Bayesian methods and optimal experimental design for gene mapping by radiation hybrids

K. LANGE* AND M. BOEHNKE†

* *Department of Biomathematics, School of Medicine, University of California, Los Angeles, CA 90024*

† *Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109*

SUMMARY

Radiation hybrid mapping is a somatic cell technique for ordering human loci along a chromosome and estimating the physical distance between adjacent loci. The present paper considers a realistic model of fragment generation and retention. This model assumes that fragments are generated in the ancestral cell of a clone according to a Poisson breakage process along the chromosome. Once generated, fragments are independently retained in the clone with a common retention probability. Based on this and less restrictive models, statistical criteria such as minimum obligate breaks, maximum likelihood, and Bayesian posterior probabilities can be used to decide order. Distances can be estimated by either maximum likelihood or Bayesian posterior means. The model also permits rational design of radiation dose for optimal statistical precision. A brief examination of some real data illustrates our criteria and computational algorithms.

INTRODUCTION

In the mid-seventies Goss & Harris (1975) developed a new method for mapping human chromosomes. This method was based on irradiating human cells, rescuing some of the irradiated cells by hybridization to rodent cells, and analysing the hybrid cells for surviving fragments of a particular human chromosome. For various technical reasons, few geneticists followed up on this promising technique, and it lay dormant for several years until revived by Cox *et al.* (1990). The more sophisticated and successful version of radiation hybrid mapping introduced by Cox *et al.* (1990) and discussed below raises many fascinating statistical problems. These problems run the gamut from computation and optimization of complicated likelihoods to Bayesian inference for locus orders and optimal design of radiation dose levels. The present paper summarizes these issues and offers some tentative solutions.

In the hands of Cox *et al.* (1990), a radiation hybrid experiment starts with a human–rodent hybrid cell line. This cell line incorporates a full rodent genome and a single copy of one of the human chromosomes. To fragment the human chromosome, the cell line is subjected to an intense dose of X-rays, which naturally also fragments the rodent chromosomes. The repair mechanisms of the cell rapidly heal chromosome breaks, and the human chromosome fragments are typically translocated or inserted into rodent chromosomes. However, the damage done by irradiation is lethal to the cell line unless further action is taken to rescue individual cells. The remedy is to fuse the irradiated cells with cells from a second unirradiated rodent cell line. The second cell line contains only rodent chromosomes, so no confusion about the source of the

human chromosome fragments can arise for a new hybrid cell created by the fusion of two cells from the two different cell lines.

The new hybrid cells have no particular growth advantage over the more numerous unfused cells of the second cell line. However, if cells from the second cell line lack the enzyme hypoxanthine phosphoribosyl transferase (HPRT), both the unfused and the hybrid cells can be grown in HAT medium, which kills the unfused cells (Cox *et al.* 1990). The selection process leaves a few hybrid cells, and each of the hybrid cells serves as a progenitor of a clone of identical cells.

Each clone can be assayed for the presence or absence of various human genes on the original human chromosome. In practice, the cells of a clone generally contain from 20 to 60% of the human chromosome fragments generated by the irradiation of its ancestral human–rodent hybrid cell (Cox *et al.* 1990; Burmeister *et al.* 1991). The basic premise of radiation hybrid mapping is that the closer two loci are on the human chromosome, the less likely that irradiation will cause a break between them. Thus, close loci will tend to be concordantly retained or lost in the hybrid cells, while distant loci will tend to be independently retained or lost. The retention patterns from the various hybrid clones therefore give important clues for determining locus order and for estimating the distances between adjacent loci for a given order.

At this point we offer some guidance for reading the remainder of the paper. Because of the necessary, but heavy mathematical machinery developed, many readers may wish to peruse the next modelling section and then skip directly to the applications section and the discussion. These last two sections illustrate our methods and contain general conclusions. Of course, we invite the mathematically inclined to read the whole paper.

MODELS FOR RADIATION HYBRIDS

The breakage phenomenon for a particular human chromosome can be reasonably modelled by a Poisson process. The preliminary evidence of Cox *et al.* (1990) suggests that this Poisson breakage process is roughly homogeneous along the chromosome. For their data on human chromosome 21, Cox *et al.* (1990) found that 8000 rad of radiation produced on average about four breaks per cell. The intensity λ characterizing the Poisson process is formally defined as the breakage probability per unit length. Assuming an estimated length of 4×10^4 kb for chromosome 21, $\lambda \approx 4/(4 \times 10^4) = 10^{-4}$ breaks per kb when a cell is exposed to 8000 rad (Cox *et al.* 1990).

For any two loci the simple mapping function

$$1 - \theta = e^{\lambda\delta} \tag{1}$$

relates the probability θ of at least one break between the loci to the physical distance δ between them. When $\lambda\delta$ is small, $\theta \approx \lambda\delta$. This is analogous to the approximate linear relationship between recombination fraction and map distance for small distances in genetic recombination experiments. Indeed, except for minor notational differences, equation (1) is Haldane's (1919) classical mapping function for recombination without interference.

In addition to breakage, fragment retention must be taken into account when analysing radiation hybrid data. A reasonable assumption is that different fragments are retained independently. For the purposes of this exposition, we will make the further assumption that there is a common fragment retention probability r. Boehnke *et al.* (1991) consider at length

more complicated models for fragment retention. For instance, the fragment bearing the centromere of the chromosome may be more often retained than other fragments. This is biologically plausible because the centromere is involved in coordination of chromosome migration during cell division. However, these more complicated models appear to make little difference in ultimate conclusions.

In a radiation hybrid experiment, a certain number of clones are scored at several loci. For example, in the Cox *et al.* (1990) chromosome 21 data, 99 clones were scored at 14 loci. In some of the clones, only a subset of the loci were scored. One of their typical clones can be represented as (0, 0, 0, 0, 0, 1, 0, 0, 0, 0, ?, 0, 0, 1). A '1' in a given position of this observation vector indicates that the corresponding human locus was present in the hybrid clone, a '0' indicates that the locus was absent, and a '?' indicates that the locus was untyped in the clone.

Computing the minimum number of obligate breaks per order allows comparisons of different orders (Boehnke *et al.* 1991; Boehnke, 1992; Bishop & Crockford 1992; Weeks *et al.* 1992). If the order of the loci along the chromosome is the same as the scoring order, then the above clone requires three obligate breaks. These breaks occur whenever a run of 0's is broken by a 1 or vice versa; untyped loci are ignored in this accounting. The minimum number of obligate breaks for each clone can be summed over all clones to give a grand sum for a given order. This grand sum serves as a criterion for comparing orders. Boehnke *et al.* (1991) discuss how this criterion can be minimized by a stepwise ordering algorithm or by standard combinatorial optimization techniques such as branch-and-bound (Reingold *et al.* 1977) and simulated annealing (Press *et al.* 1989).

MAXIMUM LIKELIHOOD METHODS

The advantage of the minimum breaks criterion is that it depends on almost no assumptions about how breaks occur and fragments are retained. The disadvantage of this nonparametric criterion is that it provides neither estimates of physical distances between loci nor comparison of relative likelihoods for competing orders. Maximum likelihood obviously remedies the latter two defects, but at the expense of introducing some of the explicit assumptions mentioned earlier. We will now briefly discuss how likelihoods are computed and maximized for a given order. Different orders can be compared on the basis of their maximum likelihoods.

