Am. J. Hum. Genet. 49:1174–1188, 1991

Statistical Methods for Multipoint Radiation Hybrid Mapping

Michael Boehnke,* Kenneth Lange,† and David R. Cox‡

*Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor; †Department of Biomathematics, University of California School of Medicine, Los Angeles; and ‡Departments of Psychiatry, and Biochemistry and Biophysics, University of California School of Medicine, San Francisco

Summary

On the basis of the earlier work of Goss and Harris, Cox et al. introduced radiation hybrid (RH) mapping, a somatic cell genetic technique for constructing fine-structure maps of human chromosomes. Radiation hybrid mapping uses X-ray breakage of chromosomes to order a set of genetic loci and to estimate distances between them. To analyze RH mapping data Cox et al. derived statistical methods that employ information on sets of two and four loci, to build an overall locus order. Here we describe alternative nonparametric and maximum-likelihood methods for the analysis of RHs that use information on many loci simultaneously, including information on partially typed hybrids. Combination of these multipoint methods provides a statistically more efficient solution to the locus-ordering problem. We illustrate our approach by applying it to RH mapping data on 14 markers in 99 radiation hybrids for the proximal long arm of human chromosome 21.

Introduction

Building on the earlier work of Goss and Harris (1975, 1977*a*, 1977*b*), Cox and colleagues (Cox et al. 1990; Burmeister et al. 1991) recently demonstrated that radiation hybrid (RH) mapping provides a powerful method for fine-structure mapping of human chromosomes. In RH mapping, a rodent-human somatic cell hybrid containing a single human chromosome is lethally irradiated with X-rays, breaking the chromosomes into several fragments. Fragment-bearing hybrid cells are nonviable but can be fused with a normal rodent cell line. If this normal cell line is deficient in the enzyme HPRT, growth in HAT medium selects against the normal, nonfused rodent cells. Each hybrid clone arising from fusion of the two cell lines contains a unique set of fragments from the original human chromosome, and a clone can be typed for the presence or absence of human DNA markers.

The basic premise of RH mapping is that the closer two loci are on a chromosome, the less likely it is that radiation will induce a break between them. Thus, markers close together on a chromosome demonstrate correlated retention patterns in the hybrid clones, while loci far apart are retained nearly independently.

When chromosomes are irradiated, the average number of chromosome breaks is an increasing function of the X-ray dose (Goss and Harris 1977*a*); Cox et al. showed that X-irradiation with a dose of 8,000 rads generates sufficient numbers of fragments to construct maps at the 200–500-kb level of resolution. At this radiation dose, RH mapping bridges the resolution gap between linkage mapping or in situ hybridization, on the one hand, and physical mapping by pulsed-field gel electrophoresis (PFGE), on the other.

Cox et al. (1990) used the method of moments to estimate the distance between two loci. They assumed independent retention of fragments and random breakage along the chromosome. Under the second of these assumptions, breakage can be modeled as a Poisson process (e.g., see Karlin and Taylor 1975). Thus, the breakage probability θ for a given interval can be converted to an additive distance D by the formula $D = -\log(1-\theta)$, in analogy to Haldane's (1919) no-interference mapping function. The resulting units of distance for D are called Rays (Cox et al. 1990). For their chromosome 21 data and X-irradi-

Received April 29, 1991; revision received July 29, 1991.

Address for correspondence and reprints: Michael Boehnke, Ph.D., Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI 48109.

0002-9297/91/4906-0005$02.00

ation of 8,000 rads, Cox et al. noted a linear relationship between physical distance and Rays, with one centiRay approximately equaling 50 kb.

To order loci along the chromosome, Cox et al. choose the order that minimizes the sum of the distance estimates between adjacent linked loci in the map (also see Falk 1991, and in press). Having chosen an order, Cox et al. estimate the local support for the order by comparing likelihoods for the four-locus orders in which the internal two loci are interchanged. For this purpose, they calculate likelihoods for four-locus orders at the parameter estimates obtained from the various two-locus analyses.

In the current paper, we describe two alternative ordering methods for RH mapping that make use of information on many loci simultaneously, including information on partially typed hybrids. The first of these multipoint methods is nonparametric; it orders loci by minimizing the number of obligate chromosome breaks required to explain the hybrid data (Bishop and Crockford, in press; Boehnke, in press; Weeks et al., in press). The second method relies on maximum likelihood; it provides estimates of the distances between adjacent loci and the relative likelihoods of the different orders under various models. For the maximum-likelihood approach we consider a variety of models for fragment retention. These models range in complexity, from assuming that all retention probabilities are equal (Bishop and Crockford, in press; Boehnke, in press; Chakravarti and Reefer, in press; Green, in press) to assuming that all retention probabilities may differ (Cox et al. 1990). Each maximum-likelihood model depends on the assumptions of independent fragment retention and random chromosome breakage suggested by Cox et al. (1990). Both the minimum-breaks and the maximum-likelihood methods consider all loci simultaneously rather than restricting attention to two or four loci at a time (Cox et al. 1990). To achieve this for moderate to large numbers of loci requires special combinatorial and numerical techniques.

The minimum-breaks and maximum-likelihood methods may be used separately as distinct approaches to identify the best locus order. Alternatively, the minimum-breaks method can provide a preliminary list of candidate orders for evaluation by the computationally more intensive maximum-likelihood method. We illustrate both these methods with an analysis of RH mapping data on 14 markers in 99 RHs for the proximal long arm of human chromosome 21 (Cox et al. 1990).

Material and Methods

Data

In an RH mapping experiment, let loci $A_1, A_2, \ldots, A_N$ be typed on H radiation hybrids. For a specific locus order, we write the observation vector for a hybrid as $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, where $x_i = 1, 0,$ or ?, depending on whether marker i is typed and retained, typed and not retained, or not typed, respectively. For example, for $N = 14$, a possible observation vector is $\mathbf{x} = (1,1,1,?,0,0,?,0,0,0,1,1,1,1)$.

Minimum Number of Obligate Chromosome Breaks

Since the closer that two loci are on a chromosome, the less likely it is that a break will occur between them, a reasonable ordering strategy is to minimize the number of obligate chromosome breaks implied by the RH mapping data (Bishop and Crockford, in press; Boehnke, in press; Weeks et al., in press). For example, the hybrid $\mathbf{x} = (1,1,1,?,0,0,?,0,0,0,1,1,1,1)$ requires *at least* two breaks in the order given: one break between loci 3 and 5 and the other between loci 10 and 11. Other breaks may have occurred, but at least two are required to explain the hybrid in the order given. To count the number of obligate breaks for a hybrid, we count the number of times that 0 is immediately followed by 1 or that 1 is immediately followed by 0; in this count, ?'s are ignored. The minimum-breaks approach to RH mapping is analogous to minimizing the number of recombinants to infer order in genetic linkage mapping (Thompson 1987). It is also closely related to the maximum-parsimony method for reconstructing evolutionary trees (Edwards and Cavalli-Sforza 1964).

Minimizing the Number of Obligate Chromosome Breaks

For a modest number of loci N, counting and comparing numbers of obligate breaks for all $N!/2$ locus orders is feasible. This rapidly becomes impractical as N gets large; if $N = 14$, the number of locus orders is more than 43 billion. Thus, alternatives to exhaustive enumeration are required. We consider three such alternatives: (1) a branch-and-bound approach, (2) a simulated annealing approach, and (3) a stepwise locus-ordering approach in which locus orders are built one locus at a time, with partial locus orders kept and extended only if they require at most K breaks more than are required by the current best partial order. Branch-and-bound guarantees that the best minimum-breaks order is found but is not always computationally feasible when the number of loci N is large. The other two approaches are computationally less demanding for large N, but neither guarantees that the best solution is found.

Branch-and-bound.—Branch-and-bound (e.g., see Nijenhuis and Wilf 1978) is an approach to systematically eliminate large numbers of nonoptimal solutions to a problem, without actually considering each solution in detail. This is achieved by identifying early in the process a candidate solution that is optimal or nearly so and then eliminating solutions that are inferior either to the candidate solution or to better solutions encountered subsequently. In the locus-ordering context, we construct locus orders one locus at a time, and as soon as a partial locus order requires more breaks than the current best complete order, all complete orders consistent with the partial order are eliminated. In practice, a list of all orders that differ by K or fewer breaks from the current best complete order may be retained.

