the two intervals considered and the distance between the intervals (in terms of the stationary metric) but not their actual locations. This also implies that the intensity, $\mu = \lim_{h \rightarrow 0} h^{-1} \Pr(\text{at least 1 point in } [t, t+h])$, where h is measured in terms of the stationary metric, is the same for all t . Furthermore, for the models considered here the conditional intensity can be defined by $\mu(z) = \lim_{\delta \rightarrow 0} \lim_{\epsilon \rightarrow 0} \epsilon^{-1} \Pr(\text{at least 1 point in } [t+z, t+z+\epsilon] | \text{at least 1 point in } [t, t+\delta])$, where ϵ is measured in terms of the stationary metric. Intuitively, this is the intensity at

$t+z$ conditional on a point at t , and by stationarity this depends only on z . When the conditional intensity exists, the coincidence function can be defined as the coincidence between two intervals, in the limit when the stationary widths of the intervals are allowed to go to zero, as a function of the stationary distance between the two intervals. The coincidence function is given by the ratio of the conditional intensity to the unconditional intensity

$$C(z) = \frac{\mu(z)}{\mu}.$$

Formulae of this type appear in FOSS *et al.* (1993) and LANDE and STAHL (1993). Note that coincidence is not a complete description of interference but measures interference only between pairs of intervals.

Multilocus linkage analysis: When one has a panel of genetic markers, one may be interested in making a linkage map, that is, in ordering the markers and calculating genetic distances between them. Alternatively, one may be interested in using linkage analysis to locate a new marker or gene of interest on a previously determined marker map. In both cases the most informative kind of analysis is multilocus linkage analysis, in which one considers recombination patterns among all the marker loci simultaneously.

The *Drosophila* dataset of MORGAN *et al.* (1935) consists of counts of recombination events among nine marker loci on the X chromosome. The nine loci under consideration are *scute*, *echinus*, *crossveinless*, *cut*, *vermillion*, *sable*, *forked*, *carnation* and *bobbed*, and they correspond to observed fly phenotypes. The actual positions of the loci on the X chromosome are assumed to be fixed but unknown, and counts of recombination events are made based on the observed physical characteristics of the fly, which are associated with alleles at the nine loci. In *Drosophila melanogaster* recombination occurs only in females. Consider the X chromosome inherited by a fly from its mother. This X chromosome will be some mixture of the two maternal grandparental X chromosomes because of crossing over between them. For each of the nine loci, it can be determined whether the offspring fly inherited its maternal grandmother or maternal grandfather's allele at that locus. Thus, there are 2^9 possible observed outcomes. However, each outcome has a complementary outcome that is considered equivalent, in terms of recombination, namely the one in which all the grandmaternal and grandpaternal alleles are switched. Therefore, there are 2^8 possible recombination outcomes, and the number of times each occurs out of 16,136 meioses is recorded in the MORGAN *et al.* dataset.

If the nine loci are ordered, we let A_j denote the interval between loci j and $j+1$. We let $x = (x_1, \dots, x_8)$, $x_j = 0$ or 1 , $j = 1, \dots, 8$, denote the event of a

Modeling Interference

recombination in each interval A_j for which $x_j = 1$, and no recombination in each interval A_j for which $x_j = 0$. Then each of the 2^8 possible recombination outcomes would correspond to one of the 2^8 possible x 's.

In what follows we assume the order of the loci to be known. If the order were not known, the procedures described below could be repeated for each of a small number of candidate orders, and the estimated order would be the one whose maximized likelihood was highest. We assume that each event x occurs with some fixed probability p_x . These events correspond to what is observed in the data. For each x , we wish to calculate its probability p_x under each of the point process models we consider in order to fit the models to the data by maximum likelihood.

Recall that event x occurs when a given chromatid is involved in an odd number of crossovers in each of the intervals A_j for which $x_j = 1$ and an even number of crossovers in each of the intervals A_j for which $x_j = 0$. A set of related events will be denoted $y = (y_1, \dots, y_8)$, $y_j = 0$ or 1 , $j = 1, \dots, 8$. y is the event that, on the bundle of four chromatids, at least one crossover occurs in each of the intervals A_j for which $y_j = 1$ and no crossovers occur in each of the intervals A_j for which $y_j = 0$. Note that y is an event occurring on the bundle of four chromatids, whereas x refers to just one of the four chromatids. We let q_y denote the probability of the event y . The assumption of no chromatid interference gives a correspondence between these two sets of probabilities (WEINSTEIN 1936), namely

$$p_x = \sum_{y: 1 \geq y \geq x} \frac{1}{2^{y \cdot 1}} q_y \quad \text{for all } x,$$

and inverting,

$$q_y = 2^{y \cdot 1} \times \sum_{x: 1 \geq x \geq y} (-1)^{(x-y) \cdot 1} p_x \quad \text{for all } y,$$

where, for example, $y \cdot 1 = \sum_{j=1}^8 y_j$, and $1 \geq y \geq x$ means $1 \geq y_j \geq x_j$ for all j . Thus, to get the recombination probabilities under different point process models, assuming no chromatid interference, it would suffice to calculate probabilities for the simpler events, y , corresponding to zero or nonzero crossovers in the locus intervals, on the bundle of four chromatids. Once we can calculate the q_y 's in terms of the unknown parameters and from these, the p_x 's, we write down the log-likelihood as

$$\text{constant} + \sum_x a_x \log(p_x),$$

where $x = (x_1, \dots, x_8)$, $x_j = 0$ or 1 , a_x is the number of times the event x occurs in the data and where the constant is irrelevant because it does not involve the unknown parameters. We would then maximize the likelihood over the unknown parameters.

MODELS

It is sometimes assumed that the problem of modeling interference is equivalent to that of formulating a map function, *i.e.*, a function from $[0, \infty)$ to $[0, 1/2]$ that maps expected number of crossovers to chance of a recombination. This is not the case. First, a map function is not a full-fledged model. By itself, a map function does not provide probabilities for multilocus recombination events when more than three loci are involved, and so it cannot be fit to data such as that described above. In an attempt to remedy this problem, LIBERMAN and KARLIN (1984) introduced multilocus map functions, which are defined as above, but with the additional property that the map function relationship between expected number of crossovers and chance of a recombination should hold on unions of disjoint intervals, as well as on intervals. Unfortunately, these multilocus map functions correspond to an extremely limited class of models under the assumption of no chromatid interference. EVANS *et al.* (1992) showed that under this assumption the only models that have multilocus map functions are the count-location models described by KARLIN and LIBERMAN (1979) and RISCH and LANGE (1979). These models have the undesirable property that the coincidence function is constant, *i.e.*, the level of interference, as measured by coincidence does not vary at all with the genetic distance between the intervals considered. In actual data coincidence seems to be close to zero for near intervals and close to one for more distant intervals (see Figure 2). For this reason the count-location model does not appear well suited to modeling interference. It is, however, easy to fit and has been previously fit to the *Drosophila* data of MORGAN *et al.* (RISCH and LANGE, 1983). We include it here for comparison. The model will be described in more detail below.

Poisson model: COX and ISHAM (1980) provide an introduction to point processes. The simplest crossover point process model we consider is the Poisson process. This model was proposed for recombination by HALDANE (1919) and continues to be virtually the only model used in linkage analysis. The model allows no interference at all, *i.e.*, crossing over in disjoint intervals is independent, or the presence of one crossover does not alter the chance of others occurring nearby. Thus, the coincidence function has the constant value 1. In the general formulation of the Poisson model, we let μ_t denote the intensity of the process at physical location t along the bundle of four chromatids, *i.e.*,

$$\mu_t = \lim_{h \rightarrow 0} \frac{P(\text{at least 1 point in } (t, t+h])}{h}.$$

We shall not consider cases in which this limit does not exist, nor in which μ_t is not integrable. Then the chance

M. S. McPeck and T. P. Speed

of no crossovers in a given interval $[a, b]$ is $\exp(-\int_a^b \mu_t dt)$, or $e^{-\mu(b-a)}$ in the homogeneous case where $\mu_t = \mu$ for all t . To get the probability of a given event y , *i.e.*, the event of at least one crossover in each interval for which $y_i = 1$ and no crossovers in each interval for which $y_i = 0$, we multiply the appropriate probabilities for each interval using independence.

If we let $l_1, \dots, l_b$ denote the physical locations of the nine marker loci, then from recombination data the estimable quantities in the Poisson model are the genetic distances (*i.e.*, expected numbers of crossovers on a single chromatid) between the markers, namely,

$$d_1 = \frac{1}{2} \int_{l_1}^{l_2} \mu_t dt, \quad d_2 = \frac{1}{2} \int_{l_2}^{l_3} \mu_t dt, \quad \dots, \\ d_8 = \frac{1}{2} \int_{l_8}^{l_9} \mu_t dt.$$

(The factor of $1/2$ occurs because each crossover on the bundle of four chromatids is assumed to involve a given chromatid with chance $1/2$.) Thus, from the data one cannot infer anything about possible inhomogeneity of the Poisson intensity, because this cannot be separated from the unknown locations of the marker loci. Similarly, one cannot estimate the actual locations of the markers nor the physical distances between them. Associating the interval $[0, 1]$ on the real line with the chromosomal segment between the outermost loci under consideration and letting μ_t and $0 = l_1 < \dots < l_b = 1$ be as before, let $\mu = \int_0^1 \mu_t dt$ and $M(z) = \mu^{-1} \int_0^z \mu_t dt$, for all $z \in [0, 1]$. Then $M: [0, 1] \rightarrow [0, 1]$ is monotone nondecreasing and onto (*i.e.*, for every point $w \in [0, 1]$ there is some $z \in [0, 1]$ such that $w = M(z)$), hence continuous. M can be thought of as a continuous cumulative distribution function (cdf) on $[0, 1]$. If we transform each point of the Poisson process by M , then the resulting process is homogeneous Poisson with intensity μ . The transformed homogeneous process on $[0, 1]$ with marker locations $M(l_1) < \dots < M(l_b)$ and the original inhomogeneous process on $[0, 1]$ with marker locations $l_1 < \dots < l_b$ would both produce recombination data with the same distribution. Thus, without loss of generality, we may consider only homogeneous Poisson processes, that is, we let $\mu_t = \mu$ for all t . The estimable parameters of the model are the genetic distances, $d_1, \dots, d_8$, between the markers, and in terms of these the probability of the event y is

$$\prod_{i=1}^8 e^{-2d_i y_i} (1 - e^{-2d_i})^{(1-y_i)}.$$

Note that μ can be written in terms of this parametrization: $\mu = 2 \sum_{i=1}^8 d_i$.

Gamma model: The distances between crossovers in the homogeneous Poisson model are independent ex-

ponential random variables, or equivalently, gamma with shape parameter 1. One way to generalize the Poisson model would be to consider renewal processes with general gamma interarrivals, thus adding an extra parameter to the model. On the real line the stationary renewal process with gamma interarrivals can be formulated as follows. Given a point at location w , the density of the distance to the next point to the right is

$$f_I(z) = \frac{1}{\Gamma(\gamma)} \mu^\gamma z^{\gamma-1} e^{-\mu z},$$

where I stands for interarrival, $\Gamma(\gamma) = \int_0^\infty \mu^\gamma s^{\gamma-1} e^{-\mu s} ds$, and this is independent of the occurrence of any points to the left of w . The distribution of the next point to the left of w is the same; there is no directionality. The distribution of the distance to the first point to the right (or equivalently left) of w , when it is not assumed that a point has occurred at w , has density

$$f_{1st}(z) = \frac{\mu}{\Gamma(\gamma+1)} \int_z^\infty \mu^\gamma s^{\gamma-1} e^{-\mu s} ds.$$

The intensity of the process is then μ/γ . Letting $\gamma = 1$, we would get the Poisson model. The coincidence function for the gamma model is

$$C(z) = \frac{\gamma}{\mu} \sum_{n=1}^\infty f_I^n(z),$$

where $f_I^n(z) = \Gamma(\gamma n)^{-1} \mu^{\gamma n} z^{\gamma n-1} e^{-\mu z}$.

We associate the interval $[0, 1]$ with the bundle of four chromatids, and we consider the above process restricted to $[0, 1]$. Then the chance of at least one crossover occurring on the bundle of four chromatids is $\alpha = \int_0^1 f_{1st}(z) dz$, and given that at least one crossover occurs, the density of the distance from the 0 end of the bundle to the first crossover is $\alpha^{-1} f_{1st}(z)$. Given a crossover at location $w \in [0, 1]$, the chance that another crossover occurs between w and 1 is $\beta(w) = \int_0^{1-w} f_I(z) dz$. Given a crossover at location w , and given that at least one crossover occurs between w and 1, the density of the distance from w to the next crossover is $\beta(w)^{-1} f_I(z)$. Note that all distances are in terms of a metric whose relationship with physical distance is unknown but which is a constant multiple of genetic distance.

We consider the gamma model on the bundle of four chromatids, and the points are then independently thinned, each with chance $1/2$, (*i.e.*, each point has chance $1/2$ to be deleted, independently of which other points are deleted) to get the crossover point process on a single chromatid. We do not restrict consideration to the particular case of gamma with integer shape parameter, but the use of the gamma renewal process with integer shape parameter to model crossover occurrence has a long history. FISHER *et al.* (1947) proposed that

Modeling Interference

the crossover point process along a single chromatid be viewed as a renewal process with interarrival density $f(z) = \frac{1}{2}\pi \operatorname{sech}(\frac{1}{2}\pi z) \tanh(\frac{1}{2}\pi z)$. OWEN (1949, 1950) found that a gamma interarrival density with shape parameter 2 and scale parameter 2 was mathematically tractable and closely approximated the renewal process proposed by FISHER *et al.* CARTER and ROBERTSON (1952) used the gamma with shape parameter 2 as a four-strand model and applied various models of chromatid interference to get the chance of recombination across an interval for a single chromatid. PAYNE (1956) considered the gamma with integer shape parameter as a two-strand model. He compared coincidence curves for the gammas with shapes 2 and 3 to data. COBBS (1978) considered the gamma with integer shape parameter as a four-strand model, assuming no chromatid interference. He fit the model to *Neurospora* and *Drosophila* data by comparing the observed distribution of the number of crossovers in a segment (ignoring those that could not be observed) to the number predicted by the model. STAM (1979) considered various mathematical aspects of the same model, and FOSS *et al.* (1993) also considered this model, comparing the coincidence curves from the model to coincidence curves for *Drosophila* and *Neurospora* data. In this paper we calculate approximate probabilities of the different possible multilocus recombination events under the model, and we fit the model to the data by maximum likelihood. We are able to compare observed to expected frequencies of recombination events. This subsumes both the comparisons of distributions of number of events and of coincidence curves.