Because different clones are independent, it suffices to demonstrate how to compute the likelihood for a single clone. Let $\mathbf{X} = (X_1, \ldots, X_m)$ be the observation vector for a clone potentially typed at m loci. The component X_i is defined as 0, 1, or ?, depending on what is observed at the ith locus. We can gain a feel for computing the likelihood of $\mathbf{X}$ by considering two simple cases. If $m = 1$ and $X_1 \neq ?$, then X_1 follows the binomial distribution

$$\Pr(X_1) = r^{X_1}(1-r)^{1-X_1} \tag{2}$$

for retention or non-retention. When $m = 2$ and both loci are typed, the likelihood must reflect breakage as well as retention. It is straightforward to show that

$$\left.\begin{aligned}
\Pr(X_1 = 0, X_2 = 0) &= (1-\theta)(1-r) + \theta(1-r)^2 \\
&= (1-\theta r)(1-r) \\
\Pr(X_1 = 1, X_2 = 0) &= \theta r(1-r) \\
\Pr(X_1 = 0, X_2 = 1) &= \Pr(X_1 = 1, X_2 = 0) \\
\Pr(X_1 = 1, X_2 = 1) &= (1-\theta)r + \theta r^2 \\
&= (1-\theta+\theta r)r.
\end{aligned}\right\} \tag{3}$$

For instance, $\Pr(X_1 = 1, X_2 = 0)$ is the product of the probability θ of breakage between the loci times the probability $r(1-r)$ that the fragment containing the first locus is retained and the fragment containing the second locus is lost. Note the substitution throughout (3) of θ for $1-e^{-\lambda\delta}$. This substitution has the advantage of allowing the Poisson breakage process to be nonhomogeneous. In any case, only the product $\lambda\delta$ of the two parameters λ and δ is identifiable.

Generalization of the above likelihood expressions to more loci involves two subtleties. First, the sheer number of terms accounting for all possible breakage and retention patterns quickly becomes unwieldy. Second, missing data can no longer be ignored. The key to efficient likelihood computation is to recognize that the likelihood splits into simple factors based on a hidden Markov property of the underlying model. To expose this factorization property, assume that the loci $1, \ldots, m$ occur in numerical order along the chromosome. Let θ_i be the breakage probability on the interval connecting loci i and $i+1$, and suppose only loci $1 \leqslant t_1 < t_2 < \ldots < t_n \leqslant m$ are typed. Then

$$\Pr(\mathbf{X}) = \Pr(X_{t_1}) \prod_{i=2}^{n} \Pr(X_{t_i} | X_{t_1}, \ldots, X_{t_{i-1}}). \tag{4}$$

Now $\Pr(X_{t_1})$ is immediately available from (2). In the degenerate case $n = 1$, the product in (4) is taken as 1. Otherwise, the ith term in the above product satisfies the Markov property

$$\Pr(X_{t_i} | X_{t_1}, \ldots, X_{t_{i-1}}) = \Pr(X_{t_i} | X_{t_{i-1}}). \tag{5}$$

Indeed when $X_{t_i} = X_{t_{i-1}}$,

$$\Pr(X_{t_i} | X_{t_1}, \ldots, X_{t_{i-1}}) = \left[1 - \prod_{j=t_{i-1}}^{t_i-1} (1-\theta_j)\right] r^{X_{t_i}} (1-r)^{1-X_{t_i}} + \prod_{j=t_{i-1}}^{t_i-1} (1-\theta_j). \tag{6}$$

The first term on the right of (6) involves conditioning on at least one break between loci t_{i-1} and t_i. Here the retention fate of locus t_i is no longer tied to that of locus t_{i-1}. The second term involves conditioning on the complementary event. When $X_{t_i} \neq X_{t_{i-1}}$, we have the simpler expression

$$\Pr(X_{t_i} | X_{t_1}, \ldots, X_{t_{i-1}}) = \left[1 - \prod_{j=t_{i-1}}^{t_i-1} (1-\theta_j)\right] r^{X_{t_i}} (1-r)^{1-X_{t_i}}$$

since a break must occur somewhere between the two loci. These factorization results generalize to what Boehnke *et al.* (1991) call left-end point retention models.

Boehnke *et al.* (1991) also show how to implement the EM algorithm (Dempster *et al.* 1977) for maximum likelihood estimation of the m parameters $\theta_1, \ldots, \theta_{m-1}$ and r. Let $\gamma_i = \theta_i$ for $i = 1, \ldots, m-1$ and $\gamma_m = r$. All of these parameters can be viewed as success probabilities for binomial trials. As a consequence (Weeks & Lange, 1989), the EM updates take either of the equivalent generic forms

$$\begin{aligned}\gamma_i^{n+1} &= \frac{E(\text{no. of successes} \,|\, \text{obs}, \boldsymbol{\gamma}^n)}{E(\text{no. of trials} \,|\, \text{obs}, \boldsymbol{\gamma}^n)} \\ &= \gamma_i^n + \frac{\gamma_i^n (1-\gamma_i^n) \partial \ln [L(\text{obs}) \,|\, \boldsymbol{\gamma}^n)] / \partial \gamma_i}{E(\text{no. of trials} \,|\, \text{obs}, \boldsymbol{\gamma}^n)},\end{aligned}$$

where obs denotes the observations $\mathbf{X}$ over all clones, and L is the likelihood function. The second form of the update requires less thought since only mechanical differentiations are

involved in forming the score. If the number of clones is H, then the expected number of trials appearing in the denominator is H for θ_i and $H(1+\Sigma_{i=1}^{m-1}\theta_i^{n+1})$ for r. (Note that the formula in Boehnke *et al.* (1991) for the expected number of fragments erroneously employs θ_i^n instead of θ_i^{n+1}.) Since the amount of missing data can be relatively small, the EM algorithm makes for very fast optimization. On a 486 25 MHz computer, it takes about 0·5 s to maximize a 14-locus likelihood for the 99 clones of the Cox *et al.* (1990) data under a specific order.

Asymptotic standard errors for the parameter estimates are available as a by-product of the EM algorithm. Meng & Rubin (1991) show how to compute the observed information matrix by numerically differentiating the EM algorithm map. When there is no missing data, it is possible to compute explicitly the expected information matrix $\mathbf{J}$; according to classical large sample theory, $\mathbf{J}^{-1}$ provides approximate variances and covariances for the maximum likelihood estimates (Rao, 1973). In the case of two loci, denote $\Pr(X_1 = i, X_2 = j)$ by p_{ij} for brevity. Then a typical entry $J_{\alpha\beta}$ of $\mathbf{J}$ for a single clone is given by

$$J_{\alpha\beta} = \sum_{i=0}^{1}\sum_{j=0}^{1}\frac{1}{p_{ij}}\left[\frac{\partial p_{ij}}{\partial\alpha}\right]\left[\frac{\partial p_{ij}}{\partial\beta}\right]. \tag{7}$$

Straightforward, but tedious, calculations based on (3) and (7) yield the entries

$$\begin{aligned}
J_{\theta\theta} &= \frac{r(1-r)(2-\theta)}{(1-\theta r)\,\theta(1-\theta+\theta r)}\\
J_{\theta r} &= J_{r\theta}\\
&= \frac{(1-2r)(1-\theta)}{(1-\theta r)(1-\theta+\theta r)}\\
J_{rr} &= \frac{1}{r(1-r)}+\theta\left[\frac{1-r}{(1-\theta r)r}+\frac{r}{(1-r)(1-\theta+\theta r)}\right].
\end{aligned}$$

The expected information from H independent clones is $H\times\mathbf{J}$.