The branch-and-bound approach works because the criterion of interest—namely, minimum obligate breaks—never decreases as loci are added to an order. To prove this fact it is enough to demonstrate it for a single radiation hybrid $\mathbf{x}$. Let A_m be the locus to be added to the current partial order. If A_m is untyped, or if there are no other typed loci yet in the order, then adding A_m requires no new obligate breaks. If A_m is placed so that typed locus A_j is to one side of it but no typed locus is on its other side, then the number of obligate breaks is not altered if $x_j = x_m$, and it increases by one if $x_j + x_m = 1$. If A_m is placed between two adjacent typed loci A_j and A_k, and if $x_j \neq x_k$ or $x_j = x_k = x_m$, then there is no change in the number of obligate breaks; if $x_j = x_k \neq x_m$, then the number of obligate breaks increases by two. Thus, in every case, adding a new locus to an order cannot decrease the number of obligate breaks for a hybrid. Thompson (1987) proved an analogous result for minimizing the number of recombinants in linkage mapping.

To prime branch-and-bound, we use a greedy algorithm (Goodman and Hedetniemi 1977) to generate a good initial candidate order. Beginning with any of the $N(N-1)/2$ locus pairs, we determine the next locus to add to the current partial locus order by examining each unplaced locus and each possible position for it. The optimal position for an unplaced locus is the position that requires the smallest increase in the number of obligate breaks. The unplaced locus with the greatest difference between the mean number of breaks required by addition at nonoptimal positions and the number of breaks required by addition at its optimal position is then added at its optimal position; ties are broken randomly. Alternatively, the unplaced locus with the greatest difference between the number of breaks required by addition at its optimal and the number of breaks required by addition at its next best position could be added at its optimal position. The purpose of either greedy algorithm is to add at each stage that locus having strongest support for its optimal position. Carrying out this procedure for all $N(N-1)/2$ possible locus pairs makes it possible to identify a high-quality candidate order for the minimum-breaks criterion.

Stepwise locus ordering.—While the branch-and-bound approach allows elimination of many possible orders, the number of orders evaluated still may scale exponentially in the number of loci N. For situations where branch-and-bound is impractical, a close alternative to branch-and-bound is to build orders one locus at a time but to keep under consideration only those partial orders that are within K breaks of the current best *partial* order. When a partial order of the same length as the current best partial order is eliminated from consideration, all complete orders descended from it are also eliminated. In general, this approach considers many fewer partial orders at some risk of missing the overall best order. Larger values of the constant K increase the chance of identifying the best order but imply a heavier computational burden. The result of stepwise locus ordering will again be a list of orders which should include the best order(s). Choosing at each step to add that locus whose position is most strongly supported by the data, and/or beginning with an anchor map of well-placed loci, improve the chances of success for stepwise locus ordering. Similar approaches are often taken to construct multipoint linkage maps (e.g., see Barker et al. 1987).

Simulated annealing.—A final approach to minimize the number of obligate breaks is to use simulated annealing (Kirkpatrick et al. 1983; Press et al. 1989). Simulated annealing is motivated by the analogy of crystal formation in a cooling liquid. When cooled slowly, the molecules of a liquid settle into the minimum energy state for that system. If cooling is rapid, the minimum energy state may not be reached; instead, the system ends up in a polycrystalline or amorphous state of somewhat greater energy.

To simulate this annealing process, we identify the $N!/2$ locus orders with the states of a nonstationary Markov chain (e.g., see Karlin and Taylor 1975). The possible transitions for the Markov chain are block inversions of the current locus order. For example, if we are in the state corresponding to locus order 1-2-3-4-5-6-7-8-9-10, we may invert the block 5-6-7-8 to yield the new order 1-2-3-4-8-7-6-5-9-10. We choose the transitions for the chain uniformly from all $N(N-1)/2$ possible block inversions of loci. To each state i we associate an objective criterion E_i, here the minimum number of obligate breaks for the set of hybrids. E is called "energy" for short and is the quantity to be minimized.

The Markov chain is executed by proposal and acceptance steps. In state (locus order) i, propose a step to state (locus order) j according to the uniform transition probability of $2/[N(N-1)]$. If $E_j \leqslant E_i$, then accept the proposal. If $E_j > E_i$, then accept the proposal with probability $\exp[-(E_j - E_i)/T]$, where $T > 0$ represents temperature. Simultaneously taking many steps and gradually letting the temperature T approach zero, the process should stop in a state with minimum or nearly minimum energy. Keeping a list of the best orders encountered is a useful adjunct.

The essence of simulated annealing is that, early on in the process, steps leading to increased energy (number of obligate breaks) are often taken. This protects against prematurely being trapped in a local minimum. Later steps converge to the presumed global minimum. Although simulated annealing does not guarantee that the best solution will be found, practical experience has shown it to yield reasonable solutions to a wide variety of combinatorial optimization problems (Press et al. 1989).

We implemented simulated annealing by starting with a temperature T of 1,000. With this initial temperature, proposed transitions to a new order requiring 100 or fewer additional breaks are taken at least 90% of the time. We then let T decrease toward zero in 100 stages, at each stage multiplying T by 0.90. With N loci, a new stage is entered either after $10N$ proposed steps are accepted or after $100N$ steps are proposed, whichever comes first (Press et al. 1989). In the final stages of annealing, transitions to locus orders that require more breaks are effectively impossible.

Clearly, branch-and-bound is the approach of choice when feasible, since it guarantees that the best locus order is found. When branch-and-bound is impractical, simulated annealing and stepwise locus ordering can be used; for a large number of loci N, these approaches require substantially less computation and seem to work well (see Application).

Maximum-Likelihood Ordering for RH Data

Models and notation.—RH mapping by the minimum-obligate-breaks criterion is attractive because of its intuitive logic, its lack of restrictive assumptions, and its straightforward computation. However, the minimum-breaks method provides neither estimates of distances between loci nor comparison of relative likelihoods for competing locus orders. An alternative approach is to construct a model for the observed RH mapping data and to estimate model parameters by maximum likelihood.

Our basic model generalizes that of Cox et al. (1990). We assume (1) that chromosome breakage occurs randomly along the chromosome and so may be modeled as a possibly nonhomogeneous Poisson process (e.g., see Karlin and Taylor 1975) and (2) that, in a hybrid, fragments are retained independently. We define the breakage probability θ_i as the probability of at least one chromosome break between loci A_i and A_{i+1} $(1 \leqslant i \leqslant N-1)$, and we define the retention probability r_{ij} as the probability that a fragment including exactly loci $A_i, A_{i+1}, \ldots, A_j$ $(i \leqslant j)$ is retained in a hybrid; simpler submodels can be defined by placing restrictions on the retention probabilities (see below). Implicit in this notation is the assumption that the loci occur in the order $A_1, A_2, \ldots, A_N$. Cox et al. presented this general model in the simplest case of two loci. In that case, there are four possible types of hybrids, with probabilities.

$$\begin{aligned}
P(x_1=1, x_2=1) &= \theta_1 r_{11} r_{22} + (1-\theta_1) r_{12}\,; \\
P(x_1=1, x_2=0) &= \theta_1 r_{11}(1-r_{22})\,; \\
P(x_1=0, x_2=1) &= \theta_1 (1-r_{11}) r_{22}\,; \\
P(x_1=0, x_2=0) &= \theta_1 (1-r_{11})(1-r_{22}) + \\
&\quad (1-\theta_1)(1-r_{12})\,.
\end{aligned}$$

Note that, for the first and last categories above, the multiple terms in the probability reflect the inherent uncertainty about whether a break has occurred between the two loci. Much greater ambiguity regarding the number and locations of the chromosome breaks holds for larger numbers of loci N (see below). All possible chromosome breakage patterns consistent with a hybrid must be considered in order to calculate correctly the likelihood for the hybrid.