The gamma interarrival process with integer shape parameter has been used in the literature so often largely because of relative mathematical tractability. However, FOSS *et al.* (1993) propose it to explain certain empirical observations concerning recombination and gene conversion (a nonreciprocal exchange of genetic material between homologous chromosomes). First, gene conversion is associated with a high frequency of recombination of flanking markers (MORTIMER and FOGEL 1974). Second, gene conversions seem to occur independently in disjoint intervals, but gene conversions accompanied by recombination do not; rather, they appear to inhibit each other (MORTIMER and FOGEL 1974). The model proposed by FOSS *et al.* (see also STAHL 1979; MORTIMER and FOGEL 1974) is that initial precrossover events occur along the chromosome according to a Poisson process. Each such event results in a gene conversion and may or may not result in a crossover as well. Their model holds that every $(m+1)$ st initial event results in a crossover. If the first event has chance $1/(m+1)$ to form a crossover, then the model is equivalent to the stationary gamma interarrival model with shape parameter $m+1$.

We note that, as in the case of the Poisson model, if

we fit the stationary gamma renewal process, we are also allowing for the possibility that the true physical process may be nonstationary. The class of models covered by the analysis consists of those for which a monotone nondecreasing onto transformation $M: [0, 1] \rightarrow [0, 1]$ (i.e., a continuous, cdf on $[0, 1]$) exists that maps the model to a stationary model. The models corresponding to almost-everywhere differentiable M 's are those that can be formulated as follows. Given a point at w , the density of the distance to the first point to the right of w is

$$f^w(z) = \frac{\mu_z}{\Gamma(\gamma)} \left(\int_w^z \mu_t ds \right)^{\gamma-1} e^{-\int_w^z \mu_t dt},$$

where $\mu_t = \mu \times dM(t)/dt$. The density of the distance to the first point to the right of a given location w , when it is not assumed that a point has occurred at w , is

$$f_{1st}^w(z) = \int_{-\infty}^w \frac{\mu_t}{\gamma} f^t(z) dt.$$

The intensity of the process at location t is $\gamma^{-1}\mu_t$. Note that the coincidence function is identical to that for the stationary case.

The estimable parameters are the shape parameter γ and the genetic distances $d_1, \dots, d_g$. Note that μ can be written, in terms of this parametrization, as $\mu = 2\gamma \sum_{i=1}^g d_i$. It has not proved possible to write down an explicit expression for the likelihood of the data under the gamma model, except in the case of integer-shape parameter (ZHAO *et al.* 1995a). Instead, the likelihood has been simulated, as described below.

Hard-core model: Another generalization of the Poisson model would be to have the points follow a Poisson process but with no two points allowed to be closer than a certain fixed distance h . This is known as a hard-core model (see e.g., STOYAN *et al.* 1987). On the real line this is just a stationary renewal process with interarrivals distributed as exponential plus a constant. That is, given a point at location w , the distance to the next point to the right (equivalently, left) has density

$$f_1(z) = \begin{cases} \mu e^{-\mu(z-h)} & \text{for } z > h \\ 0 & \text{otherwise.} \end{cases}$$

The distribution of the distance to the first point to the right (equivalently, left) of w , when it is not assumed that a point has occurred at w , has density

$$f_{1st}(z) = \begin{cases} \frac{\mu}{1 + \mu h} & \text{if } z \leq h \\ \frac{\mu e^{-\mu(z-h)}}{1 + \mu h} & \text{if } z > h. \end{cases}$$

The intensity of the process is $\mu/(1 + \mu h)$. Letting h

M. S. McPeck and T. P. Speed

$= 0$, we would get the Poisson model. The coincidence function for the hard-core model is

$$C(z) = \frac{\mu}{1 + \mu h} \sum_{n=1}^{\infty} f_I^n(z),$$

where

$$f_I^n(z) = \begin{cases} \Gamma(n)^{-1} \mu^n (z - nh)^{n-1} e^{-\mu(z-nh)} & \text{if } z \geq nh \\ 0 & \text{otherwise.} \end{cases}$$

Again, we associate the interval $[0, 1]$ with the bundle of four chromatids. The restriction of this stationary renewal process to the interval $[0, 1]$ works exactly as in the gamma case. As before, this hard-core model is equivalent, in terms of the data, to any model resulting from a transformation of the interval by a continuous cdf M . The models corresponding to almost-everywhere differentiable M 's are those that can be formulated as follows. Given a point at w , the density of the distance to the first point to the right of w is

$$f_I^w(z) = \begin{cases} \mu_z e^{-\int_w^z \mu_s ds} & \text{if } z > w + h_w \\ 0 & \text{otherwise,} \end{cases}$$

where $\mu_t = \mu \times dM(t)/dt$, and $h_w > 0$ is any solution of $\int_w^{w+h_w} \mu_s ds = B$, where $B = h\mu$. The density of the distance to the first point to the right of a given location w , when it is not assumed that a point has occurred at w , is

$$f_{1st}^w(z) = \begin{cases} \frac{\mu_z}{1+B} & \text{if } z < w + h_w \\ \mu_z e^{B - \int_0^z \mu_s ds} & \text{if } z \geq w + h_w. \end{cases}$$

The intensity of the process at location t is $\mu_t/(1+B)$.

The estimable parameters are the hard-core genetic distance, $h_g = 1/2 B/(1+B)$, and the genetic distances between the points, $d_1, \dots, d_n$. Note that μ can be expressed in terms of this parametrization as $\mu = (1 - 2h_g)^{-1} \sum_{i=1}^n d_i$. As the number of loci increases, the likelihood quickly becomes very complicated, and no tractable general form has been found.

King-Mortimer model: In one version of a model suggested by KING and MORTIMER (1990), points are put down on a segment of the real line according to a Poisson process with parameter μ , just as in the original stationary Poisson model described above. Starting from time 0, each point independently waits an exponential amount of time with parameter λ before starting polymer growth, *i.e.*, the density of the time to polymer growth is $f(t) = \lambda e^{-\lambda t}$. When a polymer starts to grow from a point, it grows in both directions at a constant rate g . If a polymer from one point hits another point that has not yet started to grow, the latter point is deleted. At first glance this process appears to be nearly

identical to the one-dimensional version of the Johnson-Mehl model (JOHNSON and MEHL 1939, but MEIJERING 1953 is more readable) for random tessellations. In that model points are born over time and immediately start growing polymers (called "crystals" by JOHNSON and MEHL 1939) at rate g in both directions. In the Johnson-Mehl model the rate of birth in an unpolymerized interval $[w, w + h_1)$ during an interval of time $[t, t + h_2)$ is $\rho h_1 h_2 + o(h_1 h_2)$. That is, in any infinitesimal interval of time $[t, t + h_2)$, it is a Poisson process, with parameter ρh_2 , on the unpolymerized part of the line. In any small unpolymerized interval $[w, w + h_1)$ along the line, one waits an exponential amount of time, with parameter ρh_1 , for a point to appear and start growing. However, in the King-Mortimer model the chance that a polymer starts to grow in a small unpolymerized interval $[w, w + h_1)$ during an interval of time $[t, t + h_2)$ can be shown to be $\mu \lambda e^{-\lambda t} h_1 h_2 + o(h_1 h_2)$, *i.e.*, it varies over time, going monotonically to zero.

In the King-Mortimer model there is an initial Poisson distribution of points, then the polymerizing process takes place over time, and then there is some resulting final distribution of points. Intuitively, it is clear that the final distribution of points would not be affected if time were slowed down or speeded up by some constant multiple c , as long as the initial Poisson distribution of points remained the same. This would be equivalent to changing λ to $c\lambda$ and changing g to cg . Therefore, it is not surprising that the final distribution of points is determined by μ and g/λ only. Thus, without loss of generality, we may take $\lambda = 1$.

KING and MORTIMER (1990) choose parameter values for which their simulations appear to resemble observed data, but they do not actually fit the model. They consider the process as defined on a line segment. In that case points are more likely to occur near the ends of the segment, because there are no points outside the segment to interfere with them. This is also true in the case of the hard-core model defined on a segment. We instead consider the process as defined on the real line and then restricted to $[0, 1]$. In that case the points near the ends of the segment behave as if other points lying outside the segment could interfere with them, and the model is stationary. As before, a nonstationary model resulting from a transformation of the interval $[0, 1]$ by a continuous cdf M is indistinguishable, in terms of the data, from the stationary model. This nonstationarity is of a different type from that introduced by considering the process only on an interval, as KING and MORTIMER (1990) have done. The nonstationary models with almost-everywhere differentiable M 's covered by this analysis are those in which the initial Poisson distribution has inhomogeneous parameter $\mu_t = \mu dM(t)/dt$, and the growth rate of a polymer that has

Modeling Interference

reached position t is $g_t = g dM(t) / dt$. We note that the stationary process is not a renewal process; given a point at $w \in [0, 1]$, the distribution of the distance to the next point to the right of w is not independent of the occurrence of other points to the left of w .

To specify the stationary King-Mortimer model on $[0, 1]$, it is convenient to determine the distribution of the time when a polymer growing from somewhere outside $[0, 1]$ would reach that interval. To do this, we ignore, for the moment, points inside $[0, 1]$ that could potentially polymerize and prevent any outside polymers from ever reaching $[0, 1]$. To calculate the distribution of the time when a polymer from $(1, \infty)$ reaches 1, we let X_i , for each $i \geq 1$, be the distance between the i th and $(i - 1)$ st initial points to the right of 1. If X_i is the distance from 1 to the first initial point to the right of 1, then the X_i are independent exponential(μ). Let Z_i be the waiting time for polymer growth of the i th initial point to the right of 1. The Z_i are independent exponential(1). Let $S_i = g^{-1} \sum_{j=1}^i X_j$. Then the chance that no polymer arrives at 1 from the right before time t is

$$P(S_i > t) \\ + \sum_{i=1}^{\infty} P(S_i \leq t < S_{i+1}, S_j + Z_j > t, \text{ for all } j \leq i).$$

The i th term of the sum is found to be

$$e^{-\mu g t} \frac{1}{i!} [\mu g (1 - e^{-t})]^i,$$

and so the required probability is

$$e^{\mu g (1 - e^{-t})}.$$

One minus this quantity is the cdf of the time that a polymer first arrives from the right (equivalently, left) at any given point. This cdf is useful for simulating the process. Substituting $2g$ in for g gives the cdf of the time that a polymer first arrives at any given point from either the right or the left, ignoring the possibility that the point itself grows first and prevents the polymer's arrival. From this we can calculate the intensity of the point process, which is just the intensity, μ , of the initial Poisson process multiplied by the chance that a given initial point starts to grow before another polymer arrives. This is

$$\mu \int_0^{\infty} e^{-t} e^{2\mu g (1 - e^{-t})} dt = \mu \left[\left(\frac{e}{a} \right)^a \Gamma(a, a) - \frac{1}{a} \right],$$

where $a = 2\mu g$ and $\Gamma(x, y) = \int_0^{\infty} t^{x-1} e^{-t} dt$, known as the incomplete gamma function. The estimable parameters under this version of the King-Mortimer model are the genetic distances, $d_1, \dots, d_8$, and the expected time for a polymer to traverse the distance between

two adjacent points in the initial Poisson process, as a multiple of expected waiting time for a polymer to start growing: $T = (\mu g)^{-1} = 2a^{-1}$. Note that μ can be expressed in terms of this parametrization:

$$\mu = 2 \sum_{i=1}^8 d_i / \left[\left(\frac{eT}{2} \right)^{2/T} \Gamma(2/T, 2/T) - T/2 \right].$$

No closed-form expression is known for the likelihood of recombination data under this model, nor for the coincidence function. Both of these are obtained here by simulation.

K-M II model: The second King-Mortimer model (KING and MORTIMER 1990), which shall be called here K-M II, is the same as the previous model except that a parameter is added for termination of polymer growth, once it has started. That is, under the set-up of the previous model, each polymer now grows for an independent exponential($1/\nu$) amount of time, or until it hits another polymer, whichever happens first. The possibility of early termination of a polymer has the effect of allowing interference to operate over a smaller range than before. Interference can then be made more intense, yet more localized, than in the ordinary King-Mortimer model. For this model there is no known expression for the intensity, the likelihood of recombination data nor for the coincidence function. The parameters estimated are the genetic distances, $d_1, \dots, d_8$, the expected time for a polymer to traverse the distance between two adjacent points in the initial Poisson configuration (given that growth does not terminate up to that time), T , and the expected time to terminate polymer growth, ν , where each of these times is expressed as a multiple of the expected waiting time for a polymer to start growing.

Count-location model: In the count-location model (KARLIN and LIBERMAN 1979, called the "generalized no interference model" by RISCH and LANGE 1979) the number of crossovers is chosen according to some count distribution c_i where c_i is given by $c_0, c_1, \dots$, and $c_i = P(i \text{ crossovers})$. Given the number of crossovers, their locations are independent and identically distributed (i.i.d.) along the bundle of chromatids, according to a location distribution ν that does not vary with the count. The count-location model can be thought of as a modification of the Poisson process model, in which the number of crossovers occurring is no longer Poisson, but their locations are again i.i.d. as in the Poisson model. In the count-location model the coincidence function is constant:

$$C(z) = \sum_{i=2}^{\infty} i(i-1) c_i / \left(\sum_{j=1}^{\infty} j c_j \right)^2.$$

The version of the count-location model considered here allows no more than three crossovers to occur

M. S. McPeck and T. P. Speed

TABLE 1
Observed and expected counts under the fitted models for five loci with chi-square statistics

Recombination event	Observed data	Expected Poisson	Expected gamma	Expected hard-core	Expected King-Mortimer	Expected count-location	Expected K-M II
0000	10431	11014.1	10497.1	10837.0	10783.6	10434.0	10443.1
1000	771	596.6	738.9	682.7	686.1	777.6	767.4
0100	1579	1246.9	1537.6	1360.6	1390.0	1556.1	1560.4
0010	1221	931.3	1214.0	1058.7	1064.2	1184.7	1210.7
0001	1994	1664.1	1979.7	1685.2	1811.7	2036.4	2018.0
1100	4	67.5	1.5	13.5	25.2	16.0	7.3
1010	7	50.4	12.0	69.3	35.8	11.8	12.4
0110	4	105.4	4.0	35.1	43.1	23.4	13.2
1001	46	90.1	73.0	106.3	85.2	20.2	32.9
0101	53	188.4	68.0	208.1	138.5	42.1	48.6
0011	25	140.7	10.1	71.7	64.8	31.5	20.8
1110	1	40.3	0.1	7.9	7.8	2.3	1.3
1101	1	40.3	0.1	7.9	7.8	2.3	1.3
1011	1	40.3	0.1	7.9	7.8	2.3	1.3
0111	1	40.3	0.1	7.9	7.8	2.3	1.3
1111	1	40.3	0.1	7.9	7.8	2.3	1.3
Chi-square	—	773	51	420	267	67	17
df	—	8	7	7	7	6	6

Triple and quadruple recombination events are quite rare. These events were pooled in the analysis because accurate probabilities for the events could not be computed by simulation.

simultaneously, *i.e.*, $\sum_{i=1}^3 c_i = 1$ and $c_i = 0$ for $i > 3$. The location distribution ν may be taken to be uniform without loss of generality, because the class of models covered by this analysis will include all those with location distributions having a continuous cdf. The intensity of the process is $\sum_{i=0}^3 i c_i$. The model can be parameterized in terms of the genetic distances, $d_1, \dots, d_8$, and c_0 and c_1 . The likelihood can be calculated explicitly in terms of these parameters.