Note that the presence of the factor $1-2r$ in $J_{\theta r}$ renders the expected information matrix diagonal when $r=\frac{1}{2}$. The above expression for $J_{\theta\theta}$ also allows us to verify the intuitive notion that information on θ is maximized when $r=\frac{1}{2}$. It is obvious that $J_{\theta\theta}\to 0$ as $r\to 0$ or 1. To prove that $J_{\theta\theta}$ assumes its maximum at $r=\frac{1}{2}$ when θ is fixed, it suffices to show that $\ln J_{\theta\theta}$ has a unique stationary point at $r=\frac{1}{2}$. But this fact follows from

$$\begin{aligned}
\frac{\partial \ln J_{\theta\theta}}{\partial r} &= \frac{1}{r}-\frac{1}{1-r}+\frac{\theta}{1-\theta r}-\frac{\theta}{1-\theta+\theta r}\\
&= \frac{1-2r}{r(1-r)}-\frac{(1-2r)\,\theta^2}{(1-\theta r)(1-\theta+\theta r)}\\
&= \frac{(1-2r)(1-\theta)}{r(1-r)(1-\theta r)(1-\theta+\theta r)}.
\end{aligned}$$

As a rather crude bound we also have

$$\begin{aligned}
J_{\theta\theta} &\leqslant \frac{2r(1-r)}{\theta(1-\theta)}\\
&\leqslant \frac{1}{2\theta(1-\theta)}.
\end{aligned}$$

Thus, the asymptotic standard error for θ will be at least $\sqrt{2} \approx 1{\cdot}4$ times greater than that calculated for a simple binomial experiment with success probability θ.

For more than two loci, the likelihood factorization (4) and the Markovian property (5) lead to a more or less transparent generalization of the two locus results for $\mathbf{J}$. If in a given clone all m loci are typed, then the loglikelihood of the observation vector $\mathbf{X} = (X_1, \ldots, X_m)$ for the clone becomes

$$\ln[\Pr(\mathbf{X})] = \ln[\Pr(X_1)] + \sum_{i=1}^{m-1} \ln[\Pr(X_{i+1} \mid X_i)].$$

Now $\Pr(X_1)$ depends only on r, and $\Pr(X_{i+1} \mid X_i)$ depends only on r and the breakage probability θ_i for the ith interval. In fact, $\Pr(X_{i+1} \mid X_i)$ assumes exactly one of the four forms in (3), except that a factor of r or $1-r$ is missing in each instance. Upon taking logarithms and partial derivatives, it is clear that in computing $J_{\theta_i\theta_i}$ and $J_{\theta_i r}$ only the factor $\Pr(X_{i+1} \mid X_i)$ is relevant. Thus, our previous expressions for $J_{\theta\theta}$ and $J_{\theta r}$ carry over from the two locus case. From the representation

$$J_{\theta_i\theta_j} = E\left(-\frac{\partial^2 \ln[\Pr(X)]}{\partial\theta_i \partial\theta_j}\right),$$

we also deduce that the θ portion of $\mathbf{J}$ is diagonal. This makes it possible to invert $\mathbf{J}$ explicitly. For brevity we omit displaying $\mathbf{J}^{-1}$. In any event, the maximum likelihood estimates $\hat{\theta}_i$ of the various θ_i are asymptotically uncorrelated, and hence independent, as long as $r = \frac{1}{2}$ or r is not jointly estimated with the θ_i. Finally, our previous expression for J_{rr} must be amended to

$$J_{rr} = \frac{1}{r(1-r)} + \sum_{i=1}^{m-1} \theta_i \left[\frac{1-r}{r(1-\theta_i r)} + \frac{r}{(1-r)(1-\theta_i+\theta_i r)}\right].$$

A BAYESIAN METROPOLIS ALGORITHM

The maximum likelihood method provides no easily interpretable measure of the certainty that the order identified as best is, in fact, the correct order (Guerra *et al.* 1992). If the likelihood ratio comparing the best order to the next best order exceeds 1000, we might feel very comfortable in declaring the best order to be the correct order. However, this likelihood ratio does not translate into a significance test since one order is not a smoothly parametrized subhypothesis of the other order. Moderate likelihood ratios fall into a grey zone and leave us uncertain about the correct order of the loci. From a Bayesian perspective, these conceptual issues disappear. Specification of posterior probabilities for order is a natural consequence of adopting the Bayesian perspective.

In this section we will illustrate the application of a Metropolis algorithm (Metropolis *et al.* 1953; Kalos & Whitlock, 1986) for computing the posterior probability of locus order and the posterior mean distances between loci. Although the Metropolis algorithm predates the Gibbs sampler, it is not as well known in statistical circles. This is a pity since it complements other Monte Carlo techniques such as the Gibbs sampler (Geman & Geman, 1984), data augmentation (Tanner & Wong, 1987), and importance sampling (Rubin, 1988). The Metropolis algorithm is generally considered the inspiration for simulated annealing (Kirkpatrick *et al.* 1983). In common with simulated annealing, it involves setting up a Markov chain. However, instead of being used for combinatorial optimization, the Metropolis algorithm is tailored to sampling

from the equilibrium distribution of the Markov chain. If by construction the Markov chain is ergodic, then in the limit sample averages tend to expected values. With the correct definition of the chain, the equilibrium distribution will also coincide with the posterior distribution of the parameters given the data. This construction should be relevant to other Bayesian models.

As a prelude to our particular concrete problem, consider the following abstract scheme. We are given a parameter space Γ, whose typical point is denoted by γ. On this parameter space we impose a prior density $f(\gamma)$ with respect to some measure μ. In our particular example with m loci, Γ will be the m-dimensional unit cube $[0,1]^m$, $f(\gamma)$ will be identically 1, and μ will be Lebesgue measure. The component γ_i of γ gives the position of the ith locus. Note that confining the m loci to the unit interval $[0,1]$ entails a rescaling of distance for the postulated region occupied by the loci. This rescaling must be matched by a corresponding rescaling of the intensity λ of the underlying Poisson process. The Poisson process we now take to be homogeneous.

Returning to the abstract scheme, denote the observations upon which we base inferences by obs. These observations can be summarized by a likelihood function $L(\text{obs}\,|\,\gamma)$ depending jointly on obs and γ. To define a Metropolis algorithm on Γ we divide transitions into two stages, a proposal stage and an acceptance stage. For the proposal stage we postulate the existence of a proposal density $t(\gamma^*\,|\,\gamma)$ relative to μ for choosing the next point γ^* given the current point γ. The density $t(\gamma^*\,|\,\gamma)$ should satisfy $t(\gamma\,|\,\gamma^*) > 0$ whenever $t(\gamma^*\,|\,\gamma) > 0$. This requirement is intimately tied to the crucial property of detailed balance that will be discussed shortly. The acceptance stage determines whether a proposed move is actually taken. Acceptance occurs based on comparing the acceptance probability $\alpha(\gamma^*\,|\,\gamma)$ to a random number uniformly drawn from [0 1]. If the random number is less than $\alpha(\gamma^*\,|\,\gamma)$, then the proposed move is taken; otherwise the Markov chain declines the proposed move and remains in the current state. At the next step of the chain a new destination point is proposed, and the cycle of proposal and acceptance/non-acceptance is repeated. Eventually enough steps of the Markov chain are taken to permit accurate approximation of expectations by sample averages.