Likelihood for a given locus order.—Let $\mathbf{b} = (b_1, b_2, \ldots, b_{N-1})$ be the breakage vector for a hybrid, where $b_i = 1$ if there is at least one chromosome break on the interval between loci A_i and A_{i+1}, and $b_i = 0$ if there is no break. These breaks may or may not be directly observable. For a given breakage vector **b**, let $n(\mathbf{b}) = \sum_{i=1}^{N-1} b_i$ be the number of breaks in the hybrid. Finally, let $\mathrm{L}(\mathbf{b})$ be the ordered list of terminal loci for the fragments: $\mathrm{L}(\mathbf{b}) = \{j: b_j = 1\} \cup \{0,N\} = \{0 = \ell_0 < \ell_1 < \ldots \leqslant \ell_{n(\mathbf{b})+1} = N\}$. Then the probability of a hybrid **x** is

$$P(\mathbf{x}) = \sum_{\mathbf{b}} P(\mathbf{x}|\mathbf{b})\, P(\mathbf{b})\,, \tag{1}$$

where

$$P(\mathbf{b}) = \prod_{i=1}^{N-1} \theta_i^{b_i}(1-\theta_i)^{1-b_i}\,;$$

$$P(\mathbf{x}|\mathbf{b}) = \prod_{i=1}^{n(\mathbf{b})+1} Q_i\,.$$

Here, $Q_i = 1$ if none of the loci on fragment i were typed; $Q_i = r_{\ell_{i-1}+1,\ell_i}$ if all typed loci on fragment i were retained; $Q_{\mathrm{i}} = 1 - r_{\ell_{i-1}+1,\ell_i}$ if no typed loci on fragment i were retained; and $Q_i = 0$ otherwise.

The number of breakage vectors **b** consistent with the hybrid observation vector **x**—and hence the number of nonzero terms in equation (1)—varies depending on the number of breaks implied by **x** and on how precisely those breaks are positioned. Given adjacent typed loci i and $j > i$, there are 2^{j-i} possible vectors $(b_i, b_{i+1}, \ldots, b_{j-1})$ consistent with $x_i = x_j$, and there are $2^{j-i} - 1$ such vectors consistent with $x_i \neq x_j$, since the vector $(0,0, \ldots, 0)$ is inconsistent with $x_i \neq x_j$. In general, the number of vectors **b** consistent with a hybrid **x** is a product of terms of these forms. For example, for $\mathbf{x} = (1,1,1,?,0,0,?,0,0,0,1,1,1,1)$, $2 \cdot 2 \cdot 3 \cdot 2 \cdot 4 \cdot 2 \cdot 2 \cdot 1 \cdot 2 \cdot 2 \cdot 2 = 3{,}072$ of the $2^{13} = 8{,}192$ possible terms are nonzero.

Retention Probability Models

General model.—The general model of Cox et al. (1990) allows all $N(N+1)/2$ retention probabilities to differ. Together with $N-1$ breakage probabilities, this results in a total of $(N^2 + 3N - 2)/2$ parameters. The number of parameters is equal to the number of complete observation classes 2^N if the number of loci N is 2 or 3. For larger N, there are more possible observation classes than parameters. In practice, many of these classes will be unobserved, owing to relatively modest numbers of hybrids H; typically $H \approx 100$. Since the number of parameters grows rapidly, a prudent tactic is to consolidate the retention probabilities.

Moving-average model.—One simplification is to model a fragment retention probability as the average value of locus-specific retention probabilities for the loci on the fragment. For this moving-average model, let r_k be the probability that a fragment containing locus k only is retained in a hybrid ($1 \leqslant k \leqslant N$), and assume that $r_{ij} = \sum_{k=i}^{j} r_k/(j-i+1)$. This model involves a total of $2N-1$ parameters; it partially captures what happens if retention probability varies continuously along the chromosome.

Neither the general model nor the moving-average model is very restrictive in its assumptions about fragment retention. Unfortunately, both models require significant computation as the number of loci N gets large; calculating the likelihood for some hybrids requires summation of 2^{N-1} terms (see above) and must be repeated for multiple hybrids, for multiple iterations of a maximization routine, and for a potentially very large number of locus orders.

Markovian models.—There exists a class of retention probability models for which the likelihood can be calculated much more simply (Boehnke, in press). To describe these models, we first reexpress the probability of a radiation hybrid **x** as

$$P(\mathbf{x}) = P(x_{t_1}) \prod_{k=2}^{n} P(x_{t_k}|x_{t_1}, \ldots, x_{t_{k-1}})\,,$$

where $t = (t_1, t_2, \ldots, t_n)$ is the set of indices of the loci typed for hybrid **x**. For some models., $P(x_{t_k}|x_{t_1}, \ldots, x_{t_{k-1}}) = P(x_{t_k}|x_{t_{k-1}})$. We call these models Markovian, because, for such models, $x_{t_1}, x_{t_2}, \ldots, x_{t_n}$ can be viewed as a nonhomogeneous Markov chain (e.g., see Karlin and Taylor 1975) that takes only a finite number of steps. For any such model, the likelihood simplifies to

$$P(\mathbf{x}) = P(x_{t_1}) \prod_{k=2}^{n} P(x_{t_k}|x_{t_{k-1}})\,, \tag{2}$$

and the number of operations required to calculate $P(\mathbf{x})$ is a linear function of the number of loci N (see below).

Equal retention probability model.—The simplest retention probability model assumes that all retention probabilities r_{ij} are equal to some common value r ($1 \leqslant i, j \leqslant N$) (Bishop and Crockford, in press; Boehnke, in press; Chakravarti and Reefer, in press; Green, in press). This Markovian model includes a total of N parameters.

Centromeric or telomeric model.—In some RH-mapping data sets the proximity of a fragment to the centromere appears to have an effect on the retention probability for that fragment (Benham et al. 1989; Cox et al. 1990). Such a situation can be modeled most simply by setting the retention probability $r_{ij} =$

r_1 if $i = 1$ and $r_{ij} = r_2$ if $i \neq 1$; here, locus A_1 is assumed to be nearest the centromere (Bishop and Crockford, in press; Boehnke, in press; Lawrence and Morton, in press). This Markovian model requires a total of $N+1$ parameters. Further, since, for this model, orientation along the chromosome matters, consideration of $N!$ locus orders is required. This same model could equally well be used to model a telomeric effect on fragment retention.

Left-endpoint model. — The last two models are special cases of the more general Markovian model that sets the retention probability $r_{ij} = r_i$ for all $1 \leqslant i, j \leqslant N$ (Boehnke, in press). This model includes $2N-1$ parameters. If retention does not vary too much along the chromosome or if fragments tend not to include many loci, then this model is similar to the moving-average model. However, its likelihood is much simpler to compute.

Likelihood for the Markovian Models

To calculate $P(\mathbf{x})$ according to equation (2), we must evaluate terms of the form $P(x_j)$ and $P(x_j|x_i)$ ($i < j$), for typed loci A_i and A_j. We do so for the left-endpoint model; the other Markovian models are special cases of this more general model. First,

$$P(x_j) = \sum_{i=1}^{j} r_i^{x_j}(1-r_i)^{1-x_j}\theta_{i-1}\prod_{k=i}^{j-1}(1-\theta_k)\,, \qquad (3)$$

where we define $\theta_0 = 1$. The term corresponding to $i = 1$ in the sum in equation (3) is the probability of x_j when there are no breaks between A_1 and A_j; for $i > 1$ the ith term in the sum is the probability of x_j if the nearest break prior to A_j is between A_{i-1} and A_i. Note that if $j = 1$, then $P(x_1) = r_1^{x_1}(1-r_1)^{1-x_1}$. Similarly,

$$\begin{aligned} P(x_j|x_i) = {} & \delta_{x_ix_j}\prod_{k=i}^{j-1}(1-\theta_k) + \\ & \sum_{m=i+1}^{j} r_m^{x_j}(1-r_m)^{1-x_j}\theta_{m-1}\prod_{k=m}^{j-1}(1-\theta_k)\,, \end{aligned} \qquad (4)$$

where $\delta_{x_ix_j} = 1$ if $x_i = x_j$ and $\delta_{x_ix_j} = 0$ if $x_i \neq x_j$. The first term in equation (4) is the conditional probability of x_j given x_i if there are no breaks between A_i and A_j. The terms in the sum give that conditional probability if the nearest break prior to A_j is between A_{m-1} and A_m. Note that if $j = i+1$, then

$$P(x_{i+1}|x_i) = \begin{cases} 1-\theta_i + r_{i+1}\theta_i & (x_i,x_{i+1}) = (1,1) \\ (1-r_{i+1})\,\theta_i & (x_i,x_{i+1}) = (1,0) \\ r_{i+1}\theta_i & (x_i,x_{i+1}) = (0,1) \\ 1-\theta_i + (1-r_{i+1})\,\theta_i & (x_i,x_{i+1}) = (0,0) \end{cases}.$$

The log likelihood for a hybrid $\mathbf{x}$ is the sum of the logarithms of (a) one term of the type in equation (3) and (b) $n-1$ terms of the type in equation (4). Given a set of H independent hybrids $\mathbf{x} = (\mathbf{x}_1,\mathbf{x}_2, \ldots, \mathbf{x}_H)$, the joint log likelihood for the hybrids is the sum of the log likelihoods for the individual hybrids.

