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, 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 $\Re$ or a subset of $\Re$ is known as a *counting process*.

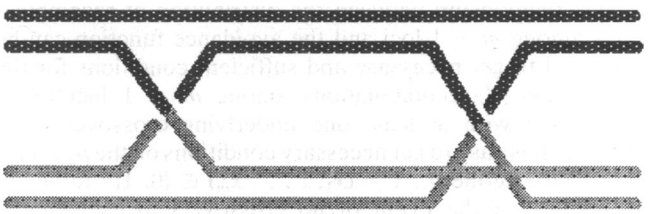

FIG. 1. Three-stand double crossover.

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

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

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, $x = (x_1, \ldots, x_m) \in \{0, 1\}^m$, denote the probability of the event of x_j recombinations in A_j, $j = 1, \ldots, m$. 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.

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 $y = (y_1, \ldots, y_m) \in \{0, 1\}^m$ and $y \cdot x = \Sigma_{j=1}^m y_j x_j$. That is, (Z_x) is the $\mathbb{Z}_2^m$ Fourier transform of (p_x), and (p_x) is the inverse $\mathbb{Z}_2^m$ 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, $x = (x_1, \ldots, x_m) \in \{0, 1\}^m$ 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 $x_j = 1; j = 1, \ldots, m$. 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

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

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

$$p_x = \sum_{y \ge x} \frac{1}{2^{y \cdot 1}} q_y.$$

Alternatively, the relation between the ps and the qs may be proved from *Lemma 1* of ref. 8, by induction. Inverting, we have

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

and

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

Now it is clearly necessary that we have

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

and

(*ii*) $$\text{for all } x,\ q_x \ge 0.$$

Expressed in terms of the p_x values, these conditions are

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

and

(*ii*) $$\text{for all } x,\ 0 \le \sum_{y \ge x} (-1)^{(y-x) \cdot 1} p_y.$$

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

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

and

(*ii*) $$\text{for all } x,\ 0 \le \sum_{y \ge x} (-1)^{(y-x) \cdot 1} Z_y.$$

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 $q_x \ge 0$ and $\Sigma_x q_x = 1$, 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*) $$Z(\emptyset) = 1,$$

and

(*ii*) $$\text{for all } x,\ Z_x \ge 0.$$

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* ½m(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 $\mathbf{f} = (f_1, \ldots, f_{m+1})$, where $\mathbf{f}$ is a permutation of $(1, \ldots, m + 1)$ and fit the data by maximum likelihood, assuming no interference, we will be fitting m parameters $(\theta_{f_1,f_2}, \ldots, \theta_{f_m,f_{m+1}})$, 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 $\mathbf{f}$, with probability 1 for n sufficiently large, the collection $\{\hat{\theta}_{f_1,f_2}, \ldots, \hat{\theta}_{f_m,f_{m+1}}\}$ dominates $\{\hat{\theta}_{1,2}, \ldots, \hat{\theta}_{m,m+1}\}$. 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.

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 θ 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*) $$M(0) = 0$$

and

(*ii*) $$\text{for all } x,\ 0 \le \sum_{y\in\{0,1\}^m} (-1)^{y\cdot x}(1 - 2M(\gamma_y)),$$

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

(*ii'*) $$\text{for all } x,\ 0 \le \sum_{y\ge x} (-1)^{(y-x)\cdot 1}(1 - 2M(\gamma_y)).$$

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, $j = 1, \ldots, m$, and let p_x, $x = (x_1, \ldots, x_m) \in \{0, 1\}^m$ be defined as before. Assume that the constraints of the general model described above hold with *strict* inequalities.

The data consist of r_x^n, $x = (x_1, \ldots, x_m) \in \{0, 1\}^m$, where r_x^n denotes the number of meioses in which x_j recombinations occurred in A_j, $j = 1, \ldots, m$. 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., $(\mathbf{r}^n)$ is distributed as Multinomial(n, $\mathbf{p}$), where $\mathbf{r}^n$ is the vector of r_x^n values and $\mathbf{p}$ is the vector of p_x values.

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

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

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

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

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 $\mathbf{p}$. Assume that these constraints on $\mathbf{r}^n$ hold. In terms of the c_x^n values, we have $\hat\theta_j = \tfrac{1}{2}n^{-1}\sum_{x:x_j=1} c_x^n$ and $\hat\theta_{f_j,f_{j+1}} = \tfrac{1}{2}n^{-1}\sum_{x\in I} c_x^n$, where $I = \{x: x_k = 1$ for at least one A_k lying between f_j and $f_{j+1}\}$. Note that we have $\hat\theta_j$ and $\hat\theta_{f_j,f_{j+1}}$ both $\le \tfrac{1}{2}$ under the constraints. From this representation, we can see that if A_k lies between f_j and f_{j+1}, then $\hat\theta_k \le \hat\theta_{f_j,f_{j+1}}$, 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 $0 \le x \le \tfrac{1}{2}$. Thus, if $L_{\mathbf{f}}$ is the maximized likelihood under order $\mathbf{f}$, and L_{true} is the maximized likelihood under the true order, we will have

$$L_{\mathbf{f}} = n\cdot\sum_{j=1}^{m} g(\hat\theta_{f_j,f_{j+1}}) \le n\cdot\sum_{j=1}^{m} g(\hat\theta_j) = L_{\text{true}}.$$

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 $\theta_{f_j,f_{j+1}}$ 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 $\mathbf{f}$ 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 $\mathbf{f}$ 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 $\mathbf{f}$. Let $i = \min\{j \in J\}$, so f_i is an endpoint of exactly one interval in the set. Now consider the M loci $\{1, 2, \ldots, M + 1\}\backslash\{f_i\}$ and the $|J| - 1$ intervals $\{(f_j, f_{j+1}): j \in J\backslash\{i\}\}$. By the induction hypothesis, these intervals cover at least $|J| - 1$ distinct intervals in the original order of $\{1, 2, \ldots, M + 1\}\backslash\{f_i\}$. Note that if the $\{(f_j, f_{j+1}): j \in J\backslash\{i\}\}$ cover exactly k distinct intervals in the original order of $\{1, 2, \ldots, M + 1\}\backslash\{f_i\}$, then they must cover either k or $k + 1$ distinct intervals in the original order of $\{1, 2, \ldots, M + 1\}$. If at least $|J|$ intervals of the original order of $\{1, \ldots, M + 1\}$ are covered by $\{(f_j, f_{j+1}): j \in J\backslash\{i\}\}$, then we are done, so assume without loss of generality that exactly $|J| - 1$ distinct intervals of the original orders of both $\{1, \ldots, M + 1\}$ and $\{1, \ldots, M + 1\}\backslash\{f_i\}$ 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 $\{(f_j, f_{j+1}): j \in J\backslash\{i\}\}$, 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.

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.
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.