Properly defining $a(\gamma^*\,|\,\gamma)$ is the key to compelling the equilibrium distribution of the Markov chain to be the posterior density. Our definition is

$$a(\gamma^*\,|\,\gamma) = \min\left[1, \frac{L(\text{obs}\,|\,\gamma^*)f(\gamma^*)\,t(\gamma\,|\,\gamma^*)}{L(\text{obs}\,|\,\gamma)f(\gamma)\,t(\gamma^*\,|\,\gamma)}\right].$$

The posterior density of γ given obs amounts to

$$p(\gamma\,|\,\text{obs}) = \frac{L(\text{obs}\,|\,\gamma)f(\gamma)}{\int L(\text{obs}\,|\,\gamma')f(\gamma')\,d\mu(\gamma')}. \tag{8}$$

It is straightforward to show that $p(\gamma\,|\,\text{obs})$ satisfies the law of detailed balance:

$$p(\gamma\,|\,\text{obs})\,q(\gamma^*\,|\,\gamma) = p(\gamma^*\,|\,\text{obs})\,q(\gamma\,|\,\gamma^*), \tag{9}$$

where

$$q(\gamma^*\,|\,\gamma) = t(\gamma^*\,|\,\gamma)\,a(\gamma^*\,|\,\gamma)$$

is the transition density from γ to γ^*. Detailed balance says that at equilibrium, the rate of probabilistic flow from γ to γ^* matches the rate of probabilistic flow in the reverse direction from γ^* to γ. In proving detailed balance, we can assume without loss of generality that

$$L(\text{obs}\,|\,\gamma^*)f(\gamma^*)\,t(\gamma\,|\,\gamma^*) \leqslant L(\text{obs}\,|\,\gamma)f(\gamma)\,t(\gamma^*\,|\,\gamma).$$

Then invoking (8) yields

$$\begin{aligned} p(\gamma \mid \text{obs})\, q(\gamma^* \mid \gamma) &= p(\gamma \mid \text{obs})\, t(\gamma^* \mid \gamma) \frac{L(\text{obs} \mid \gamma^*) f(\gamma^*)\, t(\gamma \mid \gamma^*)}{L(\text{obs} \mid \gamma) f(\gamma)\, t(\gamma^* \mid \gamma)} \\ &= p(\gamma^* \mid \text{obs})\, t(\gamma \mid \gamma^*) \\ &= p(\gamma^* \mid \text{obs})\, q(\gamma \mid \gamma^*), \end{aligned}$$

proving detailed balance.

From detailed balance it is easy to deduce that $p(\gamma \mid \text{obs})$ is an invariant distribution for the Markov chain. (Some authors use the equivalent term stationary instead of invariant.) Indeed, integrating (9) against a bounded, continuous function $g(\gamma^*)$ gives

$$\int\int g(\gamma^*)\, q(\gamma^* \mid \gamma)\, d\mu(\gamma^*)\, p(\gamma \mid \text{obs})\, d\mu(\gamma) = \int\int g(\gamma^*)\, q(\gamma \mid \gamma^*)\, d\mu(\gamma)\, p(\gamma^* \mid \text{obs})\, d\mu(\gamma^*). \tag{10}$$

Now let

$$c(\gamma) = 1 - \int q(\gamma^* \mid \gamma)\, d\mu(\gamma^*)$$

be the probability of rejecting the proposed move from the current point γ. Adding

$$\int g(\gamma)\, c(\gamma)\, p(\gamma \mid \text{obs})\, d\mu(\gamma)$$

to both sides of (10) yields the equality

$$\int [g(\gamma)\, c(\gamma) + \int g(\gamma^*)\, q(\gamma^* \mid \gamma)\, d\mu(\gamma^*)]\, p(\gamma \mid \text{obs})\, d\mu(\gamma) = \int g(\gamma)\, p(\gamma \mid \text{obs})\, d\mu(\gamma). \tag{11}$$

If $\mathbf{U}_k$ is the position of the Markov chain at step k, then (11) can be restated in terms of expectations as

$$\begin{aligned} E[g(\mathbf{U}_{k+1})] &= E(E[g(\mathbf{U}_{k+1}) \mid \mathbf{U}_k]) \\ &= E[g(\mathbf{U}_k)]. \end{aligned}$$

But this last equality demonstrates that $\mathbf{U}_{k+1}$ and $\mathbf{U}_k$ have the same distribution, and this is precisely what invariance means.

The above brief description of the Metropolis method does not specify the nature of the proposal density $t(\gamma^* \mid \gamma)$. In our experience, it is crucial that the Markov chain be able to break away from a local maximum of $p(\gamma \mid \text{obs})$ and move across a deep valley to another local maximum. In the simulated annealing solution to the travelling salesman problem (Press *et al.* 1989), new permutations are proposed by taking an existing permutation, choosing at random a block of cities from it, and then inverting the order of cities within the block. For example, if the current permutation mandates visiting ten cities in the order 7–3–4–5–1–9–2–8–10–6, we might propose to invert the cities between positions 3 and 7 to give the new permutation 7–3–2–9–1–5–4–8–10–6. We will adopt the same tactic here with one minor modification. Because locus position is important as well as locus order, we also propose new positions for the inverted loci. If we decide to move n loci, the simplest scheme for doing this in a fashion consistent with their inverted order is to sample uniformly n points between the two flanking loci for the block – loci 3 and 8 in the above cities example – and position the inverted loci according to the order statistics of this uniform sample. If one end of the inverted block is the

first or last locus of the existing order, then the sampling interval extends to either 0 or 1, whichever is appropriate.

One could justifiably complain that the above proposal mechanism is *ad hoc*. Furthermore, unless the proposal block encompasses all loci, the proposal density is not even a legitimate density with respect to Lebesgue measure. We deal with this latter concern in the Appendix. Note that if the two endpoints of the inversion block coincide, we are simply repositioning the single locus chosen between its two flanking neighbours. Because of the wide variation in block size permitted, our proposal mechanism does effect a range of readjustments of the loci. One could contemplate other proposal mechanisms. For instance, one might randomly resample one locus at a time, but not restrict its new position to be consistent with the current order. This simpler resampling tactic would make it difficult to achieve the large scale rearrangements necessary to pass between the competing orders having large posterior probabilities. Once the chain reaches a favourable order, almost all proposed one locus rearrangements of the order will be insufficiently radical to be accepted.

Invariance of the posterior density is not enough. The further condition of ergodicity must be imposed to insure that sample averages taken over many consecutive steps of the Markov chain tend to expected values with respect to $p(\gamma \mid \text{obs})$ (Rosenblatt, 1971; Karlin & Taylor, 1975). The chain is ergodic if every invariant set of states has equilibrium probability 0 or 1, and a set of states is invariant if escape from the set is impossible under the transition mechanisms of the chain. Thus, ergodicity rules out all but trivial invariant sets. Now a sufficient condition for ergodicity can be stated by defining the n-step transition probability $\Pr^{(n)}(\gamma, A)$ (Rosenblatt, 1971). Starting from the point γ, this is the probability that the chain occurs in the set A after n steps. If for each A with $\mu(A) > 0$ it is possible to choose for μ-almost all γ an integer n_γ such that $\Pr^{(n_\gamma)}(\gamma, A) > 0$, then the chain is ergodic for the invariant density $p(\gamma \mid \text{obs})$. This condition is the analogue of complete communication among states for a discrete state Markov chain. Intuitively speaking, it requires that the transition mechanisms of our continuous state Markov chain sufficiently stir up the points. Part of the Appendix will be devoted to verifying this condition for ergodicity.

In our simulations of the posterior density, we hold the retention probability r fixed. Presumably, we could also put a prior on r and include it in the Metropolis procedure. However, r tends to be fairly well estimated by maximum likelihood, and no compelling evidence suggests a reasonable prior for r. In the Cox *et al.* (1990) data, r is close to 0·5, with some decline in progressing from the centromere to the telomere of the chromosome. In other data sets, r more often ranges from 0·2 to 0·3 (Burmeister *et al.* 1991; D. Cox, personal communication). The starting position of the Metropolis algorithm is also probably not critical. If the algorithm works well, equilibrium should be reached quickly. The best maximum likelihood order furnishes a possible starting point, but we might wish to avoid biasing the early steps of the chain toward the answer we wish to confirm.