Likelihood Maximization by the EM Algorithm

EM updates. — In general, some method of iteration is required to maximize the log likelihoods for RH mapping. We have chosen to use the EM algorithm (Dempster et al. 1977). This algorithm is natural in contexts where the observed data can be considered as a subset of some hypothetical complete data. In the RH-mapping context, the complete data would be the locations of the breaks and the retention status for each of the resulting fragments. The retention patterns of the typed loci constitute the observed or incomplete data.

Given the complete data, the breakage probabilities and the retention probabilities can be viewed as success probabilities for partially hidden binomial experiments and can be estimated by sample proportions; the moving-average model is an exception to this rule (see below). Since we have only the incomplete data, it is reasonable to update an estimate of an entry p_i of the parameter vector $\mathbf{p}$ according to the general EM update formula

$$p_i^{\mathrm{new}} = \frac{\mathrm{E}(\text{no. of successes}|\mathbf{X},\mathbf{p}^{\mathrm{old}})}{\mathrm{E}(\text{no. of trials}|\mathbf{X},\mathbf{p}^{\mathrm{old}})}\,.$$

Here $\mathbf{p}^{\mathrm{old}}$ and $\mathbf{p}^{\mathrm{new}}$ are the current and updated parameter estimates, respectively, and the numbers of successes and trials refer to the binomial experiment corresponding to parameter p_i. An equivalent EM update formula (Weeks and Lange 1989) that tends to be easier to compute is

$$p_i^{\mathrm{new}} = p_i^{\mathrm{old}} + \frac{p_i^{\mathrm{old}}(1-p_i^{\mathrm{old}})\,d\log\mathrm{L}(\mathbf{p}^{\mathrm{old}})/dp_i}{\mathrm{E}(\text{no. of trials}|\mathbf{X},\mathbf{p}^{\mathrm{old}})}\,, \qquad (5)$$

where $\mathrm{L}(\mathbf{p})$ is the likelihood of the observed data. At each iteration of the EM algorithm, we update the estimates of the breakage and retention probabilities by using formula (5). The EM algorithm is guaranteed to increase $\log\mathrm{L}(\mathbf{p})$ (Dempster et al. 1977); iterations continue until $\log\mathrm{L}(\mathbf{p})$ stabilizes and the parameter estimates appear to converge.

The update formula (5) requires both the expected number of trials for a binomial experiment in the denominator and a derivative in the numerator. The expected number of trials conditional on the data and the old parameter estimates is easily calculated for each type of parameter. For a breakage probability θ_i, the expected number of trials is simply H, the number of hybrids. For the retention probability r of the equal retention probability model, the expected number of trials is the expected total number of fragments, $H[1 + \sum_{i=1}^{N-1} \theta_i]$. For the centromeric model, the expected number of trials is H for r_1 and $H \sum_{i=1}^{N-1} \theta_i$ for r_2. For the left-endpoint model, the expected number of trials for retaining a fragment starting with locus i is $H\theta_{i-1}$. Finally, for a general retention probability r_{ij}, the expected number of trials equals the expected number of fragments including loci $A_i, \ldots, A_j$: $H\theta_{i-1}\theta_j \prod_{k=i}^{j-1} (1 - \theta_k)$. In all these cases the displayed θ_i's are the current estimates and $\theta_0 = \theta_N = 1$. The derivatives required for formula (5) are not difficult to calculate. For the moving-average model, an alternative approach is required, since the retention probabilities are linear combinations of different subsets of the parameters. We have implemented the moving-average model by using a variable-metric approach to iterative maximization (Lange et al. 1988).

Parameter initial values.—Use of an iterative maximization scheme requires initial estimates for the parameters. For the general model, we use the moment estimates suggested by Cox et al. (1990). For the remaining models, we initially estimate a retention probability as the sample proportion of retained loci among the appropriate typed loci. For example, for the centromeric model, data for the first locus are used to estimate r_1, and data for all other loci are used to estimate r_2. Given initial retention probability estimates, we estimate breakage probabilities for adjacent locus pairs by two-locus maximum likelihood. Full maximum-likelihood estimation is not feasible for two-locus RH data for the general retention probability model, since the number of parameters equals the number of complete observational classes (see above).

Without loss of generality, consider θ_1. Let $s_{ij} = 1 - r_{ij}$, and let n_{ij} be the number of hybrids for which $(x_1, x_2) = (i, j)$, $i,j \ \varepsilon \ \{0,1\}$. When terms that do not involve θ_1, are ignored, the log likelihood for θ_1 is

$$\begin{aligned} \log L(\theta_1) = {} & n_{11} \log [\theta_1 r_{11} r_{22} + (1 - \theta_1) r_{12}] + \\ & (n_{10} + n_{01}) \log (\theta_1) + \\ & n_{00} \log [\theta_1 s_{11} s_{22} + (1 - \theta_1) s_{12}] \, . \end{aligned}$$

Setting the derivative of $\log L(\theta_1)$ to zero and defining $n = n_{11} + n_{10} + n_{01} + n_{00}$ results in the quadratic equation

$$n(r_{11}r_{22} - r_{12})(s_{11}s_{22} - s_{12})\theta_1^2 + [(n - n_{00})(r_{11}r_{22} - r_{12})s_{12} + (n - n_{11}) r_{12}(s_{11}s_{22} - s_{12})]\,\theta_1 + (n_{10} + n_{01}) r_{12} s_{12} = 0 \, , \quad (6)$$

which can be solved to yield the initial estimate of θ_1. Chakravarti and Reefer (in press) have solved equation (6) for the equal retention probability model.

Choosing the Maximum-Likelihood Order

Choosing the best maximum-likelihood locus order requires finding the locus order with the largest maximum likelihood. In principle, the same approaches used to identify the best minimum-breaks order can again be used to identify the best maximum-likelihood order. Here, computational efficiency is even more critical, since maximizing the likelihood requires substantially more time than counting obligate chromosome breaks.

Justification of branch-and-bound and stepwise locus ordering for maximum-likelihood requires proof that adding a locus to a partial locus order cannot increase the maximum likelihood of the RH-mapping data. Proof of this fact for the equal retention probability model is given in the Appendix.

A computationally less demanding approach to maximum likelihood is to examine either (*a*) only a short list of candidate orders identified by the minimum-breaks analysis or (*b*) those orders together with related orders obtained by one or more block inversions of loci (Weeks and Lange 1989). This combined approach keeps the number of orders manageable and can be particularly helpful either when there are large numbers of loci or when a Markovian model does not appear to be consistent with the data. The disadvantage of this combined approach is that it could in principle miss the best maximum-likelihood order. More experience with this approach is required, although it appears promising (see Application).

Model Choice

In principle, locus orders could be compared under each of the possible retention probability models, with the best maximum-likelihood order being chosen for each such model. If each retention probability model yielded the same best order, then the best-fitting model could be selected by likelihood-ratio tests; the validity of significance levels for this test is conditional on this order being the correct one. In practice, it is computationally less burdensome to identify a set of best candidate orders under one retention probability model and then to calculate maximum likelihoods for those orders under other retention probability models. Comparison of models for the various orders can then be done by likelihood-ratio testing.

Retention sample proportions observed for each locus can suggest whether nonequal retention probabilities will be necessary. However, since these sample proportions are dependent, treating them as independent in a simple statistical procedure such as a χ^2 goodness-of-fit test yields invalid significance levels.