To fit any of these models to recombination data, it is necessary to be able to calculate q_y for any y , that is, the probability of any combination of zero and nonzero crossover counts in the eight marker intervals, as described above. This is trivial for the Poisson model and the count-location model. For the other models this involves integrating the density over all possible realizations compatible with y , assuming that the density is known or can be computed. So far, these integrals appear virtually impossible to do exactly, except in the special case of gamma with integer shape parameter. As a result, they have been calculated here by Monte Carlo methods. For the Poisson, gamma, count-location and K-M II models, the recombination probabilities were computed for events among all nine loci, whereas for the other models only a subset of five of the nine loci were used.

For a given model the probability q_y depends not only on the model parameters but also on the relative locations of the loci. For each choice of model parameters, a large number n of realizations of the point pro-

cess are simulated. Using these simulations, the desired probability q_y can be estimated, for any set of locus locations, by the observed frequency of the event y . From the q 's the estimates of the probabilities of the observed recombination events can be computed, and from these, the estimated likelihood of the observed data is computed. Recombination events in which three or more recombinations took place were so rare in this data that their probabilities are very difficult to estimate. These rare events were pooled in this analysis. Holding the point process parameters fixed, the likelihood can then be numerically maximized over the locus locations by the Nelder-Mead search algorithm (NELDER and MEAD 1965). This maximization is done

TABLE 2
Fitted point process parameters for five loci

Model	Parameter(s) ^a
Gamma	$\gamma = 6.41$ (0.280)
Hard-core	$h_c = 0.116$ (0.004)
King-Mortimer I	$T = 0.038$ (0.005)
Count-location	$c_0 = 0.311$ (0.005), $c_1 = 0.641$ (0.007) (redundant parameters: $c_2 = 0.0448$ (0.005), $c_3 = 0.0032$ (0.002))
K-M II	$T = 1.01 \times 10^{-4}$ (2.44×10^{-5}), $\nu = 0.244$ (0.187)

^a Estimated standard deviations in parentheses.

Modeling Interference

TABLE 3

Fitted genetic distances between adjacent markers for five loci

Genetic distances	Poisson	Gamma	Hard-core	K-M I	Count-location	K-M II
(sc)-(ec)	0.054 (0.002)	0.051 (0.002)	0.055 (0.002)	0.052 (0.002)	0.052 (0.002)	0.050 (0.003)
(ec)-(cv)	0.114 (0.003)	0.100 (0.003)	0.100 (0.003)	0.101 (0.002)	0.104 (0.002)	0.102 (0.005)
(cv)-(ct)	0.085 (0.003)	0.077 (0.003)	0.077 (0.002)	0.076 (0.002)	0.079 (0.002)	0.079 (0.003)
(ct)-(v)	0.152 (0.004)	0.132 (0.004)	0.129 (0.003)	0.135 (0.003)	0.136 (0.003)	0.132 (0.007)

Estimated standard deviations in parentheses.

using the same set of n simulated realizations. For each choice of model parameters, the process can be repeated, generating a new set of n realizations and maximizing the estimated likelihood over the locus locations. The Nelder-Mead search algorithm can then be used to maximize over the model parameters. This is an extension of a method described in DIGGLE and GRATTON (1984). The number of realizations simulated, n , was taken to be 160,000 initially and was increased to 960,000 as a maximum was approached.

There are standard algorithms for simulating from exponential and gamma distributions (see RIPLEY 1987). In addition to these, we must be able to simulate from the first-point distributions of the gamma and the hard-core models, and also, in the King-Mortimer model, the distribution of the first time that a polymer from outside the $[0, 1]$ interval first hits 0 from below (or equivalently, hits 1 from above).

In the gamma case if the shape parameter is an integer, say γ , then a random variable from the first-point distribution can be realized by simulating the sum of k independent exponential random variables, where k is chosen uniformly on $1, \dots, \gamma$. For noninteger γ 's one can do rejection sampling using the next highest integer (see RIPLEY 1987 for a discussion of rejection sampling). For the hard-core model with hard-core distance h and Poisson parameter μ , the first point will be uniform on $[0, h]$ with probability $h/(1/\mu + h)$ and exponentially distributed on $[h, \infty)$ with probability $(1/\mu)/(1/\mu + h)$.

To simulate from the King-Mortimer model on an

interval, one must be able to simulate the distribution of the first time that a polymer from outside the interval reaches the interval from the left, or identically, from the right. The distribution, which was derived in the last section, can be simulated by rejection sampling using exponential(μ) for $\mu g \leq 1$ and exponential($1/g$) for $\mu g \geq 1$. For the K-M II model even this distribution is not known but is simulated by looking at the process over a very large interval and then restricting to a small interval, where edge effects may be assumed to be negligible.

RESULTS

SPEED *et al.* (1992) found that as a result of assuming no chromatid interference, certain constraints must be satisfied by the probabilities of recombination events. Namely,

$$0 \leq \sum_{y=x} (-1)^{(y-x)-1} p_y \quad \text{for all } x.$$

Furthermore, for any set of p 's that satisfies these constraints, there is a no-chromatid-interference model that is compatible with them. It is interesting to note that for the five loci considered, the observed frequencies of the recombination events satisfy all eight nontrivial constraints of the ones required above. Only a single observation prevents this from being true for the 37 nontrivial constraints on all nine loci. Of course, this is not proof that the assumption of no chromatid interference is correct but simply that the data are not incompatible with this assumption.

TABLE 4

Point-process parameters and chi-square statistics for nine-locus fit

Model	Parameter(s) ^a	Chi-square	d.f.
Poisson	None	1672	30
Gamma	$\gamma = 4.94$ (0.124)	107	29
Count-location	$a_0 = 0.057$ (0.008), $c_1 = 0.434$ (0.017)	889	28
	redundant parameters: $c_2 = 0.434$ (0.017), $c_3 = 0.075$ (0.008)		
K-M II	$T = 1.84 \times 10^{-4}$ (7.30×10^{-6}), $v = 1.08$ (0.337)	294	28

^a Estimated standard deviations in parentheses.

M. S. McPeck and T. P. Speed

TABLE 5
Genetic distances for nine-locus fit

Genetic distances	Poisson	Gamma	Count-location	K-M II
(sc)-(er)	0.054 (0.002)	0.053 (0.002)	0.053 (0.002)	0.046 (0.002)
(er)-(cv)	0.114 (0.003)	0.098 (0.002)	0.107 (0.003)	0.096 (0.004)
(cv)-(cl)	0.085 (0.003)	0.075 (0.002)	0.082 (0.002)	0.078 (0.003)
(cl)-(v)	0.152 (0.004)	0.133 (0.002)	0.141 (0.003)	0.141 (0.005)
(v)-(s)	0.089 (0.003)	0.085 (0.002)	0.086 (0.002)	0.088 (0.004)
(s)-(f)	0.184 (0.004)	0.157 (0.003)	0.171 (0.004)	0.163 (0.007)
(f)-(car)	0.081 (0.002)	0.075 (0.003)	0.078 (0.002)	0.071 (0.003)
(car)-(bb)	0.046 (0.002)	0.044 (0.002)	0.044 (0.002)	0.039 (0.002)

Table 1 shows Pearson chi-square statistics for the five-locus models, and Table 4 shows them for the nine-locus models. In all cases the *p* values are smaller than 0.01, implying significant misfit of all models. However, with the addition of a single parameter, the chi-square statistic in the five-locus case is brought down from 773 (Poisson model) to 51 (gamma model), and in the nine-locus case from 1672 (Poisson model) to 107 (gamma model). This represents a tremendous improvement. In the five-locus case the best-fitting model is the K-M II model, which adds two parameters to the

Poisson model and has a chi-square statistic of 17, compared to 773 for the Poisson model. The K-M II model is outperformed by the gamma model in the nine-locus case. One interpretation is that K-M II does a reasonably good job of modeling interference between nearby markers but that it is not flexible enough to simultaneously model both close and medium-range interference, as required in the nine-locus dataset. Tables 2 and 3 contain fitted parameter values, with estimated standard deviations, for the five-locus models, and Tables 4 and 5 show them for the nine-locus models.

FIGURE 1.—Distribution of the number of crossovers among the first five loci of the MORGAN *et al.* (1935) dataset, for each of the models under consideration, relative to the data distribution. (a) The chance of zero crossovers for each of the models minus 0.310, the proportion of cases in which zero recombinations occurred among the first five loci in the data. (b–d) Similar bar graphs for the probabilities of 1, 2 and 3 or more crossovers, respectively. The data proportions to which these numbers are compared are 0.656, 0.0337 and 0.000496, respectively.

FIGURE 2.—Plots of coincidence *vs.* genetic distance for the five-locus models. The points represent coincidence calculated from the data for each pair of atomic intervals among the nine loci. The corresponding genetic distance for each pair is calculated between the midpoints of the two intervals. The curves represent five-locus model coincidences. a is the Poisson model, b is the hard-core model, c the King-Mortimer model, d the gamma model, e the count-location model and f the K-M II model.

The results shown for the Poisson are from a calculation rather than from simulation, but the simulated results are virtually identical. In the case of the gamma model, the likelihood can actually be calculated in the case of integer-shape parameter (ZHAO *et al.* 1995a). In the nine-locus case using this likelihood calculation, the best fit is obtained using shape parameter 5, which agrees with the results given here for the nine-locus case. This would correspond to four gene conversion events between each crossover in the framework described in FOSS *et al.* (1993). The count-location model has previously been fit to the MORGAN *et al.* (1935) dataset by RUSCH and LANGE (1983), using a calculated, rather than simulated, likelihood. In their calculation they did not pool events in which three or more recombinations occurred. Thus, they maximized a slightly different likelihood than the one maximized here. Still, the results are very close to those obtained here. Note that the estimates of genetic distances and their standard deviations were fairly similar for all models. The Poisson model gives very similar esti-

mates to the other models for small genetic distances (<10 cM) but overestimates larger genetic distances relative to the other models.

The SDs were estimated by simply taking numerical derivatives of the recombination probabilities (p_r 's in the notation given above) with respect to the estimated parameters and using these to calculate the Fisher information matrix. A better approach, not taken here, would be to smooth the likelihood surface before calculating derivatives.

The Pearson chi-square statistic measures the overall goodness-of-fit of each model. We consider two additional criteria that do not indicate goodness-of-fit as directly as the chi-square does but that are easy to interpret and may be useful for understanding the performance of the models. First, we consider the number of crossovers predicted by the models. Figure 1 shows that the five-locus Poisson model predicts too few single crossover events and too many of all other types of events. This shows up in Table 1 as an underestimate

M. S. McPeck and T. P. Speed

FIGURE 3.—Distribution of the number of crossovers among all nine loci of the MORGAN *et al.* (1935) dataset for each of the models under consideration relative to the data distribution. (a) The chance of zero crossovers for each of the models minus 0.056, the proportion of cases in which zero recombinations occurred in the data. (b–e) Similar bar graphs for the probabilities of 1, 2, 3 and 4 or more crossovers, respectively. The data proportions to which these numbers are compared are 0.485, 0.429, 0.030 and 0, respectively.

of the number of single recombination events among the five loci and an overestimate of all other types of recombination events. Figure 3 shows that in the nine-locus case, the Poisson model predicts too few single and double crossover events and too many of all other types. The gamma, count-location and K-M II models all do well at matching the distribution of the number of crossovers in the data, although the K-M II model does not do nearly as well with nine loci and with five.

The other aspect of the models which we shall consider is the coincidence function which is discussed above. Figure 2 shows coincidence *vs.* genetic distance for the five-locus models, and Figure 4 shows it for the nine-locus models. Recall that coincidence is a measure of the degree to which one crossover inhibits the formation of another nearby. For the Poisson model coincidence is one, reflecting no interference. The coincidence curves for the models have been calculated (or simulated, in the case of the King-Mortimer and K-M II models) using infinitesimal intervals. The coincidence curves of the gamma model and the King-Mortimer

model give the best approximation to the data in the five-locus case, whereas the gamma and K-M II models do best in the nine-locus case.

The strange shape of the coincidence curve for the hard-core model can be explained as follows: in the hard-core model, no points are allowed within a fixed distance of the crossover point that is assumed to lie at 0 (for the purpose of making the coincidence curve). Immediately after that fixed distance has been surpassed, one is waiting an exponential amount of time for the next point, whereas in an interval chosen at random the intensity is less than exponential because there is a chance of another crossover already having occurred nearby. This explains the spike in the coincidence curve for the hard-core model, and the hump in the coincidence curve for the gamma model can be explained similarly. As the distance from the given crossover gets close to the mode of the gamma interarrival distribution, one is more likely to have another crossover than one would be to have a crossover in a given interval chosen at random.

FIGURE 4.—Plots of coincidence *vs.* genetic distance for the nine-locus models. The points represent coincidence calculated from the data for each pair of atomic intervals among the nine loci. The corresponding genetic distance for each pair is calculated between the midpoints of the two intervals. The curves represent nine-locus model coincidences. a is the K-M II model, b is the count-location model and c is the gamma model.

It should be noted that the coincidence curve for the count-location model is constant, even if the distribution of points is i.i.d. but nonuniform. The reason is that the *x*-axis represents genetic, not physical distance, that is, it represents the expected number of crossovers and so is automatically rescaled to be uniform. Note also the anomaly that the K-M II model fits the data better (according to chi-square) in the five-locus case than in the nine-locus case, but the coincidence curve resembles the data much more closely in the nine-locus case.

DISCUSSION

Although none of the models fit the data, several of the models were able to capture certain aspects of the data. The best overall model of the ones considered was the renewal model with gamma interarrivals. It gave reasonable estimates of the numbers of crossovers and matched the interference pattern in the data as measured by the coincidence. In terms of the chi-square statistic, the gamma model represents a huge improvement over the Poisson model. Furthermore, in the case of integer shape parameter, the gamma model may prove tractable enough to use in mapping applications.

This work represents a first attempt to systematically

fit a range of point process models to recombination data. A suitable model, such as the gamma model, may be of great use in assessing the impact of interference on conventional linkage analyses and genetic mapping. It may even be possible to apply such a model in genetic mapping, thereby making more efficient use of the data.

We are grateful to B. D. RIPLEY for his advice on the Monte Carlo method and to F. W. STAHL for helpful discussions. This work was supported by National Science Foundation grants DMS-880237 and DMS-9113527.

LITERATURE CITED

- CARTER, T. C., and A. ROBERTSON, 1952 A mathematical treatment of genetical recombination using a four-strand model. *Proc. R. Soc. Lond. Ser. B* **139**: 410–426.
- COBBS, G., 1978 Renewal process approach to the theory of genetic linkage: case of no chromatid interference. *Genetics* **89**: 563–581.
- COX, D. R., and V. ISHAM, 1980 *Point Processes*. Chapman and Hall, New York.
- DIGGLE, P. J., and R. J. GRATTON, 1984 Monte Carlo methods of inference for implicit statistical models. *J. Roy. Statist. Soc. B* **46**: 193–227.
- EVANS, S. N., M. S. MCPHEE and T. P. SPEED, 1992 A characterisation of crossover models that possess map functions. *Theor. Popul. Biol.* **43**: 80–90.
- FISHER, R. A., M. F. LYON and A. R. G. OWEN, 1947 The sex chromosome in the house mouse. *Heredity* **1**: 335–365.