ANOTHER BAYESIAN APPROXIMATION

The Monte Carlo approach of the last section is not the only method of computing the requisite integrals for posterior probabilities. A direct attack is possible if some simplifying approximations are made to reduce the complicated dependencies of the various integrands.

This direct method permits computation of the joint probability of the observations and of a particular order for the loci. Consequently, two competing orders can be compared by the ratio of their posterior probabilities. Unless the number of loci m is fairly small, and these joint probabilities can be summed over all orders, the direct method does not yield the normalizing constant for the posterior probability of any given order. Of course, if only a few orders have substantial posterior probability, then the normalizing constant can be well approximated by the sum over just these most likely orders.

To compute a joint probability Pr (obs, order), we note that the prior probability of the order is Pr (order) $= 1/m!$. Thus, it suffices to compute the conditional probability Pr (obs | order), and this can be accomplished by integrating over the possible distances between adjacent pairs of loci for the order. Without loss of generality, we again take this order to match the order of the recorded observations for each clone. We also again let δ_i be the distance between loci i and $i+1$. The δs are spacings from a uniform sample of m points on $[0, 1]$. Such spacings have been thoroughly studied (Shorack & Wellner, 1986). For instance, the δ_i are known to be exchangeable random variables with common density $m(1-\delta)^{m-1}$ on $[0, 1]$ and common mean $1/(m+1)$. The correlation of any pair δ_i and $\delta_j, j \neq i$, is just $-1/m$. These facts suggest that for large enough m, the δ_i are approximately independent with common approximate density $(m+1)\, e^{-(m+1)\delta}$. The mean of this exponential density obviously equals $1/(m+1)$.

Under these approximations,

$$\begin{aligned} \Pr(\text{obs} \mid \text{order}) &= \int_{\{\delta_1+\ldots+\delta_{m-1} \leq 1\}} L(\text{obs} \mid \boldsymbol{\delta}) \times \text{density}(\boldsymbol{\delta})\, d\boldsymbol{\delta} \\ &\approx \int_0^\infty \cdots \int_0^\infty L(\text{obs} \mid \boldsymbol{\delta}) \prod_{j=1}^{m-1} (m+1)\, e^{-(m+1)\delta_j} d\delta_1 \cdots d\delta_{m-1}, \end{aligned} \tag{12}$$

where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{m-1})$ and $d\boldsymbol{\delta} = d\delta_1 \cdots d\delta_{m-1}$.

To make further progress, we note that the likelihood $L(\text{obs} \mid \boldsymbol{\delta})$ factors into separate likelihoods for the independent clones. These clone likelihoods in turn factor in Markovian fashion according to (4) and (5). At this juncture it is helpful to introduce more manageable notation. Recall that for each clone only some of the loci are typed. In particular, suppose loci $t_{h1}, \ldots, t_{hn_h}$ are typed for clone h. The triples $(h, t_{hk}, t_{h,k+1})$ play a key role in what follows. We enumerate the set of triples

$$\bigcup_{h=1}^{H} \{(h, t_{hk}, t_{h,k+1}) : k = 1, \ldots, n_h - 1\}$$

in some order from $i = 1, \ldots, N$. Now suppose $(h, t_{hk}, t_{h,k+1})$ is the ith enumerated triple. Call the right locus $t_{h,k+1}$ of this triple right_i; define left_i similarly. We will require that the enumeration satisfy $\text{right}_i \leq \text{right}_{i+1}$. We can also characterize the span of the ith triple by introducing a row vector $\mathbf{v}^i$ having $m-1$ components and jth component defined by

$$v_j^i = \begin{cases} 1 & \text{if} \quad \text{left}_i \leq j < \text{right}_i \\ 0 & \text{otherwise.} \end{cases}$$

Finally, to simplify the notation for the conditional probabilities appearing in (5), define

$$\begin{aligned} a_{i0} &= r^{x_{ht_{h,k+1}}} (1-r)^{1-x_{ht_{h,k+1}}} \\ a_{i1} &= \chi_{\{x_{ht_{h,k+1}} = x_{ht_{hk}}\}} - a_{i0}, \end{aligned}$$

where $X_{ht_{hk}}$ and $X_{ht_{h,k+1}}$ are the two relevant observations for the triple and χ is an indicator function.

This glut of notation is actually useful. For instance for the ith triple, equation (5) and its immediate aftermath reduce via equation (1) to

$$\Pr(X_{ht_{h,k+1}} | X_{ht_{hk}}) = a_{i0} + a_{i1} e^{-\lambda \Sigma_{j=1}^{m-1} v_j^i \delta_j},$$

and the likelihood becomes

$$L(\text{obs} | \boldsymbol{\delta}) = \prod_{h=1}^{H} \Pr(X_{ht_{h1}}) \prod_{i=1}^{N} [a_{i0} + a_{i1} e^{-\lambda \Sigma_{j=1}^{m-1} v_j^i \delta_j}].$$

We can insert this representation of $L(\text{obs} | \boldsymbol{\delta})$ into our approximation (12) of Pr(obs, order). Using the distributive law to choose for each i either the factor a_{i0} or the factor $a_{i1} e^{-\lambda \Sigma_{j=1}^{m-1} v_j^i \delta_j}$, it is possible to evaluate all the 1-dimensional integrals explicitly. Indicating the first choice by $j_i = 0$ and the second choice by $j_i = 1$, the integral over δ_k can be evaluated as

$$\int_0^\infty e^{-(\lambda \Sigma_{i=1}^{N} j_i v_k^i + m + 1)\delta_k} d\delta_k = \frac{1}{\lambda \Sigma_{i=1}^{N} j_i v_k^i + m + 1}.$$

This yields the approximation

$$\Pr(\text{obs} | \text{order}) \approx b \sum_{j_1=0}^{1} \cdots \sum_{j_N=0}^{1} \prod_{l=1}^{N} a_{l j_l} \prod_{k=1}^{m-1} \frac{1}{\lambda \Sigma_{i=1}^{N} j_i v_k^i + m + 1}, \tag{13}$$

where the leading constant is

$$b = (m+1)^{m-1} \prod_{h=1}^{H} \Pr(X_{ht_{h1}}).$$

The formidable multiple sum (13) can be evaluated by recursively evaluating the intermediate sums

$$S_n = b \sum_{j_1=0}^{1} \cdots \sum_{j_n=0}^{1} \prod_{l=1}^{n} a_{l j_l} \prod_{k=1}^{\text{right}_n - 1} \frac{1}{\lambda \Sigma_{i=1}^{N} j_i v_k^i + m + 1}.$$