Identification of Influential Hybrids

Once a best locus order has been selected, it is useful to check whether there are particular hybrids that were influential in distinguishing the best order from other nearly best orders. For the minimum-breaks method, this can be determined by comparing the number of chromosome breaks required for each hybrid under the best order with the number of breaks required for each hybrid under other, nearly best orders. The hybrids for which these counts differ are most responsible for the relative rankings of the orders. If only one or a few hybrids are responsible for this difference, we might choose to reexamine those hybrids for laboratory errors. We also might note whether any of the hybrids has a surprisingly large number of breaks under the best order. Finally it is worthwhile to observe whether any hybrids appear to display patterns of the sort (1,1, . . . , 1,0,1,1, . . . , 1) or (0,0, . . . 0,1,0,0, . . .,0). Although such patterns are logically possible, the discordant marker could represent a false negative or false positive.

In the maximum-likelihood context, likelihood-ratio statistics may be calculated hybrid by hybrid to compare the best order with a competing order. Each log likelihood should be evaluated at the MLEs for that order. Large log-likelihood differences identify the hybrids that were influential in determining the best maximum-likelihood order. As with minimum breaks, we can note whether one or a few hybrids were largely responsible for the inferred order.

Application

We applied these locus-ordering methods to RH data on 14 chromosome 21 markers in 99 RHs (Cox et al. 1990). These data are summarized in table 1. Two-locus RH lod scores (Cox et al. 1990) suggested that the 14 markers constituted a single linkage group. Since in no analysis of these data were the markers D21S12 and D21S111 distinguishable, ordering results in the tables exclude D21S111; order counts and analysis times are for all 14 loci.

Minimum Chromosome Breaks

We first analyzed the RH-mapping data by the minimum-breaks criterion by using the branch-and-bound approach. We obtained the best candidate order for comparison by the greedy algorithm described in the Material and Methods section. Break counts for subsequent orders generated were then compared with the 123 breaks required by the best candidate order. Partial orders requiring more than 132 breaks were eliminated from further consideration. Table 2 lists the nine 13-locus orders that required no more than 127 chromosome breaks. The overall best minimum-breaks order turned out to be the same as the best candidate order identified by the greedy algorithm, which was identical to the order arrived at by Cox et al. (1990) using a combination of likelihood-based methods and PFGE. A total of 244 14-locus orders required no more than 132 chromosome breaks. Identifying those 244 orders required visiting 311,097 partial and complete locus orders and took about 4.6 min computing time on a 486 25-MHz computer.

Since the branch-and-bound approach guaranteed that the best minimum-breaks order was found, no other approach to obtaining this best order was required for these data. However, for comparison purposes, we also used simulated annealing and stepwise locus ordering to attempt to identify the best minimum-breaks orders. On the basis of 100 different randomly chosen initial locus orders, we found that simulated annealing identified, on average, (*a*) the 35 best locus orders (i.e., on average, the 36th locus order was the best order missed) and (*b*) 141 of the 244 locus orders requiring no more than 132 chromosome breaks, while visiting 73,262 locus orders. In no case did simulated annealing fail to identify any of the seven best locus orders. For the stepwise locus-ordering method, we chose $K = 9$, so that all partial orders requiring no more than nine breaks more than the current best partial order were saved. Stepwise locus ordering with $K = 9$ identified (*a*) all 90 locus orders requiring no more than 130 breaks and (*b*) 224 of the 244 locus orders requiring no more than 132 chromosome breaks, while requiring consideration of only 15,085 locus orders. Thus, had branch-and-bound not been feasible, either of these other approaches would have identified the best minimum-breaks orders for these data.

For each of the best minimum-breaks orders, we

Table 1

Chromosome 21 RH-Mapping Data

Hybrid Type	Retention Pattern for Hybrid[b]													
(n[a])	16	48	46	4	52	11	1	18	8	AP	12	111	47	SOD
1 (1).............	0	0	0	0	0	1	0	0	0	0	0	0	1	1
2 (4).............	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3 (1).............	1	1	?	1	?	0	0	1	?	?	1	?	?	?
4 (1).............	0	0	1	1	1	1	0	0	0	0	0	?	0	0
5 (14)...........	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6 (1).............	1	1	1	1	?	0	0	0	0	0	0	?	0	?
7 (2).............	1	1	1	1	1	1	1	1	1	1	1	1	1	0
8 (1).............	1	1	1	1	0	0	0	0	0	0	0	0	0	0
9 (1).............	1	1	1	1	1	1	?	0	0	0	0	0	0	?
10 (1)...........	1	1	1	0	0	1	1	1	1	1	0	0	0	0
11 (1)...........	1	1	?	1	?	1	1	0	?	?	0	0	?	?
12 (2)...........	1	1	?	1	?	1	1	1	?	?	1	1	1	?
13 (2)...........	1	1	1	0	1	1	1	1	1	1	1	1	1	1
14 (1)...........	1	1	1	1	1	0	0	1	1	1	1	1	0	0
15 (1)...........	1	1	1	0	0	1	1	1	0	0	0	0	0	?
16 (1)...........	0	0	0	0	0	1	?	0	0	0	0	0	0	0
17 (1)...........	0	0	0	0	0	1	1	0	0	0	0	0	0	0
18 (1)...........	0	0	?	0	?	0	0	0	?	?	0	0	1	?
19 (1)...........	1	1	?	1	?	1	1	0	?	?	1	1	1	?
20 (1)...........	1	1	?	0	?	0	0	0	?	?	0	0	?	?
21 (1)...........	1	1	1	1	1	0	0	0	0	0	0	0	1	1
22 (1)...........	1	1	1	1	0	0	0	1	1	0	1	1	?	1
23 (1)...........	?	0	?	0	?	0	0	0	?	?	0	?	?	?
24 (1)...........	?	0	?	0	?	1	1	0	?	?	0	0	?	?
25 (3)...........	1	1	1	1	1	1	1	1	1	1	0	0	1	1
26 (1)...........	0	0	?	0	0	1	1	?	0	0	0	0	1	1
27 (1)...........	1	1	1	1	1	1	1	1	1	1	1	1	?	?
28 (1)...........	0	0	0	0	0	1	1	0	0	0	0	0	?	?
29 (1)...........	1	1	1	1	?	?	?	0	0	0	0	?	0	0
30 (1)...........	1	1	1	1	1	1	1	1	0	0	0	0	0	1
31 (1)...........	1	1	1	1	1	1	1	0	1	1	1	1	1	?
32 (1)...........	1	1	1	1	1	1	1	1	1	1	1	1	0	0
33 (1)...........	0	1	1	0	1	0	0	0	0	0	0	0	0	0
34 (1)...........	0	0	0	0	0	0	0	1	1	1	1	1	1	1
35 (1)...........	1	1	1	1	1	1	1	0	0	0	0	?	0	0
36 (1)...........	1	1	?	1	?	0	0	0	?	?	0	?	?	?
37 (1)...........	1	1	1	1	1	1	1	0	1	1	1	1	0	0
38 (1)...........	0	0	0	0	0	0	0	0	0	0	0	0	0	1
39 (1)...........	0	0	0	0	0	0	0	0	0	0	0	0	0	?
40 (1)...........	0	0	0	0	?	1	0	0	?	0	1	1	0	0
41 (1)...........	0	0	0	0	0	1	0	0	0	0	0	?	0	0
42 (1)...........	1	1	1	1	1	1	1	1	0	0	0	?	0	0
43 (2)...........	1	1	1	1	1	1	1	1	1	1	1	?	1	1
44 (1)...........	1	1	1	1	1	1	1	0	0	0	0	0	0	1
45 (1)...........	1	1	?	0	?	0	0	0	?	?	?	?	?	0
46 (1)...........	0	0	0	0	0	1	1	0	1	0	0	0	0	0
47 (1)...........	0	0	0	0	1	1	0	0	1	1	0	0	0	1
48 (1)...........	1	0	0	0	0	1	0	0	1	0	0	0	0	0
49 (1)...........	1	1	1	1	1	1	1	1	1	0	0	?	0	1
50 (1)...........	1	1	1	1	0	1	0	0	0	0	0	0	0	0
51 (1)...........	1	1	1	1	0	0	0	0	0	0	0	0	1	1
52 (1)...........	0	0	0	0	0	0	0	0	0	0	0	0	1	1
53 (1)...........	1	1	0	0	0	0	0	0	0	0	0	0	0	0