M. S. McPeck and T. P. Speed

- 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**: 38–45.
- HALDANE, J. B. S., 1919 The combination of linkage values, and the calculation of distances between the loci of linked factors. *J. Genet.* **8**: 299–309.
- JOHNSON, W. A., and R. F. MEHL, 1939 Reaction kinetics in processes of nucleation and growth. *Trans. A.I.M.M.E.* **135**: 416–458.
- KARLIN, S., and U. LIBERMAN, 1979 A natural class of multilocus recombination processes and related measures of crossover interference. *Adv. Appl. Probab.* **11**: 479–501.
- LANDE, R., and F. W. STAHL, 1993 Chiasma interference and the distribution of exchanges in *Drosophila melanogaster*. *Cold Spring Harbor Symp. Quant. Biol.* **58**: 543–552.
- LIBERMAN, U., and S. KARLIN, 1984 Theoretical models of genetic map functions. *Theor. Popul. Biol.* **25**: 331–346.
- KING, J. S., and R. K. MORTIMER, 1990 A polymerization model of chiasma interference and corresponding computer simulation. *Genetics* **126**: 1127–1138.
- MEIJERING, J. L., 1953 Interface area, edge length, and number of vertices in crystal aggregates with random nucleation. *Philips Res. Rep.* **8**: 270–290.
- MORGAN, T. H., C. B. BRIDGES and J. SCHULTZ, 1935 Constitution of the germinal material in relation to heredity. *Carnegie Instit. Washington Publ.* **34**: 284–291.
- MORTIMER, R. K., and S. FOGEL, 1974 Genetical interference and gene conversion, pp. 263–275 in *Mechanisms in Recombination*, edited by R. F. GRELL. Plenum, New York.
- MULLER, H. J., 1916 The mechanism of crossing-over. *Am. Nat.* **50**: 193–221, 284–305, 350–366, 421–434.
- NELDER, J. A., and R. MEAD, 1965 A simplex method for function minimisation. *Comput. J.* **7**: 303–313.
- OWEN, A. R. G., 1949 The theory of genetical recombination. I. Long-chromosome arms. *Proc. R. Soc. Lond. Ser. B* **136**: 67–94.
- OWEN, A. R. G., 1950 The theory of genetical recombination. *Adv. Genet.* **3**: 117–157.
- PASCOE, L., and N. E. MORTON, 1987 The use of map functions in multipoint mapping. *Am. J. Hum. Genet.* **40**: 174–188.
- PAYNE, L. C., 1956 The theory of genetical recombination: a general formulation for a certain class of intercept length distributions appropriate to the discussion of multiple linkage. *Proc. R. Soc. Lond. Ser. B* **144**: 528–544.
- RIPLEY, B. D. R., 1987 *Stochastic Simulation*. John Wiley and Sons, New York.
- RISCH, N., and K. LANGE, 1979 An alternative model of recombination and interference. *Ann. Hum. Genet.* **43**: 61–70.
- RISCH, N., and K. LANGE, 1983 Statistical analysis of multilocus recombination. *Biometrics* **39**: 949–963.
- SPEED, T. P., M. S. MCPECK and S. N. EVANS, 1992 Robustness of the no-interference model for ordering genetic markers. *Proc. Natl. Acad. Sci. USA* **89**: 3103–3106.
- STAHL, F. W., 1979 *Genetic Recombination. Thinking About it in Phage and Fungi*. W. H. Freeman, San Francisco.
- STAM, P., 1979 Interference in genetic crossing over and chromosome mapping. *Genetics* **92**: 573–594.
- STOYAN, D., W. S. KENDALL and J. MECKE, 1987 *Stochastic Geometry and its Applications*. Wiley, New York.
- STURTEVANT, A. H., 1915 The behavior of the chromosomes as studied through linkage. *Z. Indukt. Abstammungs. Vererbungsleh.* **13**: 234–287.
- WEINSTEIN, A., 1936 The theory of multiple-strand crossing over. *Genetics* **21**: 155–199.
- ZHAO, H., T. P. SPEED and M. S. MCPECK, 1995a Statistical analysis of crossover interference using the chi-square model. *Genetics* **139**: 000–000.
- ZHAO, H., M. S. MCPECK and T. P. SPEED, 1995b Statistical analysis of chromatid interference. *Genetics* **139**: 000–000.

Communicating editor: B. S. WEIR

Statistical Analysis of Crossover Interference Using the Chi-Square Model

Hongyu Zhao, Terence P. Speed and Mary Sara McPeck¹

Department of Statistics, University of California, Berkeley, California 94720

Manuscript received December 17, 1993

Accepted for publication October 19, 1994

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 that 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 that have only a small number of observations. Finally, comparisons are made between the chi-square model and some other tractable models in the literature.

CROSSOVER interference has been observed in almost all organisms studied, although there is little consistent evidence of chromatid interference even within the same organism (ZHAO *et al.* 1995). In what follows we assume no chromatid interference (NCI).

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 NCI, 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 $(i_1 i_2 \dots i_n)$, where $i_i = 0, 1, 2$ corresponds to parental ditype, tetratype and nonparental ditype, respectively, between $\cdot \cdot l_j$ and $\cdot \cdot l_{j+1}$.

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 nonsister chromatids to be involved in each crossover.

The chi-square model for crossovers has a long history; see BAILEY (1961). McPECK and SPEED (1995) 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. Whereas it has generally been of interest due to its mathematical tractability, the chi-square model has also been suggested as a plausible 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 $\dots Cx(Co)^m Cx(Co)^m \dots$. 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.

Corresponding author: Hongyu Zhao, Department of Statistics, University of California, Berkeley, CA 94720.
E-mail: zhao@stat.berkeley.edu

¹ Present address: Department of Statistics, University of Chicago, 5734 University Ave., Chicago, IL 60637.
E-mail: mcpeek@galton.uchicago.edu

H. Zhao, T. P. Speed and M. S. McPeck

ESTIMATION UNDER THE CHI-SQUARE MODEL

Given a set of markers $\cdot i_1, \dots, \cdot i_{n+1}$ 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, $x_1, x_2, \dots, x_n$, 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, j th entry is $e^{-y} y^{pk+j-1} / (pk + j - i)!$. Then the probability of k_j crossovers between $\cdot i_j$ and $\cdot i_{j+1}$, $j = 1, \dots, n$, is (see APPENDIX, Lemma):

$$\frac{1}{p} \mathbf{1D}_{k_1}(y_1) \mathbf{D}_{k_2}(y_2) \cdots \mathbf{D}_{k_n}(y_n) \mathbf{1}', \quad \text{where}$$

$$\mathbf{1} = (1, 1, \dots, 1).$$

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 $(i_1 i_2 \cdots i_n)$. We consider the two cases separately.

For single spore data, given two consecutive markers $\cdot i_j$ and $\cdot i_{j+1}$, we can observe a recombination or nonrecombination between them. If no crossovers occur between $\cdot i_j$ and $\cdot i_{j+1}$, no strand in the bundle will show any recombination between these markers. MATHER (1935) proved that under the assumption of NCI, 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 $(i_1 i_2 \cdots i_n)$, where $i_j = 0$ or 1. Define

$$\mathbf{N}_j = \mathbf{D}_0(y_j) + \frac{1}{2} \sum_{s \geq 1} \mathbf{D}_s(y_j)$$

$$\mathbf{R}_j = \frac{1}{2} \sum_{s \geq 1} \mathbf{D}_s(y_j).$$

Then the probability of recombination pattern $(i_1 i_2 \cdots i_n)$ is (see APPENDIX, Theorem 1)

$$P(i_1 i_2 \cdots i_n) = \frac{1}{p} \mathbf{1M}_1 \mathbf{M}_2 \cdots \mathbf{M}_n \mathbf{1}',$$

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 diatype, tetratype and nonparental diatype, respectively, between a fixed pair of markers. Given $k \geq 1$ crossovers between two loci, under the assumption of NCI, the conditional probabilities $p_0^{(k)}, p_1^{(k)}$ and $p_2^{(k)}$ of a tetrad being of parental diatype, tetratype and nonparental diatype, respectively, are given by MATHER (1935):

$$p_0^{(k)} = \frac{1}{3} \left(\frac{1}{2} + (-\frac{1}{2})^k \right)$$

$$p_1^{(k)} = \frac{2}{3} \left(1 - (-\frac{1}{2})^k \right)$$

$$p_2^{(k)} = \frac{1}{3} \left(\frac{1}{2} + (-\frac{1}{2})^k \right).$$

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

$$\mathbf{P}_j = \mathbf{D}_0(y_j) + \sum_{s \geq 2} \frac{1}{3} \left(\frac{1}{2} + (-\frac{1}{2})^s \right) \mathbf{D}_s(y_j)$$

$$\mathbf{T}_j = \mathbf{D}_1(y_j) + \sum_{s \geq 2} \frac{2}{3} \left(1 - (-\frac{1}{2})^s \right) \mathbf{D}_s(y_j)$$

$$\mathbf{N}_j = \sum_{s \geq 2} \frac{1}{3} \left(\frac{1}{2} + (-\frac{1}{2})^s \right) \mathbf{D}_s(y_j).$$

Then the probability of the tetrad pattern $(i_1 i_2 \cdots i_n)$ can be written as (see APPENDIX, Theorem 2)

$$P(i_1 i_2 \cdots i_n) = \frac{1}{p} \mathbf{1M}_1 \mathbf{M}_2 \cdots \mathbf{M}_n \mathbf{1}',$$

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 and based upon the above formulae, the likelihood of the observations, up to a constant factor, can be calculated in terms of the parameters as $\prod P(i_1 i_2 \cdots i_n)^{x_{i_1 i_2 \cdots i_n}}$, where $x_{i_1 i_2 \cdots i_n}$ is the observed frequency of single spores or tetrads with pattern $(i_1 i_2 \cdots i_n)$. 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 $n \rightarrow \infty$,

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

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 that 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 (1936) and MORGAN *et al.* (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

Chi-Square Model for Interference

TABLE 1

Observed and expected counts under the $Cx(Co)^4$ model

<i>sc-e</i>	<i>e-cv</i>	<i>cv-cl</i>	<i>cl-v</i>	<i>v-g</i>	<i>gf</i>	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
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	0	1	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

$Cx(Co)^4$ 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. 0, no recombination; 1, recombination. The estimated P value is <0.001 . Data of WEINSTEIN (1936).

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 *et al.*'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 (1936). The $Cx(Co)^4$ model

TABLE 2

Estimated genetic distances with standard errors under the $Cx(Co)^4$ model

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

$Cx(Co)^4$ model gives the best fit among $Cx(Co)^m$ models. The estimated P value is <0.001 . Data of MORGAN *et al.* (1935).

again gives the best fit to the data among the $Cx(Co)^m$ class. The results are given in Table 2. MCPECK and SPEED (1995) fit a broader class of models, in which m is allowed to be noninteger, to the MORGAN *et al.* dataset and estimated $m = 3.94$, which agrees very well with the integer value $m = 4$.

Among the loci that 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 nonhomogeneity 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). However, 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 *N. 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

H. Zhao, T. P. Speed and M. S. McPeck

TABLE 3

Observed and expected counts under the $Cx(Co)^2$ model

<i>cr-th</i>	<i>th-ni</i>	<i>ni-au</i>	<i>au-ni</i>	<i>ni-os</i>	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
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

$Cx(Co)^2$ 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. 0, parental ditype; 1, tetratype; 2, nonparental ditype. The estimated P value is 0.69. Data of PERKINS (1962).

$Cx(Co)^m$ 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 *et al.* (1962) consider seven markers on linkage group 1 of *Neurospora crassa*, three on the left arm and four on the right arm. Altogether 2920 offspring 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 *S. cerevisiae*, but, perhaps because of 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 that 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, whereas $k > 1$ 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_2 by FOSS *et al.* (1993). SNOW fits BARRATT *et al.*'s model to tetrad data involving 34 pairs of markers on 12 chromosomes in *S. 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 multilo-

TABLE 4

Estimated genetic distances with standard errors and estimated P values

	Experiment 1	Experiment 2	Experiment 3	Experiment 4
<i>histI-inos</i>	4.33 (0.33)	7.07 (0.40)	6.50 (0.35)	5.40 (0.29)
<i>inos-bis</i>	4.97 (0.36)	5.27 (0.35)	6.02 (0.36)	6.23 (0.32)
<i>bis-pab2</i>	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

$Cx(Co)^2$ is the best model among $Cx(Co)^m$ models for all four experiments. Data of STRICKLAND (1961).