Before we specify an algorithm, let us clarify some of the issues involved. First, the second product of S_n,

$$\prod_{k=1}^{\text{right}_n - 1} \frac{1}{\lambda \Sigma_{i=1}^{N} j_i v_k^i + m + 1}, \tag{14}$$

extends only from 1 to $\text{right}_n - 1$. This limited range is motivated by our desire to involve as few loci as possible in each intermediate sum and accounts for the convention $\text{right}_i \leqslant \text{right}_{i+1}$ in enumerating triples. Second, the displayed indices $j_1, \ldots, j_n$ do not fully determine the sums $\lambda \Sigma_{i=1}^{N} j_i v_k^i + m + 1$ in the denominators of the product; the partial sums $w_k = \Sigma_{i=n+1}^{N} j_i v_k^i$ depend on the unspecified indices $j_{n+1}, \ldots, j_N$. Let $\mathbf{w}$ be the vector with components the partial sums $w_k = \Sigma_{i=n+1}^{N} j_i v_k^i$, $k+1, \ldots, \text{right}_n - 1$, generated by a specific choice of the indices $(j_{n+1}, \ldots, j_N)$. It often happens that two different choices of the unspecified indices determine the same vector $\mathbf{w}$. For instance, an index $j_i, i > n$, is irrelevant if all $v_k^i = 0$ for $k \leqslant \text{right}_n - 1$. The crux of the matter is that S_n depends only on the partial sums w_k that are the components of $\mathbf{w}$ and not

the particular indices $(j_{n+1}, \ldots, j_N)$ generating these partial sums. We will write $S_n(\mathbf{w})$ to emphasize this dependence. Now denote by $\mathscr{C}_n$ the collection of distinct vectors $\mathbf{w}$ arising from one or more choices of $(j_{n+1}, \ldots, j_N)$. At each stage n, we will need to calculate and store $S_n(\mathbf{w})$ for each $\mathbf{w} \in \mathscr{C}_n$. The cardinality of $\mathscr{C}_n$ generally should be far less than the number 2^{N-n} of index sets $(j_{n+1}, \ldots, j_N)$.

Let us next consider how the collections $\mathscr{C}_n$ might be constructed efficiently by a backwards recurrence. In this process, we will construct the $\mathscr{C}_n$ in groups, with each group corresponding to a different value of right_n. First we construct $\mathscr{C}_1$ and all remaining $\mathscr{C}_n$ having $\text{right}_n = \text{right}_1 = 2$, then we construct the group of $\mathscr{C}_n$ having $\text{right}_n = 3$, and so forth. Note here we implicitly assume that all numbers from 2 to m are represented among the values of right_n. If this were not the case, then some locus would never be typed, or if it were, then it would never be typed simultaneously with any other locus. Obviously, it would be impossible to order such a locus relative to the other loci.

Now suppose $n = 1$ or $\text{right}_{n-1} < \text{right}_n$. To obtain $\mathscr{C}_n$ and all other $\mathscr{C}_j$ with $\text{right}_j = \text{right}_n$, observe that among the triples from $n+1$ to N, there is some subsequence $k_1, \ldots, k_e$ defined by the requirement $\text{left}_{k_j} < \text{right}_n$. Exactly these triples contribute to $\mathscr{C}_n$ since the corresponding spanning vectors $\mathbf{v}^{k_j}$ have 1's entry $\text{right}_n - 1$. Truncate $\mathbf{v}^{k_j}$ by deleting all but its first $\text{right}_n - 1$ entries. The resulting vector, $\mathbf{u}^{k_j}$, is an element of $\mathscr{C}_n$. We begin the recurrence with $\mathscr{D}_{k_e} = \{\mathbf{0}\}$, where now $\mathbf{0}$ is the 0 vector with $\text{right}_n - 1$ entries. We then recursively define the further collections

$$\mathscr{D}_{k_{j-1}} = \mathscr{D}_{k_j} \bigcup \{\mathbf{u}^{k_j} + \mathbf{w} : \mathbf{w} \in \mathscr{D}_{k_j}\},$$

where $\mathbf{u}^{k_j} + \mathbf{w}$ denotes the obvious vector sum. We can keep a single list of vectors for the collections $\mathscr{D}_{k_j}$ by adding the new vectors of $\mathscr{D}_{k_{j-1}}$ to the bottom of the list for $\mathscr{D}_{k_j}$. All intermediate collections are then readily available in the final list $\mathscr{D}_{k_0}$, which coincides with $\mathscr{C}_n$. Proceeding forward from $\mathscr{C}_n$, the collections $\mathscr{C}_{k_j}$ and $\mathscr{D}_{k_j}$ coincide as long as right does not jump. For instance, $\mathscr{C}_{n+1}$ coincides with $\mathscr{D}_{n+1}$ provided $\text{right}_{n+1} = \text{right}_n$. Once right jumps, $\mathscr{C}_{k_j}$ and $\mathscr{D}_{k_j}$ contain vectors of different lengths. It is also clear from this construction that within the group the inclusion $\mathscr{C}_j \subset \mathscr{C}_{j+1}$ always holds. Subsequent groups can be constructed by the same backwards recurrence until $\mathscr{C}_N = \{\mathbf{0}\}$ is reached.

To give a concrete illustration of the above construction, suppose that there are $m = 3$ loci and $H = 4$ clones. Taking the same order for observations and loci, suppose the clones are $(1, ?, 0)$, $(0, ?, 1)$, $(1, 1, 1)$, and $(?, 0, 1)$. Clones 1, 2, and 4 generate 1 spanning vector each. These are respectively, $\mathbf{v}^2 = (1, 1)$, $\mathbf{v}^3 = (1, 1)$, and $\mathbf{v}^5 = (0, 1)$. Clone 3 generates the 2 spanning vectors $\mathbf{v}^1 = (1, 0)$ and $\mathbf{v}^4 = (0, 1)$. The collections generated by the backward recurrence relations are:

$$\begin{aligned}
\mathscr{C}_1 &= \{(0), (1), (2)\} \\
\mathscr{C}_2 &= \{(0,0), (0,1), (0,2), (1,1), (1,2), (1,3)\} \\
\mathscr{C}_3 &= \{(0,0), (0,1), (0,2)\} \\
\mathscr{C}_4 &= \{(0,0), (0,1)\} \\
\mathscr{C}_5 &= \{(0,0)\}.
\end{aligned}$$

The initial sum $S_1(\mathbf{w})$ is trivial to compute for each $\mathbf{w} \in \mathscr{C}_1$ since it involves only one summation. To define a forward recurrence relation between the sets of sums S_{n-1} and S_n, suppose first that $\text{right}_{n-1} = \text{right}_n$. Then $\mathscr{C}_{n-1}$ and $\mathscr{C}_n$ are collections of vectors of equal

lengths. As above, let $\mathbf{u}^n$ be the vector of length $\text{right}_n - 1$ agreeing with the first $\text{right}_n - 1$ entries of the spanning vector $\mathbf{v}^n$. If $\mathbf{w}$ belongs to $\mathscr{C}_n$, then both $\mathbf{w}$ and $\mathbf{u}^n + \mathbf{w}$ belong to $\mathscr{C}_{n-1}$. The recurrence relation

$$S_n(\mathbf{w}) = a_{n0} S_{n-1}(\mathbf{w}) + a_{n1} S_{n-1}(\mathbf{u}^n + \mathbf{w})$$

follows immediately from the distributive law.