(*continued*)

Table 1 (continued)

Hybrid Type	Retention Pattern for Hybrid[b]													
(n[a])	16	48	46	4	52	11	1	18	8	AP	12	111	47	SOD
54 (1)...........	0	0	0	1	1	0	0	1	1	1	1	1	1	1
55 (1)...........	1	1	0	0	0	0	0	0	1	1	0	0	0	0
56 (1)...........	?	0	?	0	?	0	0	0	?	?	?	0	?	?
57 (1)...........	?	0	?	0	?	0	0	0	?	?	1	?	1	?
58 (1)...........	?	0	?	0	?	0	0	0	?	?	0	?	0	?
59 (1)...........	0	0	0	0	0	1	1	1	1	1	1	?	1	1
60 (1)...........	?	1	?	1	?	0	0	0	?	?	0	?	0	?
61 (1)...........	1	0	0	0	?	0	0	0	0	0	0	?	0	?
62 (5)...........	?	1	?	1	?	1	1	1	?	?	1	?	1	?
63 (1)...........	?	1	?	1	?	1	?	0	?	?	0	?	0	?
64 (1)...........	0	0	0	0	0	0	0	0	0	0	0	?	0	0
65 (1)...........	?	1	?	1	?	0	0	0	?	?	1	?	1	?
66 (1)...........	?	1	?	1	?	1	1	0	?	?	1	1	1	?
67 (3)...........	?	?	?	?	?	?	?	?	?	?	?	?	?	?
68 (1)...........	?	0	?	0	?	1	1	1	?	?	1	?	1	?
69 (1)...........	1	1	1	0	?	?	?	0	0	0	0	?	0	?
70 (1)...........	1	1	1	1	1	1	1	1	0	0	0	0	0	0
71 (1)...........	1	1	1	1	1	1	1	0	1	?	1	1	1	1

Source.—Cox et al. (1990).

[a] Number of occurrences.

[b] All numbered loci have D21S as prefix; AP = amyloid precursor; SOD = superoxide dismutase.

compared the number of chromosome breaks required by each hybrid under the best set of locus orders. We found that a single hybrid was responsible for the difference in the number of breaks required by the two best orders. Under the best locus order, this hybrid required two chromosome breaks; when loci D21S8 and APP were inverted to give the second best order, the hybrid required four. This hybrid was retyped, and the original typing results were confirmed. Similarly, the fourth and fifth best orders required one or two additional breaks in three and two hybrids, respectively; in contrast, the third best order required different numbers of breaks in 33 hybrids.

Maximum Likelihood

We next compared locus orders by maximum likelihood, initially using the equal retention probability model. Because maximizing the likelihood for a locus

Table 2

Locus Orders Implying Fewest Obligate Breaks for Chromosome 21 RHs

Locus Order[a]													Breaks
16	48	46	4	52	11	1	18	8	AP	12	47	SOD	123
16	48	46	4	52	11	1	18	AP	8	12	47	SOD	125
16	48	46	4	52	SOD	47	12	AP	8	18	1	11	126
48	16	46	4	52	11	1	18	8	AP	12	47	SOD	126
16	48	46	4	52	11	1	18	12	AP	8	47	SOD	127
16	48	46	4	52	11	1	18	12	8	AP	47	SOD	127
11	1	16	48	46	4	52	18	8	AP	12	47	SOD	127
16	48	46	4	52	1	11	18	8	AP	12	47	SOD	127
46	48	16	4	52	11	1	18	8	AP	12	47	SOD	127

Note.—Loci D21S12 and D21S111 were indistinguishable. Other conventions are as described in table 1.

[a] Single underlines indicate simple block inversions, compared with best locus order; double underlines indicate more complex rearrangements.

order requires substantially greater computation than counting obligate breaks, branch-and-bound was impractical for these data.

Instead, we used four other approaches in an attempt to identify the best maximum-likelihood locus orders. First, we calculated the maximum likelihood for the 90 locus orders that required no more than 132 chromosome breaks. Second, we did the same for the 7,428 distinct orders that resulted from block inversions of these 90 best minimum-breaks orders. Third, we carried out stepwise locus ordering with $K = 10$, so that all partial locus orders with maximum likelihood no more than 10^{10} times smaller than that for the current best partial order were saved at each step. Fourth, we carried out simulated annealing starting with five different initial locus orders. Since we had no global list of the best maximum-likelihood orders, we combined the lists of orders arrived at by each of these four methods, and we compared the performance of the methods in terms of the numbers of the orders they identified in the combined list.

Table 3 presents the 13 best maximum-likelihood orders identified for the equal retention probability model. The three best maximum-likelihood orders were the same as the three best minimum-breaks orders. Among the 90 best minimum-breaks orders were the 26 best maximum-likelihood orders, 36 of the 44 orders with maximum likelihoods within 10^4 times that of the best maximum-likelihood order, and 56 of the 74 orders with maximum likelihoods within 10^5 times that of the best maximum-likelihood order. Determining the maximum likelihood for a 14-locus order required on average about 0.5 s on our 486 25-MHz computer.

We next considered the 7,428 distinct locus orders that resulted from block inversions of the 90 best minimum-breaks orders. These 7,428 orders included (*a*) all 74 orders with maximum likelihoods within 10^5 times that of the best maximum-likelihood order and (*b*) 170 of the 172 orders with maximum likelihoods within 10^6 times that of the best order.

Stepwise locus ordering with $K = 10$ required evaluation of the maximum likelihood for 44,618 locus orders and identified 1,772 locus orders with maximum likelihood within 10^{10} times that of the best order. These orders included the 934 best maximum-likelihood orders, and 1,068 of the 1,070 locus orders with maximum likelihoods within 10^9 times that of the best order.

Simulated annealing on average identified the 24 best locus orders, 42 of the 44 locus orders with maximum likelihoods within 10^4 of the best locus order, and 62 of the 74 with maximum likelihoods within 10^5 of the best order and required visiting 68,924 locus orders. At a minimum, simulated annealing identified the 14 best locus orders.

We next calculated, under the centromeric and left-endpoint models, maximum likelihoods for the 200

Table 3

Best Maximum-Likelihood Locus Orders for Chromosome 21 RHs

Rank	Locus Order														$\Delta\log_{10}L$[a]	Breaks
1........	16	48	46	4	52	11	1	18	8	AP	12	47	SOD	.000	123
2........	16	48	46	4	52	11	1	18	AP	8	12	47	SOD	1.485	125
3........	16	48	46	4	52	SOD	47	12	AP	8	18	1	11	1.786	126
4........	52	4	46	48	16	11	1	18	8	AP	12	47	SOD	1.835	128
5........	16	48	46	4	52	1	11	18	8	AP	12	47	SOD	1.932	127
6........	11	1	16	48	46	4	52	18	8	AP	12	47	SOD	2.263	127
7........	11	1	52	4	46	48	16	18	8	AP	12	47	SOD	2.425	128
8........	16	48	46	4	52	11	1	18	12	AP	8	47	SOD	3.222	127
9........	16	48	46	4	52	11	1	18	12	8	AP	47	SOD	3.228	127
10......	16	48	46	4	52	SOD	47	12	8	AP	18	1	11	3.282	128
11......	52	4	46	48	16	11	1	18	AP	8	12	47	SOD	3.315	130
12......	16	48	46	4	52	1	11	18	AP	8	12	47	SOD	3.415	129
13......	1	11	16	48	46	4	52	18	8	AP	12	47	SOD	3.422	130

NOTE.—All conventions are as described in tables 1 and 2.

[a] $\log_{10}$-likelihood difference from the best order.

locus orders identified as best under the equal retention probability model; because of the large number of loci involved, use of non-Markovian models was not feasible for these data. For 136 of these 200 locus orders, the centromeric model fit the data better than did the equal retention probability model, when we tested at the .05 level; among these 136 orders were the six orders that were best under the equal retention probability model. For all 200 locus orders, the left-endpoint model gave a better fit to the data than did the equal retention probability model, even when we tested at the .005 level. However, estimates of locus-specific retention probabilities for the left-endpoint model were often at the boundary values of 0 and 1, perhaps suggesting an overparameterized model for these data.