Chi-Square Model for Interference

TABLE 5
Data of *S. cerevisiae* analyzed in SNOW (1979)

Chromosome	Gene pair	<i>m</i>	<i>n</i>	x_{χ^2}	SE	<i>p</i>	x_{ϵ}	<i>k</i>
2	<i>cyh1-gal1</i>	1	146	19.9	2.6	0.98	23.0	0.337
	<i>gal1-lys2</i>	2	383	52.7	3.5	0.71	79.5	0.488
	<i>lys2-try1</i>	3	335	34.8	1.9	0.77	50.0	0.245
	<i>try1-his7</i>	5	127	42.2	3.5	0.90	70.5	0.194
	<i>SUP45-lys2</i>	2	104	25.7	3.1	0.89	32.7	0.258
3	<i>SUP45-tyr1</i>	0	105	17.4	3.4	0.64	17.4	1.817
	<i>his4-mat1</i>	2	278	39.8	2.8	0.77	56.3	0.294
	<i>his4-leu2</i>	1	521	17.4	1.2	0.85	19.8	0.370
	<i>leu2-mat1</i>	1	481	35.8	2.2	0.73	44.6	0.429
	<i>mat1-thr4</i>	1	434	21.3	1.6	0.90	24.8	0.404
4	<i>thr4-MAL2</i>	2	286	29.2	2.2	0.82	38.2	0.294
	<i>SUP35-aro1</i>	5	101	46.1	7.5	0.99	80.9	0.204
	<i>trp1-cdc2</i>	1	205	18.2	2.1	0.92	20.8	0.284
5	<i>his1-trp2</i>	2	215	24.5	2.2	0.79	30.8	0.277
	<i>ura3-hom3</i>	5	206	34.7	2.3	0.99	53.2	0.115
6	<i>SUP11-his2</i>	1	105	22.0	3.2	0.97	25.6	0.386
7	<i>trp5-ade6</i>	0	106	82.1	4.9	0.62	83.3	1.333
	<i>ade5-tyr3</i>	1	166	73.3	10.	0.87	101.	0.764
	<i>try3-lys5</i>	0	162	8.0	1.7	0.89	0.80	2.355
	<i>cyh2-trp5</i>	3	160	44.8	3.9	0.88	70.5	0.265
	<i>leu1-ade6</i>	2	507	34.3	1.8	0.89	46.7	0.271
8	<i>MAL1-ade3</i>	1	138	52.0	6.4	0.78	68.5	0.495
	<i>pet1-CUP1</i>	3	232	46.7	3.4	0.80	74.7	0.271
	<i>thr1-CUP1</i>	3	486	24.4	1.3	0.96	31.8	0.112
	<i>CUP1-pet1</i>	3	240	36.0	2.6	0.98	52.4	0.210
9	<i>his6-lys1</i>	1	411	47.1	3.3	0.55	61.1	0.461
10	<i>SUP4-SUP7</i>	2	179	53.1	5.4	0.78	80.1	0.498
11	<i>met14-met1</i>	1	109	46.2	6.1	0.96	59.7	0.545
15	<i>met1-MAL4</i>	1	133	28.0	3.4	0.70	33.5	0.633
	<i>ser1-ade2</i>	4	210	27.4	2.0	0.98	37.5	0.098
	<i>ade2-cyh4</i>	1	236	33.9	3.0	0.78	41.9	0.397
17	<i>pet17-ade2</i>	2	232	48.8	3.9	0.79	72.8	0.367
	<i>met2-pha2</i>	2	146	36.8	3.5	0.75	50.6	0.386
	<i>pet2-pha2</i>	1	151	50.7	5.9	0.93	66.4	0.551

m, estimated *m* in the $Cx(Co)^m$ model; *n*, sample size; x_{χ^2} , estimated genetic distance from the $Cx(Co)^m$ model; SE, standard error for the estimated genetic distance; *p*, estimated *P* value; x_{ϵ} , estimated genetic distance in SNOW (1979); *k*, estimated *k* in BARRATT *et al.*'s model in SNOW (1979).

cus *S. 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 14 experiments, and *met13*, *trp5*, *cyh3* and *leu1* were used in the other 11 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 *S. cerevisiae* the ratio of gene conversions to cross-overs is ~ 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 ~ 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 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) that were from KOHLI *et al.* (1977) and those provided by Dr. P. MUNZ (personal communication). Two-point

H. Zhao, T. P. Speed and M. S. McPeck

TABLE 6
Data of *S. pombe* analyzed in SNOW (1979)

Chromosome	Gene pair	<i>m</i>	<i>n</i>	χ^2	SE	<i>p</i>	x_i	<i>k</i>
1	<i>cyh1-cdc1</i>	1	142	58.0	7.5	0.75	77.7	0.545
	<i>cdc1-leu2</i>	1	124	51.7	7.0	0.73	67.7	0.698
	<i>his1-leu2</i>	0	102	11.5	2.7	0.22	11.5	4.691
	<i>sup3-aro3</i>	0	170	75.1	10.	0.56	76.2	1.350
	<i>ura2-ade2</i>	0	364	35.8	3.1	0.75	35.7	0.888
	<i>ade2-ade4</i>	0	290	122.	20.	0.72	121.	0.869
	<i>lys3-ura1</i>	0	692	19.9	1.4	0.17	19.9	1.771
	<i>ura1-lys5</i>	0	131	38.4	5.5	0.63	32.8	0.716
	<i>pro1-ade3</i>	0	100	49.6	8.1	0.92	49.7	1.087
	<i>ade3-pro2</i>	0	589	96.9	8.5	0.94	96.7	0.985
	<i>ade7-ura5</i>	1	556	12.3	1.0	0.94	13.4	0.229
	<i>ade7-his3</i>	0	392	69.5	6.1	0.72	69.8	1.117
	<i>gtu1-his3</i>	0	212	66.8	8.0	0.81	66.5	0.911
	<i>his3-mat1</i>	1	728	80.1	6.0	0.49	111.	0.837
2	<i>tsl24-mat1</i>	0	451	96.2	9.8	0.48	97.6	1.261
	<i>leu1-his5</i>	0	371	28.2	2.5	0.91	28.2	0.951
	<i>his5-leu3</i>	0	100	62.1	11.	0.94	61.8	0.921
	<i>ade1-his4</i>	0	498	42.5	3.0	0.48	42.7	1.249
	<i>his4-trp1</i>	0	128	109.	23.	0.75	111.	1.227
	<i>ade8-arg4</i>	0	185	31.3	4.0	0.87	31.2	0.894
	<i>ade10-fur1</i>	0	142	17.6	3.0	0.70	17.5	0.581
	<i>ade10-ade6</i>	0	202	31.8	3.8	0.98	31.7	0.974
	<i>fur1-sin2</i>	0	206	25.4	3.1	0.90	25.3	0.892
	<i>fur1-min5</i>	0	199	22.4	2.9	0.85	22.4	1.177
3	<i>ade6-min5</i>	0	337	4.5	0.8	0.05	4.4	7.693
	<i>tsl5-arg1</i>	5	48	5.16	7.3	0.98	42.3	0.302
	<i>arg1-ade5</i>	1	157	6.14	7.8	0.99	82.7	0.645
	<i>arg1-aro4</i>	0	67	56.3	12.	0.34	58.0	2.523
	<i>trp3-aro4</i>	0	126	19.7	3.4	0.90	19.6	1.151
	<i>ade5-uee1</i>	0	67	69.1	15.	0.85	68.5	0.850

Symbols in this table are the same as those of Table 5.

cross tetrad data on 30 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 < 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 that is described by MCPPECK and SPEED (1995).

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 1231 fully classifiable asci of *A. nidulans* from three separate experiments. Crosses number 1 and 3 cover the same three intervals in the right arm of the *Bl* chromosome, whereas cross 2 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.

Chi-Square Model for Interference

TABLE 7

Observed and expected counts under the $Cx(Co)^2$ model				
Crossovers	Expected (male)	Observed (male)	Expected (female)	Observed (female)
0	206	196	123	130
1	115	131	102	93
2	11	4	20	20
3	0	1	1	3

When m is assumed to be the same for both male and female, $Cx(Co)^2$ gives the best fit. Data of McINNIS *et al.* (1993).

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 ~ 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 that 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 $f(\theta_1, \dots, \theta_k)$, and the test statistic for the sample is T . A Monte Carlo approximation can be carried out as follows: (1) Estimate parameters $\theta_1, \dots, \theta_k$ in the model by the method of maximum likelihood, (2) generate N observations according to the distribution $f(\theta_1, \dots, \theta_k)$ and calculate the test statistic t_i for this sample and (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 five equally spaced markers on the chromosome with genetic distance 10 cM 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

TABLE 8

Power comparison between grouped and ungrouped tests

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

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.

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 $< \alpha$. 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

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

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, whereas 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 $Cx(Co)$ in the second pair. The results are summarized in Table 8 and Table 9.

From Tables 8 and 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.

H. Zhao, T. P. Speed and M. S. McPeck

TABLE 9

Power comparison between grouped and ungrouped tests					
λ	$-1/2$	0	$1/2$	1	$3/2$
Ungrouped test					
$\alpha = 0.01$	0.009	0.086	0.123	0.077	0.048
$\alpha = 0.05$	0.069	0.213	0.278	0.235	0.017
$\alpha = 0.10$	0.139	0.321	0.407	0.393	0.308
Grouped test					
$\alpha = 0.01$	0.099	0.111	0.115	0.115	0.125
$\alpha = 0.05$	0.250	0.254	0.251	0.256	0.271
$\alpha = 0.10$	0.359	0.363	0.366	0.372	0.379

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

COMPARISON WITH THE MODEL OF GOLDGAR AND FAIN

A thorough comparison of different models is made by McPECK and SPEED (1995). 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 that has 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 NCI model that is consistent with their model on a single meiotic product (see D. GOLDSTEIN, H. ZHAO and T. P. SPEED unpublished results). 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 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 as follows: d_i , $i = 0, 1, 2$ and 3, the probabilities of 0, 1, 2 and 3 crossovers, respectively; h , 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 that 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 *D. 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 that 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 NCI. Thus, the model they describe is incompatible with the assumption of NCI. 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 NCI, 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 (STURTEVANT 1915; MULLER

TABLE 10
Comparison with GOLDGAR and FAIN's model (WEINSTEIN's data)

Interval	<i>sc-ec</i>	<i>ec-cv</i>	<i>cv-cl</i>	<i>cl-v</i>	<i>v-g</i>	<i>g-f</i>	LR
GOLDGAR and FAIN	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

Estimated genetic distances based on GOLDGAR and FAIN's model and the $Cx(Co)^4$ model, together with likelihood ratio statistics (LR).

Chi-Square Model for Interference

TABLE 11

Comparison with GOLDGAR and FAIN's model (MORGAN's data)

Interval	<i>sc-ec</i>	<i>ec-cv</i>	<i>cv-cl</i>	<i>cl-v</i>	<i>v-s</i>	<i>s-f</i>	<i>f-ca</i>	<i>ca-b</i>	LR
GOLDGAR and FAIN	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

Estimated genetic distances based on GOLDGAR and FAIN's model and the $Cx(Co)^4$ model, together with likelihood ratio statistics (LR).

1916), 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 (Figure 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 >1.5 when the distance between the two regions is bigger than 60 cM, 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 *D. melanogaster* and *N. crassa* data agree with those given by FOSS *et al.* (1993), where m was estimated by the ratio of gene conversions

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. $\circ$, the observed S value.

H. Zhao, T. P. Speed and M. S. McPeck

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. Whereas some amount of positive interference is shown in the above organisms, it is not present in two other organisms we analyzed, *S. pombe* and *A. 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 because 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-strand version of the model of GOLDBERG 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 GOLDBERG 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), biologically reasonable and generalizable to many organisms. In these respects the $Cx(Co)^m$ model seems superior. It has a biological basis and is computationally tractable. Because of its simple structure, it can be applied to a broad range of organisms.

Although the $Cx(Co)^m$ model discussed in this paper applies well to data of different organisms and gives some insight into the underlying crossover process, there is a lot of 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. McPECK and T. P. SPEED (unpublished results) 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 *et al.* (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 results) 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.

We 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

- BAILEY, N. T. J., 1961 *Introduction to the Mathematical Theory of Genetic Linkage*. Oxford University Press, London.
- BARRATT, R. W., D. NEWMAYER, D. D. PERKINS and L. GARNJOIST, 1954 Map construction in *Neurospora crassa*. *Adv. 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.
- GOLDBERG, D. E., and P. R. FAIN, 1988 Models of multilocus recombination: nonrandomness in chiasma number and crossover positions. *Am. J. Hum. Genet.* 43: 38–45.
- HALDANE, J. B. S., 1919 The combination of linkage values, and the calculation of distances between the loci of linked factors. *J. Genet.* 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. Probab.* 11: 479–501.
- KARLIN, S., and U. LIBERMAN, 1983 Measuring interference in the chiasma renewal formation process. *Adv. Appl. Probab.* 15: 471–487.
- KOHL, 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. BIASCHAK, M. B. PETERSEN, V. SHARMA *et al.*, 1993 A linkage map of human chromosome 21: 43 PCR markers at average intervals of 2.5 cM. *Genomics* 16: 562–571.
- MCPECK, M. S., and T. P. SPEED, 1995 Modeling interference in genetic recombination. *Genetics* 139: 000–000.
- MORGAN, T. H., C. B. BRIDGES and J. SCHULTZ, 1935 Report of

Chi-Square Model for Interference

- investigations on the constitution of the germinal material in relation to heredity. *Carnegie Instit. 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. *Am. Nat.* **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, UK.
- READ, T. 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. *Ann. Hum. Genet.* **43**: 61–70.
- 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. McPECK 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. R. Soc. Lond. Ser. 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. *Z. Indukt. Abstammungs-Vererbungslehre* **13**: 234–287.
- WEINSTEIN, A., 1936 The theory of multiple-strand crossing over. *Genetics* **21**: 155–199.
- ZHAO, H., M. S. McPECK and T. P. SPEED, 1995 Statistical analysis of chromatid interference. *Genetics* **139**: 1057–1065.

Communicating editor: B. S. WEIR

APPENDIX

Lemma: Under the $Cx(Co)^m$ model the probability of k_j crossovers between $\cdot l_j$ and $\cdot l_{j+1}$, $j = 1, \dots, n$ is

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

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

Proof: We start with the simplest case when there are only two markers. Because 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 y are related by $x = y/2p$, i.e., $y = 2px$.

In the following discussion suppose markers $\cdot l_1, \cdot l_2, \dots, \cdot l_n$ are laid out from left to right, and the C events occur also from left to right. The $Cx(Co)^m$ model as-

sumes that the crossover intermediate (C) events resolve in sequence like $CxCoCo \cdots CoCxCo \cdots$ and that the process is stationary, so the first C event to the right of $\cdot l_1$ has an equal chance of resolving as any of the $m+1$ elements of $Cx(Co)^m$. The occurrence of k crossovers between $\cdot l_1$ and $\cdot l_2$ might be the result of p^2 possible situations, depending on the number of Co 's before the first Cx to right of $\cdot l_1$ and the number of Co 's between $\cdot l_2$ 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 $\cdot l_1$ and $\cdot l_2$ 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 $\cdot l_1$ and $\cdot l_2$ is a Cx , and the rightmost C could be either one Cx (pk_1-p+1 C 's altogether between $\cdot l_1$ and $\cdot l_2$), the first Co after a Cx (pk_1-p+2 C 's) or the second Co after a Cx , etc. $i-1$ corresponds to the number of Co 's between $\cdot l_1$ and the first Cx .

We can write the sum in a matrix product form:

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

Each element in the first column of the matrix corresponds to the last C event between $\cdot l_1$ and $\cdot l_2$ being a Cx ; the second column corresponds to the last C being the first Co after the k_1 th Cx , the j th column to the j th C after the k_1 th Cx . Therefore, the sum of the j th ($j > 0$) column multiplied by $1/p$ is the probability that there are k_1 crossovers between $\cdot l_1$ and $\cdot l_2$, and the last C event is the $(j-1)$ th Co after the k_1 th Cx . Therefore if we define

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

then $p_{k_1}^j$ is the probability that the last C between $\cdot l_1$ and $\cdot l_2$ is the $(j-1)$ th Co after the k_1 th Cx with the exception that $p_{k_1}^0$ is the probability of the last C being the k_1 th Cx .