When $\text{right}_n = \text{right}_{n-1} + 1$, the vectors in $\mathscr{C}_n$ have one more entry than vectors in $\mathscr{C}_{n-1}$. Also the product (14) has the extra factor for $k = \text{right}_n - 1$ lacking in the corresponding product for S_{n-1}. These considerations dictate the recurrence relation

$$S_n(\mathbf{w}) = \frac{a_{n0} S_{n-1}(\mathbf{w}^*)}{\lambda w_{\text{right}_n - 1} + m + 1} + \frac{a_{n1} S_{n-1}([\mathbf{u}^n + \mathbf{w}]^*)}{\lambda(1 + w_{\text{right}_n - 1}) + m + 1}, \tag{15}$$

where $\mathbf{w} \in \mathscr{C}_n$ and where $\mathbf{w}^*$ and $[\mathbf{u}^n + \mathbf{w}]^*$ are $\mathbf{w}$ and $\mathbf{u}^n + \mathbf{w}$, respectively, truncated to $\text{right}_n - 2$ components. It is important to note that the denominators in this recurrence relation are correct since $v^i_{\text{right}_n - 1} = 0$ for $1 \leqslant i < n$, and hence

$$\sum_{i=1}^{n-1} j_i v^i_{\text{right}_n - 1} = 0.$$

Continuing the above simple example with $m = 3$ loci and $H = 4$ clones, the initialization of S_1 and the two forms of the recurrence can be illustrated by:

$$S_1((2)) = b\left[\frac{a_{10}}{2\lambda + 4} + \frac{a_{11}}{3\lambda + 4}\right]$$

$$S_2((1,1)) = \frac{a_{20} S_1((1))}{\lambda + 4} + \frac{a_{21} S_1((2))}{2\lambda + 4}$$

$$S_3((0,1)) = a_{30} S_2((0,1)) + a_{31} S_2((1,2)).$$

The major barrier to computing with the above recurrence relations is the size of the collections $\mathscr{C}_n$. If there are no missing data, then $\mathscr{C}_n$ has at most cardinality H. In this case, each spanning vector has exactly one non-zero entry, and the number of triples $k_1, \ldots, k_e$ in the backwards construction of $\mathscr{C}_n$ is at most $e = H - 1$. Each vector $\mathbf{w} \in \mathscr{C}_n$ has the form $\mathbf{w} = (0, \ldots, 0, i)$ for $0 \leqslant i \leqslant H - 1$. With a modest amount of missing data, assessing the maximum cardinality of $\mathscr{C}_n$ is more complex, but it is still possible to carry out the recursive computations.

Note finally that the above recurrences can easily be adapted to computing approximate moments of the δ_i conditional on the observations and a given order. For instance, to approximate the conditional mean of δ_k, note that

$$\int_0^\infty \delta_k e^{-(\lambda \Sigma_{i-1}^N j_i v_k^i + m + 1)\delta_k} d\delta_k = \frac{1}{(\lambda \Sigma_{i-1}^N j_i v_k^i + m + 1)^2}.$$

This dictates substituting the factor

$$\frac{1}{(\lambda \Sigma_{i-1}^N j_i v_k^i + m + 1)^2}$$

for the factor $$\frac{1}{\lambda\Sigma_{i=1}^{N} j_i v_k^i + m + 1}$$

wherever it occurs in the multiple sum S_N and the intermediate sums S_n. In particular, the denominators in the recurrence (15) should be squared when $\text{right}_n - 1 = k$. When $k = 1$, the initial sums rather than the recurrences must be changed correspondingly. These substitutions yield a new final sum that can be normalized by the original S_N to produce the approximate conditional mean.

OPTIMAL DESIGN OF RADIATION LEVELS

Our earlier computation of the expected information matrix was based on breakage probabilities θ_i rather than physical distances δ_i. To convert an information matrix entry from the $\boldsymbol{\theta}$ parameterization to the $\boldsymbol{\delta}$ parameterization, it is simplest to employ the equality

$$\begin{aligned}\frac{\partial \ln L}{\partial \delta_i} &= \frac{\partial \ln L}{\partial \theta_i}\frac{\partial \theta_i}{\partial \delta_i} \\ &= \frac{\partial \ln L}{\partial \theta_i}\lambda e^{-\lambda\delta}.\end{aligned}$$

For instance,

$$J_{\delta_i\delta_i} = \frac{r(1-r)\,(y_i e^{-y_i})^2\,(1+e^{-y_i})}{\delta_i^2\,(1-r+e^{-y_i}r)\,(1-e^{-y_i})\,(r+e^{-y_i}[1-r])} \tag{16}$$

with $y_i = \lambda\delta_i$. (The quantity y_i measures distance in expected numbers of breaks.) If $r = \frac{1}{2}$ or r is not estimated, the standard error $\sqrt{[\text{Var}\,(\hat{\delta})]}$ of the maximum likelihood estimate $\hat{\delta}_i$ of δ_i is asymptotically $(J_{\delta_i,\delta_i})^{-\frac{1}{2}}$. Because this asymptotic standard error does not depend on the particular locus i chosen, it is convenient to drop the subscript i in the following arguments.

The primary concern of most geneticists will be to find the correct order of the loci. The order of a locus relative to nearby loci will be poorly resolved if its estimated distance to one of its two flanking neighbours is small compared with the asymptotic standard error of the estimated distance. This argues that minimizing the asymptotic coefficient of variation $\sqrt{[\text{Var}\,(\hat{\delta})]}/\delta \approx (\delta^2 J_{\delta\delta})^{-\frac{1}{2}}$ over the likely range of δ is more appropriate than minimizing the asymptotic standard error over the same range.

For the moment let us take δ_i as approximately known. The function $\delta^2 J_{\delta\delta}$, which depends only on $y = \lambda\delta$ and r, is somewhat easier to deal with than the asymptotic coefficient of variation $(\delta^2 J_{\delta\delta})^{-\frac{1}{2}}$. Suppose the maximum of $\delta^2 J_{\delta\delta}$ as a function of y occurs at $y_{\max}(r)$. Then the optimal intensity λ for a given distance δ and retention probability r is $y_{\max}(r)/\delta$. Figure 1 depicts a numerical solution for $y_{\max}(r)$. Evidently from this figure, $y_{\max}(r)$ is a rather flat function of r symmetric around $r = \frac{1}{2}$. At $r = \frac{1}{2}$, $y_{\max} = 0{\cdot}80$, and as $r \to 0$ or 1, $y_{\max} \to 1{\cdot}21$. At $r = \frac{1}{4}$ or $\frac{3}{4}$, $y_{\max} = 0{\cdot}85$. These considerations suggest that if we know δ approximately, then any reasonable choice of λ should satisfy $0{\cdot}8/\delta < \lambda < 0{\cdot}9/\delta$, corresponding to the restriction $0{\cdot}55 < \theta < 0{\cdot}59$. Since in the Cox *et al.* (1990) data $r \approx \frac{1}{2}$, and this choice of r renders the joint information matrix for the δ_i and r diagonal, we will feature the lower value $0{\cdot}8/\delta$ for λ in the remaining arguments.

For m randomly distributed loci on $[0, 1]$, the optimal choice of λ is a more subtle problem.