For all three retention probability models, the same two orders were identified as best, although under the left-endpoint model the rankings of these two orders were reversed. Rankings for other orders varied somewhat between models, particularly for the left-endpoint model. A perplexing observation with regard to the left-endpoint model was that orientation along the chromosome—A_1-A_2-A_3 versus A_3-A_2-A_1—made no difference in the calculated maximum likelihood for order; fortunately, distance estimates under the different orientations were very similar though not identical. Table 4 presents distance estimates from Cox et al. (1990), together with maximum-likelihood distance estimates we obtained using the three Markovian retention probability models and the distance transformation $D = -\log(1-\theta)$. Distance estimates under the different models were remarkably similar.

Table 4

Distance Estimates (Rays) for Cox et al. Locus Order

	Estimate of Retention-Probability Model			
Locus[a]	Equal	Centromeric	Left-Endpoint	Cox et al.
D21S16				
	.076	.074	.075	.08
D21S48				
	.079	.077	.073	.09
D21S46				
	.194	.191	.178	.22
D21S4				
	.273	.274	.287	.27
D21S52				
	.644	.622	.618	.64
D21S11				
	.180	.179	.166	.17
D21S1				
	.556	.571	.572	.48
D21S18				
	.349	.361	.353	.40
D21S8				
	.111	.115	.111	.13
APP				
	.235	.249	.256	.28
D21S12				
	.362	.376	.343	.38
D21S47				
	.253	.264	.237	.26
SOD1				

[a] Loci D21S12 and D21S111 were indistinguishable.

Discussion

RH mapping has several important advantages. First, RH mapping complements existing mapping techniques. Given X-irradiation at a dose of 8,000 rads, Cox and coworkers (Cox et al. 1990; Burmeister et al. 1991) demonstrated that the level of resolution for RH mapping is intermediate between that of either linkage analysis or in situ hybridization, on the one hand, and physical mapping by PFGE, on the other. Thus, the several techniques complement and reinforce one another. Second, since RH mapping involves the analysis of a single copy of the human chromosome of interest, even nonpolymorphic markers can be used for map construction; thus, all markers are informative in every hybrid in which they are typed. Third, in contrast to recombination, for which the usual linkage mapping assumption of no interference is certainly violated, X-irradiation appears to induce breaks essentially at random, so that the no-interference assumption for RH mapping is reasonable. Fourth, some degree of experimental design is possible by modification of the X-ray dose. We can in principle ask the question: Given N markers in a region of B kilobases and a fixed fragment retention probability r, what is the optimal radiation dose necessary to order the loci? (K. Lange and M. Boehnke, unpublished data). Such questions will be important, given the wide interest in the use of RH mapping. A disadvantage of RH mapping is that it cannot be used to map disease loci.

Both of our methods for analyzing RH-mapping data have advantages and disadvantages. Both methods consider data for many loci simultaneously and take full advantage of partially typed RHs. The minimum-breaks method has the obvious advantages of computational simplicity and of requiring a minimum number of assumptions. We can expect a close relationship between it and maximum likelihood; in the context of linkage analysis, Thompson (1987) demonstrated that the analogous criterion of minimum recombinants is asymptotically equivalent to maximum likelihood under certain conditions. The disadvantage of the minimum-breaks method is that it provides neither estimates of distance nor relative likelihoods for the various locus orders.

In contrast, maximum likelihood is fully efficient and, since it assumes a parametric model, provides estimates of distances and relative likelihoods for orders. However, the maximum-likelihood method described here requires the assumption of independent retention of fragments, an assumption that may not be fully supported by the data. For example, in the proximal chromosome 21 RH-mapping data of Cox et al. (1990), the markers D21S16 and SOD1 are at opposite ends of the map, and the probability of no chromosome breaks between them is estimated as less than 4%. If independent retention of fragments and equal fragment retention probabilities are assumed, the conditional probability that SOD1 should be retained, given that D21S16 is retained, satisfies $P(x_{\mathrm{SOD1}} = 1 | x_{\mathrm{D21S16}} = 1) \approx .04 \cdot 1 + .96 \cdot P(x_{\mathrm{SOD1}} = 1) \approx .04 + P(_{\mathrm{SOD1}} = 1 | x_{\mathrm{D21S16}} = 0)$; and so it is about 4% greater than the conditional probability that SOD1 is retained, given that D21S16 is not. In fact, on the basis of the 64 hybrids in which both markers are typed, the conditional and probabilities of retaining SOD1 are estimated as $18/34 \approx .529$ and $8/30 \approx .267$, respectively. Despite the relatively small samples, there is some evidence that the corresponding true proportions differ by more than 4% ($z = 1.89$; two-sided $p = .06$) (e.g., see Mendenhall and Beaver 1991). Finally, the maximum-likelihood method involves assuming a retention probability model and comparing models by likelihood-ratio tests. In some cases, we may be forced to select poorly fitting models simply because of their computational tractability.

Particularly if the number of loci N is large, the minimum-breaks criterion can be used to generate a preliminary list of orders for further consideration by maximum likelihood. This worked very well for the chromosome 21 data of Cox et al. (1990). Alternatively, the minimum-breaks and maximum-likelihood methods can be used in parallel, and the resulting best sets of orders can be compared. This latter approach should be practical if the number of loci N is not too large, particularly if a Markovian model for the retention probabilities is consistent with the data.

The stepwise locus-ordering approach worked very well for the chromosome 21 data of Cox et al. (1990). It identified more of the best orders than did simulated annealing, while requiring evaluation of fewer locus orders. We do not know whether this is generally true. Perhaps the efficiency of simulated annealing could be improved by modification of either the initial temperature T or the speed with which T approaches zero.

For the maximum-likelihood method of RH mapping, which is the best retention probability model will likely depend on the data set. If a Markovian model such as the equal retention probability model, the centromeric model, or the left-endpoint model is consistent with the data as determined by likelihood-ratio testing, then many loci can be ordered simultaneously. If not, it may be helpful to compare orders for blocks of loci analyzed under both a more general retention probability model and a Markovian model. Our limited experience suggests that among the non-Markovian models the moving-average model is preferable to the general model, since it involves a more modest and identifiable set of parameters. Since RH mapping will often be used in concert with other mapping methods such as PFGE and linkage analysis, it may be that only a few loci will need to be ordered relative to one another by RH mapping, in which case a non-Markovian retention probability model can be employed to good effect.

Other retention probability models might also be entertained. For example, retention might be modeled as a function of the distance from the centromere (Bishop and Crockford, in press). The advantage of the maximum-likelihood approach described here is that we can, by likelihood-ratio tests, choose from a flexible class of retention probability models.

It is clear from examination of the lists of best orders (tables 2 and 3) that quantifying the support for a best locus order requires more than indicating the relative likelihoods when adjacent locus pairs are inverted. While some of the nearly best orders differ from the best order in this way (e.g., the two best orders differ by the inversion of APP and D21S8), many nearly best orders involve block inversions of a larger number of loci—or even more complex rearrangements. For this reason, we advocate presenting a list of the best orders together with their ordering criterion. This approach reflects more accurately the strength of evidence for the inferred locus order. The same considerations apply in presenting evidence for human linkage maps.

RHMAP is a package of FORTRAN programs that can be used to carry out RH mapping by the minimum obligate breaks and the maximum-likelihood methods described here. It is available from Michael Boehnke free of charge.

Acknowledgments

We thank Jennifer Barrett, Aravinda Chakravarti, Cathy Falk, Janet Warrington, and Daniel Weeks for helpful discussions and for critiquing a previous version of the manuscript. Support for this work was provided by National Institutes of Health grants R29-HG00376 (to M.B.), P30-HG00290 (to M.B.), CA16042 (to K.L.), R01-HD24610 (to D.R.C.), and P50-HG00206 (to D.R.C.).

Appendix

Use of branch-and-bound and stepwise locus ordering for maximum-likelihood RH mapping requires that the maximum likelihood cannot increase when a new locus is added to a partial locus order. We demonstrate this for the equal retention probability model by comparing the maximum likelihood for N loci with the maximum likelihood for $N-1$ loci.