Now we consider the case for three markers $\cdot l_1, \cdot l_2$ and $\cdot l_3$. Given that the first C to the right of $\cdot l_2$ is the l th Co after a Cx , the probability of k_2 crossovers between $\cdot l_2$ and $\cdot l_3$ 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 $\cdot l_2$ and the first Cx after $\cdot l_2$ is the same as the chance that the last C between $\cdot l_1$ and $\cdot l_2$ is the $(p-l-1)$ th C after a Cx

H. Zhao, T. P. Speed and M. S. McPeck

which is $p_{k_1}^{p-l-1}$. Therefore the chance of k_1 crossovers between i_1 , i_2 , and k_2 crossovers between i_2 and i_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

$$(p_{k_1}^1 \ p_{k_1}^2 \ \cdots \ p_{k_1}^p) \mathbf{D}_{k_2}(y_2) \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Recall that $(p_{k_1}^1 \ p_{k_1}^2 \ \cdots \ p_{k_1}^p) = (1 \ 1 \ \cdots \ 1) \mathbf{D}_{k_1}(y_1)$, thus the probability of k_1 crossovers between i_1 and i_2 , k_2 crossovers between i_2 and i_3 is

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

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

Theorem 1: Define

$$\mathbf{N}_j = \mathbf{D}_0(y_j) + \frac{1}{2} \sum_{s=1}^j \mathbf{D}_s(y_j),$$

$$\mathbf{R}_j = \frac{1}{2} \sum_{s=1}^j \mathbf{D}_s(y_j),$$

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

$$P(i_1 i_2 \cdots i_n) = \frac{1}{p} \mathbf{1} \mathbf{M}_1 \mathbf{M}_2 \cdots \mathbf{M}_n \mathbf{1}',$$

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 $P_{(i_1 i_2 \cdots i_n)}^{(k_1 k_2 \cdots k_n)}$ for the probability of observing recombination pattern $(i_1 i_2 \cdots i_n)$ when there are k_j crossovers in the j th interval, $j = 1, 2, \dots, n$. Then we have

$$\begin{aligned} P(i_1 i_2 \cdots i_n) &= \sum_{k_1, k_2, \dots, k_n} P_{(i_1 i_2 \cdots i_n)}^{(k_1 k_2 \cdots k_n)} \\ &= \sum_{k_1} \sum_{k_2} \cdots \sum_{k_n} p_{i_1}^{(k_1)} p_{i_2}^{(k_2)} \cdots p_{i_n}^{(k_n)} \frac{1}{p} \\ &\quad \times 1 \mathbf{D}_{k_1}(y_1) \mathbf{D}_{k_2}(y_2) \cdots \mathbf{D}_{k_n}(y_n) \mathbf{1}' \\ &= \frac{1}{p} \mathbf{1} \left(\sum_{k_1} \sum_{k_2} \cdots \sum_{k_n} p_{i_1}^{(k_1)} p_{i_2}^{(k_2)} \cdots p_{i_n}^{(k_n)} \right. \\ &\quad \times \mathbf{D}_{k_1}(y_1) \mathbf{D}_{k_2}(y_2) \cdots \mathbf{D}_{k_n}(y_n) \left. \right) \mathbf{1}' \\ &= \frac{1}{p} \mathbf{1} \left(\sum_{k_1} p_{i_1}^{(k_1)} \mathbf{D}_{k_1}(y_1) \sum_{k_2} p_{i_2}^{(k_2)} \right. \\ &\quad \times \mathbf{D}_{k_2}(y_2) \cdots \sum_{k_n} p_{i_n}^{(k_n)} \mathbf{D}_{k_n}(y_n) \left. \right) \mathbf{1}' \\ &= \frac{1}{p} \mathbf{1} \mathbf{M}_1 \mathbf{M}_2 \cdots \mathbf{M}_n \mathbf{1}'. \end{aligned}$$

Theorem 2: Define

$$\mathbf{P}_j = \mathbf{D}_0(y_j) + \sum_{s=2}^j \frac{1}{s} (1/2 + (-1/2)^s) \mathbf{D}_s(y_j)$$

$$\mathbf{T}_j = \mathbf{D}_1(y_j) + \sum_{s=2}^j \frac{2}{s} (1 - (-1/2)^s) \mathbf{D}_s(y_j)$$

$$\mathbf{N}_j = \sum_{s=2}^j \frac{1}{s} (1/2 + (-1/2)^s) \mathbf{D}_s(y_j).$$

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

$$P(i_1 i_2 \cdots i_n) = \frac{1}{p} \mathbf{1} \mathbf{M}_1 \mathbf{M}_2 \cdots \mathbf{M}_n \mathbf{1}',$$

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 diatype, tetra-type and nonparental diatype are $1/3(1/2 + (-1/2)^k)$, $2/3(1 - (-1/2)^k)$ and $1/3(1/2 + (-1/2)^k)$, respectively. Using the same method as Theorem 1, the conclusion follows.

On Genetic Map Functions

Hongyu Zhao* and Terence P. Speed†

*Department of Biostatistics, University of California, Los Angeles, California 90024 and †Department of Statistics, University of California, Berkeley, California 94720

Manuscript received June 12, 1995

Accepted for publication December 20, 1995

ABSTRACT

Various genetic map functions have been proposed to infer the unobservable genetic distance between two loci from the observable recombination fraction between them. Some map functions were found to fit data better than others. When there are more than three markers, multilocus recombination probabilities cannot be uniquely determined by the defining property of map functions, and different methods have been proposed to permit the use of map functions to analyze multilocus data. If for a given map function, there is a probability model for recombination that can give rise to it, then joint recombination probabilities can be deduced from this model. This provides another way to use map functions in multilocus analysis. In this paper we show that stationary renewal processes give rise to most of the map functions in the literature. Furthermore, we show that the interevent distributions of these renewal processes can all be approximated quite well by gamma distributions.

GENETIC maps consist of different kinds of markers positioned along a chromosome, with their relative locations measured in the map units known as centi-Morgans. During meiosis, crossing over occurs after homologous chromosomes pair and duplicate, resulting in a four-strand structure. Each crossover involves two strands of different origins (nonsister pairs). The genetic map distance in Morgans between a pair of markers on the same chromosome is the average number of crossovers occurring between these markers during meiosis on one chromatid. Because crossovers are not observable, a genetic map function $r = M(d)$ is often used to infer genetic distance (d) from the observable recombination fraction (r).

In his 1919 paper, HALDANE made three contributions to the study of genetic map functions, as well as defining them. By assuming no crossover interference (STURTEVANT 1915; MULLER 1916), he derived the map function $r = \frac{1}{2}(1 - e^{-2d})$ with inverse $d = -\frac{1}{2} \log(1 - 2r)$; he also proposed the empirical inverse map function $d = 0.7r + 0.3(-\frac{1}{2} \log(1 - 2r))$ to account for crossover interference in the data then available, and he introduced a differential equation method that permitted the construction of a variety of map functions.

A commonly used measure of the degree of crossover interference is the coincidence coefficient C involving three markers. Letting $p_{i_1 i_2}$ be the probability of i_1 recombinations in the first interval and i_2 recombinations in the second interval, $i_1, i_2 = 0, 1$, C is defined by

$$C = \frac{p_{11}}{(p_{10} + p_{11})(p_{01} + p_{11})}.$$

Corresponding author: Terence P. Speed, Department of Statistics, University of California, Berkeley, CA 94720.
E-mail: terry@stat.berkeley.edu

The case of $C = 1$ corresponds to no crossover interference, while $C < 1$ and $C > 1$ correspond to *positive* and *negative* crossover interference, respectively. Because the $p_{i_1 i_2}$ can be directly estimated from recombination data in most experimental organisms, C is often used as an empirical index of the degree of crossover interference.

One implicit assumption underlying the use of a genetic map function for one organism is that the functional relationship between genetic distances and recombination fractions does not vary across the genome in this organism. This assumption implies that C should only depend on the map lengths of the intervals between three markers of interest, say d and h . Noting that $p_{10} + p_{11} = M(d)$, $p_{01} + p_{11} = M(h)$, and $p_{11} = \{M(d) + M(h) - M(d + h)\}/2$, we have

$$C(d, h) = \frac{M(d) + M(h) - M(d + h)}{2M(d)M(h)}.$$

Letting $h \rightarrow 0$, and assuming $\lim_{h \rightarrow 0} M(h)/h = 1$, we obtain the differential equation

$$M'(d) = 1 - 2C(d)M(d). \quad (1)$$

Different choices of $C(d)$ lead to different map functions, which will be discussed later.

One of the difficulties in using genetic map functions to construct genetic maps is that when there are more than three markers, the multilocus recombination probabilities cannot be uniquely determined from the map function, a point that was made by FISHER (1947). This identifiability problem can be solved by postulating that the probability that there is an odd number of crossovers across a set of (disjoint) intervals only depends on the total length of these intervals

H. Zhao and T. P. Speed

(GEIRINGER 1944; SCHNELL 1961). The underlying assumption is that the distance between two disjoint intervals is irrelevant to the joint recombination probabilities in these two intervals. However experimental results suggest that the degree of interference varies with the distance between two intervals: the smaller the distance, the stronger the interference. Thus, the above approach is not consistent with experimental results. Nevertheless, this assumption can be and has been used to calculate joint recombination probabilities from map functions. LIBERMAN and KARLIN (1984) proved that a necessary condition to guarantee that the recombination probabilities obtained in this way are nonnegative is that $(-1)^k M^{(k)}(0) \leq 0$ for all k . They (inappropriately) termed map functions that always give rise to nonnegative recombination probabilities "multilocus feasible" and showed that many map functions that were found to fit data well did not satisfy these criteria.

If crossovers are viewed as a stochastic point process along the chromatid and a given map function can be realized from a crossover process, multi-locus recombination probabilities compatible with the map function can be obtained by assuming crossovers are generated from this point process. Two classes of point processes have been studied extensively in the literature in the context of modeling crossover interference: renewal processes, reviewed in BAILEY (1961) and McPECK and SPEED (1995), and count-location processes, studied by KARLIN and LIBERMAN (1978) and RISCH and LANGE (1979).

In this paper, we show that for most map functions in the literature there exist stationary renewal processes that give rise to them, and so these map functions are compatible with the analysis of multilocus data via this approach. Moreover, the interevent distributions of the stationary renewal processes corresponding to most map functions can be closely approximated by gamma distributions.

A special class of stationary renewal processes, called chi-square models, which have chi-square interevent distributions with even degrees of freedom, was found to give good fit to data from a variety of organisms (ZHAO *et al.* 1995b). This class of models, which evolved from an ordinary renewal process model for crossovers on a single meiotic product proposed by FISHER *et al.* (1947), was mainly studied because of its mathematical tractability (McPECK and SPEED 1995). It was reintroduced by FOSS *et al.* (1993) from a biological perspective, motivated by observations from experiments on gene conversion although there are now serious doubts concerning the appropriateness of this motivation (see FOSS and STAHL 1995). The chi-square model was conveniently denoted as $Cx(Co)^m$, which corresponds to a chi-square renewal density with $2(m+1)$ degrees of freedom. Using the method of maximum likelihood, chi-square models were fitted to different organisms by ZHAO *et al.* (1995b). The estimates of m for *Drosophila* ($m = 4$) and

Neurospora ($m = 2$) were the same as those obtained by FOSS *et al.* (1993), who estimated m from the observed ratio of Cx to Co . S. LIN and T. P. SPEED (unpublished results) fitted chi-square models to data on six loci from the CEPH consortium map of human chromosome 10, which was analyzed by WEEKS *et al.* (1994) using other models, and estimated the parameter m to be 3. Their results suggested the presence of crossover interference during human meiosis. McPECK and SPEED (1995) compared the fit of the stationary renewal process model with gamma interevent distributions (the gamma model) with that of other models using one large data set of *Drosophila* and found that the gamma model gives a better fit than all other models. This fact, together with the observation that for most map functions the interevent distributions for the corresponding stationary renewal processes can be closely approximated by gamma distributions, suggest that to some degree the chi-square model, and more generally, the gamma model, is able to capture the important features of the unobservable crossover process.

After raising the question "What is a genetic map function?", SPEED (1995) discussed in detail many issues related to map functions. It should be pointed out that other methods have also been proposed to extend map functions to handle multilocus data (OWEN 1953; MORTON and MACLEAN 1984).

MAP FUNCTIONS AND STATIONARY RENEWAL PROCESSES

When modeling crossovers as a point process, a distinction should be made between the point process on the four strand bundle, the *chiasma process*, and the point process on a single strand, the *crossover process*. Modeling the crossover process as a renewal process has a long history. Following JENNINGS (1923) and MATHER (1936, 1937), FISHER *et al.* (1947) modeled crossing over as a renewal process; that is, crossovers along a single strand were assumed to be formed as a regular sequence starting from the centromere, with the length between two adjacent crossovers always following the same distribution. Although it was known that crossovers take place when four strands (chromatids) are present during meiosis, FISHER *et al.* (1947) only modeled the crossover process on a single strand, without relating it to the chiasma process on the four-strand bundle. On the other hand, most later work modeled the chiasma process as a renewal process and related the crossover process to the chiasma process (CARTER and ROBERTSON 1952; COBBS 1978; STAM 1979). The assumption that crossovers occur starting from the centromere and progressing toward the telomere is now known not to be true (WHITEHOUSE 1982), but it is still a convenience. To connect the chiasma process to the crossover process, assumptions have to be made concerning the way nonsister pairs are involved in crossovers. Two types of

Map Functions

interference can be distinguished: *chiasma interference*, where chiasmata on the four-strand bundle do not occur independently of each other, and *chromatid interference*, where the choices of nonsister pairs involved in different chiasmata are not independent.

The observation of crossover interference on the meiotic products (single strands) can be the result of chiasma interference alone, the result of chromatid interference alone, or the result of both types of interference. It is interesting to note that the operation of two types of interference can lead to no apparent crossover interference. Consider the case when the chiasma process follows a stationary renewal process with interevent distribution being the gamma distribution with shape parameter $1/2$, and there is complete positive chromatid interference, *i.e.*, the strands involved in one chiasma are never involved in the closest chiasmata to its left and to its right. It is easy to see that the distance between two crossovers on a single meiotic product from this process follows the gamma distribution with shape parameter 1, *i.e.*, the exponential distribution. Therefore crossovers on a single strand appear to occur independently of each other. This example shows that two types of interference cannot be separated based on single-strand recombination data, where meiotic products from a single meiosis are recovered separately. Therefore tetrad data, where all meiotic products from a single meiosis are recovered together, are often used to detect chromatid interference. Since there is no strong and consistent evidence of chromatid interference (WHITEHOUSE 1982), it is generally assumed that there is no chromatid interference (NCI) in the models proposed in the literature.

The assumption of NCI imposes certain constraints on both recombination and tetrad probabilities (SPEED *et al.* 1992; ZHAO *et al.* 1995a). These constraints further impose constraints on map functions (SPEED 1995):

$$0 \leq M(d) \leq 1/2,$$

$$M'(d) \geq 0,$$

$$M''(d) \leq 0.$$

It is also obvious that if the chiasma point process is simple and stationary in the map distance metric, then $M(0) = 0$ and $M'(0) = 1$ (DALEY and VERE-JONES 1988, Section 3.3). We will consider map functions that are defined on a finite interval $[0, L]$ and those defined $[0, \infty)$ separately. By imposing one more condition, we say a function M defined on $[0, \infty)$ satisfies condition (A) if

$$M(0) = 0, \quad (A1)$$

$$M'(d) \geq 0, \quad \text{for all } d, \quad (A2)$$

$$M'(0) = 1, \quad (A3)$$

$$M''(d) \leq 0, \quad \text{for all } d, \quad (A4)$$

$$\lim_{d \rightarrow \infty} M(d) = 1/2. \quad (A5)$$

Apart from (A5), these conditions are necessary under the assumptions of NCI and that the chiasma process is a simple stationary point process. Condition (A5) postulates that two markers that are very far apart on the same chromosome can be considered very loosely linked, effectively behaving like markers on different chromosomes and segregating independently. Moreover, we have the following theorem to characterize this class of map functions. The proof is given in the APPENDIX.