Suppose that the ordered set of $N-1$ loci differs from the ordered set of N loci only by the absence of the locus in position j among the N loci. Consider the case of a single hybrid with observation vector $\mathbf{x} = (x_1, \ldots, x_N)$. For the $(N-1)$-locus model, this vector is amended to $\mathbf{x}^* = (x_1, \ldots, x_{j-1}, ?, x_{j+1}, \ldots, x_N)$. Since $\mathbf{x}$ contains all the information in $\mathbf{x}^*$, it is obvious that the N-locus probabilities satisfy

$$P_N(\mathbf{x}) \leqslant P_N(\mathbf{x}^*), \tag{A1}$$

regardless of the parameter values $(r, \theta_1, \ldots, \theta_{N-1})$. We next show that

$$P_N(\mathbf{x}^*) = P_{N-1}(\mathbf{x}^*), \tag{A2}$$

where the $(N-1)$-locus version of $\mathbf{x}^*$ on the right-hand side of equation (8) is $(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_N)$, and we define the $(N-1)$-locus parameters on the right-hand side of equation (A2) by $r^* = r$ and

$$\theta_k^* = \begin{cases} \theta_k & k < j-1 \\ \theta_{j-1} + \theta_j - \theta_{j-1}\theta_j & k = j-1 \\ \theta_{k+1} & k > j-1 \end{cases}.$$

This definition entails $1 - \theta_{j-1}{}^* = (1-\theta_{j-1})(1-\theta_j)$. In equation (2) (from the main text) with $\mathbf{x}^*$ substituted for $\mathbf{x}$, the factor $P(x_{t_1}^*)$ depends only on r and consequently must match across equation (A2). The remaining factors in equation (2) also provide matches, except possibly in the case $t_{k-1} \leqslant j < t_k$. Let B denote the event of a break between loci t_{k-1} and t_k. Then

$$\begin{aligned} P(x_{t\,k}^* | x_{t\,k-1}^*) &= P(x_{t\,k}^* | x_{t\,k-1}^*, \text{not } B)\, P(\text{not } B | x_{t\,k-1}^*) \\ &\quad + P(x_{t\,k}^* | x_{t\,k-1}^*, B)\, P(B | x_{t\,k-1}^*) \\ &= \delta_{x_{t_{k-1}} x_{t_k}} P(\text{not } B) + r^{x_{t_k}} (1-r)^{1-x_{t_k}} P(B) \end{aligned}$$

and

$$P(\text{not } B) = \prod_{i=t_{k-1}}^{t_k - 1} (1-\theta_i) = \prod_{i=t_{k-1}}^{t_k - 2} (1-\theta^*_i).$$

It follows that equation (A2) is true. Combining inequality (A1) and (A2) yields

$$P_N(\mathbf{x}) \leqslant P_{N-1}(\mathbf{x}^*). \tag{A3}$$

Taking the product of inequality (A3) over all hybrids produces the likelihood inequality

$$\mathrm{L}_N(r, \theta_1, \ldots, \theta_{N-1}) \leqslant \mathrm{L}_{N-1}(r^*, \theta_1^*, \ldots, \theta_{N-2}^*). \tag{A4}$$

In particular, inequality (A4) holds for the N-locus maximum-likelihood estimates. But $\mathrm{L}_{N-1}(r^*, \theta_1^*, \ldots, \theta_{N-2}^*)$ must in turn be no greater than L_{N-1} evaluated at the $(N-1)$-locus maximum-likelihood estimates, and the result is proved.

References

Barker D, Green P, Knowlton R, Schumm J, Langer E, Oliphant A, Willard H, et al (1987) Genetic linkage map of human chromosome 7 with 63 DNA markers. Proc Natl Acad Sci USA 84:8006–8010

Benham F, Hart K, Crolla J, Bobrow M, Francavilla M, Goodfellow PN (1989) A method for generating hybrids containing nonselected fragments of human chromosomes. Genomics 4:509–517

Bishop DT, Crockford GP. Comparisons of radiation hybrid mapping and linkage mapping. In: MacCluer JW, Chakravarti A, Cox D, Bishop DT, Bale SJ, Skolnick MH (eds) Cytogenet Cell Genet (in press)

Boehnke M. Radiation hybrid mapping by minimization of the number of obligate chromosome breaks. In: MacCluer JW, Chakravarti A, Cox D, Bishop DT, Bale SJ, Skolnick MH (eds) Cytogenet Cell Genet (in press)

Burmeister M, Kim S, Price ER, de Lange T, Tantravahi U, Myers RM, Cox DR (1991) A map of the distal region of the long arm of human chromosome 21 constructed by radiation hybrid mapping and pulsed-field gel electrophoresis. Genomics 9:19–30

Chakravarti A, Reefer JE. A theory for radiation hybrid (Goss-Harris) mapping: application to proximal 21q markers. In: MacCluer JW, Chakravarti A, Cox D, Bishop DT, Bale SJ, Skolnick MH (eds) Cytogenet Cell Genet (in press)

Cox DR, Burmeister M, Price ER, Kim S, Myers RM (1990) Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. Science 250:245–250

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc [B] 39:1–22

Edwards AWF, Cavalli-Sforza LL (1964) Reconstruction of evolutionary trees In: Heywood VH, McNeill J (eds) Phenetic and phylogenetic classification. London, Systematics Association, pp 67–76

Falk CT (1991) A simple method for ordering loci using data from radiation hybrids. Genomics 9:120–123

———. Multilocus mapping strategies on chromosome 21 data sets: comparison of results from family data, radiation hybrids and physical data. In: MacCluer JW, Chakravarti A, Cox D, Bishop DT, Bale SJ, Skolnick MH (eds) Cytogenet Cell Genet (in press)

Goodman E, Hedetniemi T (1977) Introduction to the design and analysis of algorithms. McGraw-Hill, New York

Goss SJ, Harris H (1975) New method for mapping genes in human chromosomes. Nature 255:680–684

——— (1977*a*) Gene transfer by means of cell fusion. I. Statistical mapping of the human X-chromosome by analysis of radiation-induced gene segregation. J Cell Sci 25: 17–37

——— (1977*b*) Gene transfer by means of cell fusion. II. The mapping of 8 loci on human chromosome 1 by statistical analysis of gene assortment in somatic cell hybrids. J Cell Sci 25:39–57

Green P. Construction and comparison of chromosome 21 radiation hybrid and linkage maps using CRI-MAP. In: MacCluer JW, Chakravarti A, Cox D, Bishop DT, Bale SJ, Skolnick MH (eds) Cytogenet Cell Genet (in press)

Haldane JBS (1919) The combination of linkage values, and the calculation of distance between the loci of linked factors. J Genet 8:299–309

Karlin S, Taylor HM (1975) A first course in stochastic processes, 2d ed. Academic Press, New York, pp 45–80, 117–128

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220:671–680

Lange K, Weeks D, Boehnke M (1988) Programs for pedigree analysis: MENDEL, FISHER, and dGENE. Genet Epidemiol 5:471–472

Lawrence S, Morton N. Physical mapping by multiple pairwise analysis. In: MacCluer JW, Chakravarti A, Cox D, Bishop DT, Bale SJ, Skolnick MH (eds) Cytogenet Cell Genet (in press)

Mendenhall W, Beaver RJ (1991) Introduction to probability and statistics, 8th ed. PWS-Kent, Boston, pp 305–307

Nijenhuis A, Wilf HS (1978) Combinatorial algorithms, 2d ed. Academic Press, New York, pp 240–246

Press WH, Flannery Bp, Teukolsky SA, Vetterling WT (1989) Numerical recipes: the art of scientific computing (FORTRAN version). Cambridge University Press, Cambridge, pp 326–334

Thompson EA (1987) Crossover counts and likelihood in multipoint linkage analysis. Int Math Assoc J Math Appl Med Biol 4:93–108

Weeks DE, Lange K (1989) Trials, tribulations, and triumphs of the EM algorithm in pedigree analysis. IMA J Math Appl Med Biol 6:209–232

Weeks DE, Lehner T, Ott J. Preliminary ranking procedures for multilocus ordering based on radiation hybrid data. In: MacCluer JW, Chakravarti A, Cox D, Bishop DT, Bale SJ, Skolnick MH (eds) Cytogenet Cell Genet (in press)