Theorem 1: Let M be the map function for a stationary renewal chiasma process satisfying NCI on a chromosome arm of infinite length. Then M satisfies (A). Conversely, suppose that a function M from $[0, \infty)$ into $[0, 1/2]$ satisfies (A). Then there is a stationary renewal chiasma process whose map function is M and whose renewal density is $-M''$.

For a map function M defined on $[0, L]$ where $L < \infty$, we say that M satisfies condition (B) if

$$M(0) = 0, \quad (B1)$$

$$M'(d) \geq 0, \quad \text{for all } d, \quad (B2)$$

$$M'(0) = 1, \quad (B3)$$

$$M''(d) \leq 0, \quad \text{for all } d, \quad (B4)$$

$$M'(L) = 0, \quad (B5)$$

$$M(L) = 1/2. \quad (B6)$$

We say that M satisfies condition (B)' if it satisfies (B1)–(B4) and

$$M'(L) > 0, \quad (B5)'$$

$$M(L) < 1/2. \quad (B6)'$$

For map functions defined on $[0, L]$, we have the following analogue of Theorem 1 with the proof given in the APPENDIX.

Theorem 2: Let M be the map function for a stationary renewal chiasma process satisfying NCI on a chromosome arm of finite length. Then M satisfies (B) or (B)' for any L . Conversely, suppose that a function M from $[0, L]$ into $[0, 1/2]$ satisfies (B) or (B)'. Then there is a stationary renewal chiasma process whose map function is M and whose renewal density is $-M''$ when $d \leq L$.

VARIOUS MAP FUNCTIONS

In this section, we apply our two theorems to some map functions in the literature to see if there are stationary renewal processes that can give rise to them.

HALDANE (1919): $M_{H_2}^{-1}(r) = 0.7r + 0.3(-1/2 \log(1 - 2r))$. It is easy to see that $M^{-1}(0) = 0$, $\lim_{r \rightarrow 1/2} M^{-1}(r) = \infty$, $(M^{-1})' > 0$, $(M^{-1})'(0) = 1$, $\lim_{r \rightarrow 1/2} (M^{-1})' = \infty$, and $(M^{-1})'' \geq 0$. So M_{H_2} satisfies condition (A), and there is a stationary renewal process giving rise to M_{H_2} .

LUDWIG (1934): $M_L(d) = 1/2 \sin(2d)$. Clearly this should only be considered a possible map function in

H. Zhao and T. P. Speed

FIGURE 1.—Comparison between the interevent density ($-M'_K$) of the stationary renewal process corresponding to the Kosambi map function and two gamma densities: gamma(4,2) and gamma(2v,v), where v is $1/(2 \log(2) - 1) \approx 2.6$.

the interval $[0, 1/4\pi]$, and it is easy to check condition (B) in this case, with $L = 1/4\pi$. Thus there is a stationary renewal process having it as a map function, although the chromosome arm is rather short.

KOSAMBI (1944): $M_K(d) = 1/2 \tanh(2d)$. Since $M'_K(d) = 4(e^{2d} + e^{-2d})^{-2}$, and $M_K(d) = -16(e^{2d} - e^{-2d})/(e^{2d} + e^{-2d})^3$, it is easy to check that (A) is satisfied. The interevent distribution for the corresponding stationary renewal process is $16(e^{2d} - e^{-2d})/(e^{2d} + e^{-2d})^3$. KOSAMBI (1944) found that this map function gave good fit to *Drosophila* data. FISHER *et al.* (1947) noticed that the Kosambi map function is very close to the map function from a renewal process with interevent distribution being chi-square with four degrees of freedom. Both $-M_K$ and the density of gamma(4,2) are plotted in Figure 1. The density of a gamma(b, g) variable is $[b(bx)^{b-1}e^{-bx}]/\Gamma(b)$. The mean and variance of $-M_K$ are $1/2$ and $\log(2)/2 - 1/4$. The density of gamma(2v, v), where $v = 1/(2 \log(2) - 1)$, which has the same mean and variance as $-M_K$, is also plotted in Figure 1.

CARTER and FALCONER (1951): $M_{CF}(r) = 1/4(\tanh^{-1}(2r) + \tanh^{-1}(2r))$. Here $M_{CF}(d)$ is the solution of the differential equation $M'(d) = 1 - 16M^4(d)$. This map function was found to fit mouse data better than other map functions (CARTER and FALCONER 1951). Since $M'_{CF}(d) = -64M_{CF}^3(d)(1 - 16M_{CF}^4(d))$, and $M_{CF} \leq 1/2$, (A) is satisfied. The corresponding stationary renewal process has interevent distribution $64M_{CF}^3(d)(1 - 16M_{CF}^4(d))$. The Carter and Falconer map function $-M'_{CF}(d)$ together with the density of gamma(14,7) and gamma(16,8) are plotted in Figure 2. It was found that $Cx(Co)^6$ and $Cx(Co)^4$ give the best fit for two mouse data sets in BLANK *et al.* (1988) and TODD *et al.* (1991) (ZHAO 1995); the choice of the Carter and Falconer map function would be equivalent to using the $Cx(Co)^6$ model.

STURT (1976): $M_S(d) = 1/2(1 - (1 - d/L)e^{-d(2L-1)/L})$, $0 \leq d \leq L$. This map function arises via a count-location

FIGURE 2.—Comparison between the interevent density ($-M'_{CF}$) of the stationary renewal process corresponding to the Carter and Falconer map function and two gamma densities: gamma(14,7) and gamma(16,8).

chiasma process that begins with an obligatory crossover event on the arm, followed by a Poisson-distributed number of crossover events having mean $2L - 1$. The total genetic length is thus L . This map function fails (B5) but satisfies (B6), and so no stationary renewal process can give rise to it.

RAO *et al.* (1977): $M_R^{-1}(r) = w_H M_H^{-1}(2r) + w_K M_K^{-1}(2r) + w_{CF} M_{CF}^{-1}(2r) + w_M M_M^{-1}(2r)$, where $w_H = p(1 - 2p)(1 - 4p)/3$, $w_K = -4p(1 - p)(1 - 4p)$, $w_{CF} = 32p(1 - p)(1 - 2p)/3$, $w_M = (1 - p)(1 - 2p)(1 - 4p)$, and $M_H^{-1}, M_K^{-1}, M_{CF}^{-1}$, and M_M^{-1} are the inverse of the Haldane, Kosambi, Carter-Falconer, and Morgan map functions. It is easy to check that for any p , w_H, w_K, w_{CF} , and w_M cannot all be positive. Indeed M_R does not satisfy our necessary conditions (WEEKS 1994). Following RAO *et al.*'s idea, we might try to define a map function from a set of $n > 2$ map functions by letting $M^{-1}(r) = \sum_{i=1}^n w_i(p) M_i^{-1}(r)$, where $w_i(p)$ is a polynomial in p of order $n - 1$, and $M(d)$ reduces to $M_i(d)$ when $p = p_i$ for given $0 < p_1 < p_2 < \dots < p_n < 1$. But it can be shown that for no p can the $w_i(p)$, $i = 1, 2, \dots, n$, so defined all be positive. Therefore this approach to obtaining empirical map functions from existing map functions would seem questionable.

FELSENSTEIN (1979): $M_F(d) = 1/2(1 - e^{2(K-2)d})/(1 - (K-1)e^{2(K-2)d})$. When $K > 2$, $\lim_{d \rightarrow \infty} M_F(d) = 1/2(K-1) \neq 1/2$, and so no stationary renewal process exists with this mapping function. Because $M_F(d) = (K-2)^{2(K-2)d}/(1 - (K-1)e^{2(K-2)d})^2$, $M'_F(d) = 2(K-2)^3 e^{2(K-2)d} (1 + (K-1)e^{2(K-2)d})/(1 - (K-1)e^{2(K-2)d})^3$, it is easy to verify that $M_F(d)$ satisfies (A) when $0 \leq K < 2$. In Felsenstein's map functions, K is a measure of interference, with $K > 1$ corresponding to negative interference, and $K < 1$ to positive interference. The mean of $-M'_F(d)$ is $1/2$, and the variance is $\log(2 - K)/(2(1 - K)) - 1/4$ ($1/2$ when $K = 1$). Both $-M'_F(d)$ and the gamma distribution with the same mean and variance are plotted in Figure 3 for different K s. Note the

FIGURE 3.—Comparison between the interevent density ($-M'_h$, solid curve) of the stationary renewal process corresponding to the Felsenstein map function and the gamma density (dotted curve) having the same mean and variance for $K = 0.4, 0.8, 1.2$, and 1.6 .

close agreement between Felsenstein's family and the gamma family.

KARLIN and LIBERMAN (1978, 1979): $M_{CL}(d) = \frac{1}{2} [1 - c(1 - d/L)]$. This class of map functions arise from the count-location process, where $c(s) = \sum_{k=0}^{\infty} c_k s^k$ is a probability generating function of a count variable c with distribution (c_k) , and $c'(1) = 2L$. M_{CL} is only well defined for finite L . It is easy to check that (B1)–(B4) are satisfied. Because $M_{CL}(L) = \frac{1}{2} (1 - c_0)$ and $M_{CL}(L) = c_1/2L$, from Theorem 2, there is a corresponding stationary renewal process for M_{CL} only if (1) $c_0 = 0$ and $c_1 = 0$, or (2) $c_0 > 0$ and $c_1 > 0$.

DISCUSSION

In constructing genetic maps, map functions have been used to infer the unobservable genetic distance between two markers from the observable recombination fraction between these markers. Different genetic map functions embody different degrees of crossover interference among the crossovers. It has been observed that different organisms have different degrees of crossover interference, it so is not surprising that different map functions have been found suitable for different organisms. The major disadvantage of using map functions is that joint recombination probabilities cannot be uniquely determined in terms of them when there are more than three markers. Various approaches have been proposed to extend map functions to handle multilocus data.

One widely adopted approach, which was suggested by GEIRINGER (1944) and SCHNELL (1961) and thoroughly studied by LIBERMAN and KARLIN (1984), em-

bodies the assumption that for a pair of noncontiguous intervals, the probabilities for joint recombination patterns across these intervals do not depend on the distance between these two intervals, something that is not consistent with observations. Those map functions that can be extended to multilocus data through this approach have been (inappropriately) called "multilocus feasible" (LIBERMAN and KARLIN 1984). This criterion excludes many functions that were found to fit well to recombination data, such as the Kosambi map function.

In this paper, another approach is proposed to extend map functions for the analysis of multilocus data. If for any given map function, we can find a point process model that gives rise to this map function, then multilocus joint recombination probabilities can be obtained in a way that is completely compatible with the map function. Stationary renewal processes can give rise to many map functions. From this perspective, most map functions that are not multilocus feasible according to KARLIN and LIBERMAN can in fact be extended to permit the analysis of multilocus data.

Another measure of interference, called S_4 by Foss *et al.* (1993), is formally defined as

$$S_4(d) = \lim_{h \rightarrow 0} \lim_{k \rightarrow 0} \frac{p_{11}}{(p_{10} + p_{11})(p_{01} + p_{11})},$$

where the $p_{i_{1/2}}$ are as in the definition of C , with one interval having map length h , the other interval map length k , and the two intervals being separated by a map distance d . It seems that S_4 captures more important aspects of crossover interference than does C (MCPECK and SPEED 1995). Though S_4 cannot be deter-

H. Zhao and T. P. Speed

FIGURE 4.—The interevent density of the stationary renewal process corresponding to the map function M that satisfies $M' = 1 - (2M)^n$ for $n = 1, 3, 5, 6$.

mined from a map function, it can be calculated given a particular way of generalizing map functions to multilocus data and compared to empirical estimates. For the count-location processes, S_k is constant for all d , whereas S_k has various forms for the stationary renewal processes. The values of S_k as a function of d were estimated from a large *Drosophila* data set and were very close to the S_k values under the chi-square model (FOSS *et al.* 1993, McPECK and SPEED 1995).

Map functions cannot and should not be expected to reflect chiasma or crossover interference in anything but the most superficial way. Indeed, a map function could arise from both the stationary renewal process and the count-location process, although the interference differs greatly between these two classes of models. For example, the map function $M(d) = d/(1 + 2d)$ could arise from a stationary renewal process with interevent distribution $4(1 + 2d)^{-3}$, and it could also arise from a count-location model with $c_k = 1/2^k$, where $k \geq 0$ (LIBERMAN and KARLIN 1984). Under the count-location model, the length of the chromosome defined by $c = \{c_k\}$ is $1/2$, so $M(d)$ is only defined on $[0, 1/2]$, while d can range from 0 to ∞ for the stationary renewal process model.

Note that for the stationary renewal processes corresponding to most map functions discussed, the interevent distributions can often be well approximated by gamma distributions. Recall that the Haldane, Kosambi, and Carter and Falconer map functions are all solutions to the differential Equation 1 with $C(d) = (2M)^{n-1}$. It can be shown that map functions M obtained via this

approach always satisfy (A). Figure 4 displays plots of $-M''(d)$, the interevent distribution of the corresponding stationary renewal process for $n = 1$ (Haldane map function), $n = 3, 5$, and 6 . The cases $n = 2$ and 4 , which correspond to the Kosambi map function and the Carter and Falconer map function, were plotted in Figure 1 and Figure 2. When n is large, the corresponding density becomes close to a normal distribution. This suggests that if we take the approach proposed in this paper to generalize map functions to the multilocus situation, the stationary renewal processes corresponding to most genetic map functions can be well approximated by the gamma model, or the chi-square model. This provides, to some extent, an explanation of the fact that chi-square models (and hence gamma models) are flexible enough to give good fit to data from various organisms exhibiting different degrees of interference.

It has been assumed throughout this paper that there is no chromatid interference. Under this assumption, the relation between the chiasma process and the crossover process and the relation between the chiasma process and the map function are simple [see SPEED (1995) for a comprehensive review of this and related matters]. It was shown earlier that there are situations where the presence of both chiasma interference and chromatid interference could lead to no apparent crossover interference, from which we concluded that it is impossible to separate two types of interference from single-strand recombination data. Because there is no strong and consistent evidence of chromatid interference from the study of tetrad data, the assumption of no chromatid

Map Functions

interference is generally considered valid. The use of map functions and stationary renewal processes requires that the degree of interference be the same across the chromosome, which is an obvious simplification. But a large amount of data will probably be necessary to permit the detection of nonstationarity of the underlying process.

In summary, we have shown that for most genetic map functions, there is a corresponding stationary renewal process, and that these map functions can be extended to permit the analysis of multilocus data by calculating joint recombination probabilities from their corresponding renewal processes. This provides another way of generalizing a given map function to the multilocus situation, although it seems unlikely that there are efficient methods to estimate genetic distances and other parameters from multilocus recombination data under completely general renewal processes. The calculation of multilocus recombination probabilities for stationary renewal chiasma processes is discussed in the APPENDIX. However, comparisons between the interevent distribution of general stationary renewal processes and that of chi-square distributions suggest that this class of renewal processes provides satisfactory approximations to the renewal processes corresponding to most genetic map functions in the literature. With the limited amounts of data currently available, it will probably be hard to distinguish these models, although we can look forward to many refinements in the future.

This work was supported by National Institutes of Health grant HG-01093-01. The authors thank two referees for their helpful comments.

LITERATURE CITED

- BAILEY, N. T. J., 1961 *Introduction to the Mathematical Theory of Genetic Linkage*. Oxford University Press, London.
- BLANK, R. D., G. R. CAMPBELL, A. CALABRO and P. D. EUSTACHIO, 1988 A linkage map of mouse chromosome 12: localization of *Igh* and effects of sex and interference on recombination. *Genetics* 120: 1073–1083.
- CARTER, T. C., and D. S. FALCONER, 1951 Stocks for detecting linkage in the mouse and the theory of their design. *J. Genet.* 50: 307–323.
- CARTER, T. C., and A. ROBERTSON, A., 1952 A mathematical treatment of genetical recombination using a four-strand model. *Proc. Roy. Soc. B* 139: 410–426.
- COBBS, G., 1978 Renewal process approach to the theory of genetic linkage: case of no chromatid interference. *Genetics* 89: 563–581.
- DALEY, D. J., and D. VERE-JONES, 1988 *An Introduction to the Theory of Point Processes*. Springer-Verlag, New York.
- FELLER, W., 1971 *An Introduction to Probability Theory and Its Applications*. Vol. 2, Ed. 2. John Wiley, New York.
- FELSENSTEIN, J., 1979 A mathematically tractable family of genetic mapping functions with different amounts of interference. *Genetics* 91: 769–775.
- FISHER, R. A., 1947 The theory of linkage in polysomic inheritance. *Phil. Trans. Roy. Soc. B* 233: 55–87.
- FISHER, R. A., M. F. LYON and A. R. G. OWEN, 1947 The sex chromosome in the house mouse. *Heredity* 1: 335–365.
- FOSS, E., and F. W. STAHL, 1995 A test of a counting model for chiasma interference. *Genetics* 139: 1201–1209.
- FOSS, E., R. LANDE, F. W. STAHL and C. M. STEINBERG, 1993 Chiasma interference as a function of genetic distance. *Genetics* 133: 681–691.
- GEIRINGER, H., 1944 On the probability theory of linkage in Mendelian heredity. *Ann. Math. Statist.* 15: 25–57.
- 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.
- JENNINGS, H. S., 1923 The numerical relations in the crossing over of the genes with a critical examination of the theory that the genes are arranged in a linear series. *Genetics* 8: 393–457.
- KARLIN, S., and U. LIBERMAN, 1978 Classification and comparisons of multilocus recombination distributions. *Proc. Natl. Acad. Sci. USA* 75: 6332–6336.
- 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.
- KOSAMBI, D. D., 1944 The estimation of the map distance from recombination values. *Ann. Eugen.* 12: 172–175.
- LIBERMAN, U., and S. KARLIN, 1984 Theoretical models of genetic map functions. *Theor. Pop. Bio.* 25: 331–346.
- LUDWIG, W., 1934 Über numerische beziehungen der crossoverwerte untereinander. *Z. indukt. Abstamm. Vererb.* 67: 58–95.
- MATHER, K., 1935 Reductional and equational separation of the chromosomes in bivalents and multivalents. *J. Genet.* 30: 53–78.
- MATHER, K., 1936 The determination of position in crossing over. *J. Genet.* 33: 207–235.
- MATHER, K., 1937 The determination of position in crossing over. II. The chromosome length-chiasma frequency relation. *Cytologia, Jubb. Vol.*, 514–526.
- MCPHEE, M. S., and T. P. SPEED, 1995 Modeling interference in genetic recombination. *Genetics* 139: 1031–1044.
- MORGAN, T. H., and C. B. BRIDGES, 1916 Sex-linked inheritance in *Drosophila*. Carnegie Institute of Washington.
- MORTON, N. E., and C. J. MACLEAN, 1984 Multilocus recombination frequencies. *Genet. Res.* 44: 99–108.
- MULLER, H. J., 1916 The mechanism of crossing over. *Am. Nat.* 50: 193–221, 284–305, 350–366, 421–434.
- OWEN, A. R. G., 1953 The analysis of multiple linkage data. *Heredity* 7: 247–264.
- RAO, D. C., N. E. MORTON, J. LINDSTEN, M. HULTEN, and S. YEE, 1977 A mapping function for man. *Hum. Hered.* 27: 99–104.
- RISCH, N., and K. LANGE, 1979 An alternative model of recombination and interference. *Ann. Hum. Genet.* 43: 61–70.
- SCHNELL, F. W., 1961 Some general formulations of linkage effects in inbreeding. *Genetics* 46: 947–957.
- SPEED, T. P., 1995 What is a genetic map function? in *Genetic Mapping and DNA Sequencing*, edited by T. P. SPEED and M. S. WATERMAN. Springer-Verlag, New York.
- SPEED, T. P., M. S. MCPHEE and S. N. EVANS, 1992 Robustness of the no-interference model for ordering genetic markers. *Proc. Natl. Acad. Sci. USA* 89: 3103–3106.
- STAM, P., 1979 Interference in genetic crossing over and chromosome mapping. *Genetics* 92: 573–594.
- STURTEVANT, E., 1975 A mapping function for human chromosomes. *Ann. Hum. Genet.* 40: 147–163.
- STURTEVANT, A. H., 1915 The behavior of chromosomes as studied through linkage. *Z. Indukt. Abstammungs. Vererbungs.* 13: 234–287.
- TODD, J. A., T. J. AITMAN, R. J. CORNALL, S. GHOSH, J. R. HALL *et al.*, 1991 Genetic analysis of autoimmune type 1 diabetes mellitus in mice. *Nature* 351: 542–547.
- WEEKS, D. E., 1994 Invalidity of the Rao map function for three loci. *Hum. Hered.* 44: 178–180.
- WEEKS, D. E., G. M. LATHROP and J. OTT, 1993 Multipoint mapping under genetic interference. *Hum. Hered.* 43: 86–97.
- WHITEHOUSE, H. L. K., 1982 *Genetic Recombination: Understanding the Mechanisms*. John Wiley, New York.
- ZHAO, H., 1995 *Statistical Analysis of Genetical Interference*. PhD thesis, University of California at Berkeley, Berkeley.
- ZHAO, H., M. S. MCPHEE and T. P. SPEED, 1995a Statistical analysis of chromatid interference. *Genetics* 139: 1057–1065.
- ZHAO, H., T. P. SPEED and M. S. MCPHEE, 1995b Statistical analysis of crossover interference using the chi-square model. *Genetics* 139: 1045–1056.

Communicating editor: B. S. WEIR

H. Zhao and T. P. Speed

APPENDIX

Proof of Theorem 1: Suppose the crossovers are from a stationary renewal chiasma process with interevent density f . Without loss of generality, we may assume the mean interevent distance is $\mu = 1/2$, so the metric is simply genetic distance. For any point $\mathcal{M}_1$, say, on the chromosome, the chance that the first crossover after $\mathcal{M}_1$ occurs in the small interval $(y, y + dy)$ is

$$\int_y^\infty \frac{f(t)}{\mu} dt dy = 2 \int_y^\infty f(t) dt dy.$$

The probability of no crossovers occurring before $\mathcal{M}_2$, which is map distance d from $\mathcal{M}_1$, is

$$p_0 = \int_d^\infty \left\{ 2 \int_y^\infty f(t) dt \right\} dy.$$

Using Mather's formula (MATHER 1935), which asserts that the recombination fraction r between any two markers is

$$r = \frac{1}{2} (1 - p_0),$$

where p_0 is the probability of zero crossovers occurring between these markers, we have

$$M(d) = r = \frac{1}{2} \left\{ 1 - 2 \int_d^\infty \int_y^\infty f(t) dt dy \right\}.$$

It is easy to verify that M so defined satisfies (A).

Conversely, if M satisfies (A), then

$$\int_0^\infty (-M''(t)) dt = M'(0) = 1.$$

Thus $-M''$ is a probability density function on $[0, \infty)$. If the interevent distribution in a stationary renewal process is $-M''$, then their mean is

$$\begin{aligned} \int_0^\infty (-tM''(t)) dt &= \int_0^\infty \int_t^\infty (-M''(y)) dy dt \\ &= \int_0^\infty M'(t) dt = 1/2. \end{aligned}$$

Thus, the map function generated from the stationary renewal chiasma process with interevent distribution $-M''$ is

$$\frac{1}{2} (1 - 2 \int_d^\infty \int_y^\infty (-M''(t)) dt dy) = M(d).$$

Proof of Theorem 2: Note that

$$M(L) = \frac{1}{2} \left\{ 1 - 2 \int_L^\infty \int_y^\infty f(t) dt dy \right\},$$

and

$$M'(L) = \int_L^\infty f(t) dt.$$

It follows that either (B5) and (B6) are true, or (B5)'

and (B6)' are true. The first part of this theorem can then be proved as in Theorem 1.

If $M(L) < 1/2$ and $M'(L) > 0$, we may define an extended map function $M_E(d)$ on $[0, \infty)$ as follows:

$$M_E(d) = \begin{cases} M(d) & \text{if } d < L \\ M(L) + \alpha(1 - e^{-\beta(d-L)}) & \text{if } d \geq L, \end{cases}$$

where $\alpha = 1/2 - M(L)$ and $\alpha\beta = M'(L)$. It can be verified that $M_E(d)$ so defined satisfies (A) in Theorem 1. So there is a stationary renewal process whose corresponding map function is $M_E(d)$, which coincides with $M(d)$ on $[0, L]$. If $M(L) = 1/2$ and $M'(L) = 0$, it can be easily shown that the stationary renewal process with renewal density $-M''$ gives rise to M .

Calculating multilocus recombination probabilities for stationary renewal chiasma processes: Suppose that

$\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_n$ are $n+1$ consecutive loci along a chromosome, defining n genomic intervals $I_1 = [\mathcal{M}_0, \mathcal{M}_1)$, $\dots$, $I_n = [\mathcal{M}_{n-1}, \mathcal{M}_n)$ of map lengths $d_1, d_2, \dots, d_n$, respectively. Extending the notation introduced earlier, we denote by $p_{i_1, \dots, i_n}$ the joint recombination probability of having $i_j = 0$ or 1 recombination across interval I_j , $j = 1, \dots, n$. The question we address here is the calculation of all such probabilities $\mathbf{p} = (p_{i_1, \dots, i_n})$ when the underlying chiasmata form a renewal process stationary in the genetic distance metric and NCI is assumed. To do so we make use of the so-called *linkage values*, denoted by $\mathbf{z} = (z_{i_1, \dots, i_n})$, where $z_{i_1, \dots, i_n}$ is the probability of finding no chiasmata in the union $\bigcup \{I_j : i_j = 1\}$ of those intervals for which $i_j = 1$, see SPEED *et al.* (1992) for fuller details. We also make use of some well known facts from renewal theory and refer to FELLER (1971) for derivations. Suppose that we have a stationary renewal process with interevent density f and mean interval length μ . The distance of an arbitrary but fixed point on the chromosome to the next chiasma to its left (respectively right) is called the backward (respectively forward) recurrence length (traditionally called "time"), and if these are denoted by β and ϕ , then

$$P(\beta > u, \phi > v) = \int_{u+v}^\infty \frac{1 - F(y)}{\mu} dy, \quad (2)$$

where F is the cumulative distribution function (*c.d.f.*) corresponding to f .

Now let us consider the calculation of the linkage values $\mathbf{z} = (z_{i_1, \dots, i_n})$. For $n = 2$ this is quite straightforward. Suppose that we want to calculate z_{10} , the probability of no crossovers in I_1 . We regard $\mathcal{M}_1$ as the arbitrary but fixed point in the preceding discussion, and put $u = d_1$ and $v = 0$ in (2), obtaining the formula $z_{10} = 1 - F^*(d_1)$, where F^* is the *c.d.f.* corresponding to $f^* = \mu^{-1}(1 - F)$. Similarly, $z_{01} = 1 - F^*(d_2)$, $z_{11} = 1 - F^*(d_1 + d_2)$, and $z_{00} = 1$. All of these expressions are easily computed as long as F^* is tractable.

We now turn to $n = 3$ intervals and the calculation of

Map Functions

$\mathbf{z} = (z_{ij})$. A quick run through all eight possibilities reveals that all but z_{101} can be obtained in the manner just illustrated with $n = 2$. For example, $z_{100} = 1 - F^*(d_1)$, $z_{011} = 1 - F^*(d_2 + d_3)$, etc. It turns out that the calculation of z_{101} , in general requires summing a series of multiple integrals, and that the only known cases in which these integrals simplify into something tractable are variants on thinned Poisson processes. Let us see why.

First we recall that z_{101} is the probability of no chiasmata in either I_1 or I_3 ; there may be zero, one or more in I_2 , where the count is not constrained. Thus an initial reduction of z_{101} is as follows:

$$z_{101} = z_{111} + \sum_{k=1}^{\infty} \zeta_k$$

where z_{111} is the probability of no chiasmata in $I_1 \cup I_2 \cup I_3 (= 1 - F^*(d_1 + d_2 + d_3))$, and ζ_k is the probability of k chiasmata in I_2 and none in I_1 or I_3 . We now give an expression for ζ_k that, in general, does not simplify, and we remark that we know of no substantially simpler alternative expressions in the literature on renewal processes.

If there are to be $k \geq 1$ chiasmata in I_2 , we may denote the forward recurrence interval length from $\mathcal{M}_1$ to the first event by y_1 , and the k subsequent interevent distances by $y_2, y_3, \dots, y_k, y_{k+1}$. Further we may denote by y_0 the backward recurrence distance to the first event to the left of $\mathcal{M}_1$. With this notation we can readily check that the probability ζ_k is the $(k+2)$ -fold integral of the joint density of $(y_0, y_1, \dots, y_k, y_{k+1})$ over the range $\{y_0 > d_1\} \cap \{y_1 + \dots + y_k < d_2\} \cap \{y_1 + \dots + y_k + y_{k+1} > d_2 + d_3\}$. The joint density of $y_0, \dots, y_{k+1}$ is the product

$$\mu^{-1} f^*(y_0 + y_1) \times \prod_{i=2}^{k+1} f(y_i),$$

and so our assertion is demonstrated: z_{101} is an infinite sum of multiple integrals and will have a tractable expression only when these sums and integrals simplify. For each $n \geq 3$ there is one or more $z_{i_1 \dots i_n}$ requiring such expressions, and so far it is only the class of chi-square renewal processes (ZHAO *et al.* 1995b) and a slightly more general class termed Poisson-skip processes (H. ZHAO, K. LANGE and T. P. SPEED, personal communication) that are known to yield simplifications.