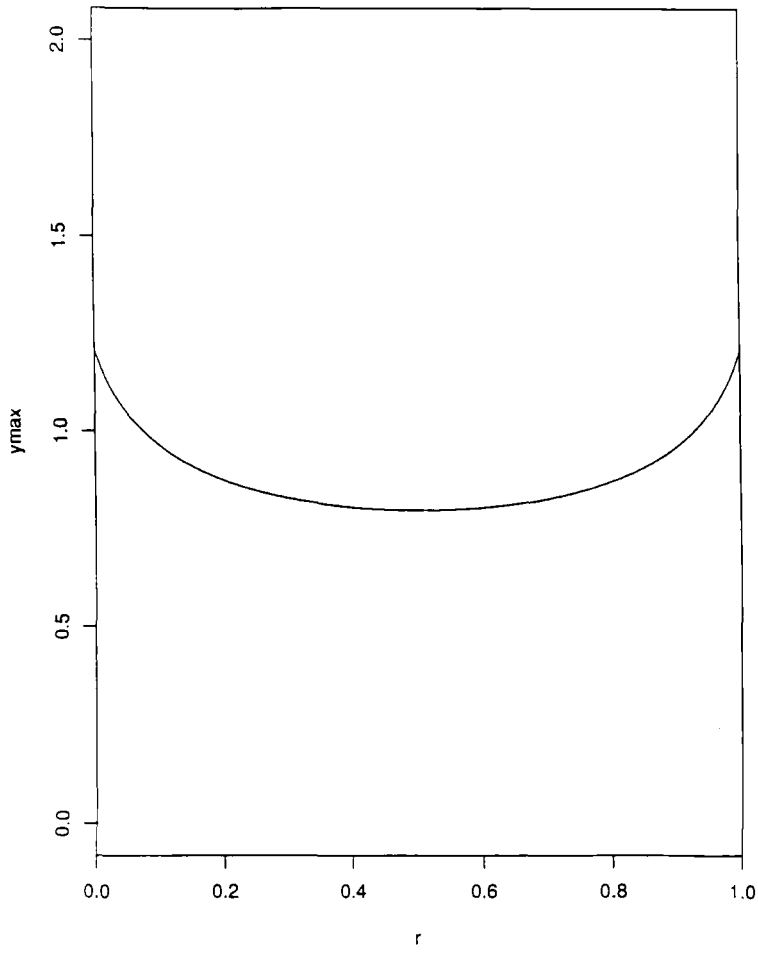

Fig. 1. The optimal expected number of breaks y_{max} between two adjacent loci as a function of the fragment retention probability r. The physical distance δ between the loci is assumed known.

The average distance between adjacent pairs of loci is $1/(m+1)$, and one might be tempted to take $\lambda \approx 0{\cdot}8\ (m+1)$. However, the average distance between adjacent loci may be far from the maximum and minimum distances between adjacent loci. Standard arguments from geometric probability theory show that the average maximum distance between adjacent loci is

$$\frac{1}{m+1}\left[1+\tfrac{1}{2}+\ldots+\frac{1}{m-1}\right] \approx \frac{\ln(m-1)}{m+1}$$

(Read, 1988). This is not too alarming, but the average minimum distance is only $[(m+1)(m-1)]^{-1}$ (Read 1988). Thus, to estimate the minimum distance between adjacent loci well, it would be better to take $\lambda \approx 0{\cdot}8\ (m+1)(m-1)$. Such a large value of λ would generate many small fragments in a typical clone, and order for the more distantly spaced loci would be hard to resolve. Clearly, some balance must be struck if a single intensity λ is used in irradiating all clones.

One possibility might be to choose λ to minimize the average value of the coefficient of variation $\sqrt{[\mathrm{Var}\,(\hat{\delta})]}/\delta$. As noted earlier, if m points are chosen randomly from $[0, 1]$, then the

Table 1. *Minima of the average coefficient of variation (17) and average standard error (18) for single-dose designs*

Number of loci m	Minimum value of (17)	Optimal λ for (17)	Minimum value of (18)	Optimal λ for (18)
3	3·54	3·26	0·692	1·98
4	3·62	4·00	0·560	2·36
5	3·68	4·73	0·471	2·74
6	3·72	5·44	0·406	3·12
8	3·77	6·86	0·319	3·87
10	3·81	8·27	0·263	4·62
12	3·84	9·66	0·223	5·37
14	3·86	11·1	0·194	6·11
16	3·88	12·4	0·172	6·86
18	3·89	13·8	0·154	7·60
20	3·90	15·2	0·140	8·34
30	3·94	22·0	0·095	12·1
40	3·96	28·9	0·072	15·8
50	3·97	35·7	0·058	19·5

spacing δ between any two adjacent points has density $m(1-\delta)^{m-1}$. Further, when $r = \frac{1}{2}$ and $y = \lambda\delta$,

$$J_{\delta\delta} = \frac{(ye^{-y})^2}{\delta^2(1-e^{-2y})}$$

$$= \frac{y^2}{\delta^2(e^{2y}-1)}.$$

Minimizing the average coefficient of variation consequently requires minimizing

$$\int_0^1 \frac{\sqrt{(e^{2\lambda\delta}-1)}}{\lambda\delta} m(1-\delta)^{m-1}\,d\delta. \tag{17}$$

Note that the integrand of (17) behaves like a constant times $\delta^{-\frac{1}{2}}$ for δ near 0. To tame this singularity, one can make the change of variable $s = \delta^{\frac{1}{2}}$ in (17). Once this is done, the required integrations and minimization can be done by standard numerical techniques. Table 1 lists the optimal intensity λ for numbers of loci m between 3 and 50 and retention probability $r = \frac{1}{2}$. It is evident from the table that the optimal λ is nearly a linear function of m for moderately large m. The regression equation $\lambda_{\text{opt}} = 1{\cdot}333 + 0{\cdot}689\,m$ provides an excellent fit. For comparison to the previously suggested value $\lambda = 0{\cdot}8\,(m+1)$, the best fitting linear regression for λ_{opt} with equal slope and intercept is $\lambda_{\text{opt}} = 0{\cdot}71\,(m+1)$. One should also bear in mind that (17) is a rather flat function of λ in the vicinity of the optimal λ. For instance, the integral (17) is within 10% of its optimum for $m = 5$ loci throughout the interval (2·7, 6·8). For 10 loci and 20 loci, the corresponding intervals are (4·9, 11·6) and (9·2, 20·8), respectively.

For the sake of comparison, the optimal intensity λ to minimize the average standard deviation of $\hat{\delta}$,

$$\int_0^1 \frac{\sqrt{(e^{2\lambda\delta}-1)}}{\lambda} m(1-\delta)^{m-1}\,d\delta, \tag{18}$$

is also listed in Table 1. Again the optimal λ is nearly linear in m; $\lambda_{\text{opt}} = 0{\cdot}884 + 0{\cdot}373\,m$ provides a very close fit. As anticipated, the optimal intensity λ for (17) tends to be quite a bit

Table 2. *Accuracy of the maximum likelihood criterion for ordering ten random loci on* $[0, 1]$ *when the retention probability* $r = 0{\cdot}5$

(Results based on 500 trials per intensity.)

Empirical probability of correct identification	Standard error of empirical probability	Intensity λ
0·505	0·022	1·0
0·817	0·017	2·0
0·908	0·013	3·0
0·914	0·013	4·0
0·924	0·012	5·0
0·924	0·012	6·0
0·891	0·014	7·0
0·864	0·015	8·0
0·820	0·017	9·0
0·817	0·017	10·0

larger than that for (18). Surprisingly, the optimal λ for (17) is slightly smaller than the $0{\cdot}8\,(m+1)$ value associated with the average interval length $1/(m+1)$. We expected that small distances δ would predominate over large distances δ in determining the optimal λ.

Since there are sizable differences between the optimal intensities determined by the two criteria, and since our analytic results are limited to pairs of loci, we undertook a simulation study to compare the criteria. We simulated experiments using either four or ten loci and retention probabilities of either 0·2 or 0·5. For the four-locus experiments, it was possible to evaluate the maximum likelihood for all $4!/2 = 12$ locus orders for a given simulation trial. For the ten-locus experiments, the maximum likelihood orders were obtained using a stepwise locus-ordering algorithm (Barker *et al.* 1987; Boehnke *et al.* 1991). In this algorithm locus orders are built one locus at a time, with a partial locus order discarded whenever its maximum likelihood is at least k times smaller than the most likely partial order constructed from the same set of loci. For the simulation study, we set $k = 10^5$. In those trials in which the true order was one of t orders tied for the maximum likelihood, credit $1/t$ was given to the true order.

Results of the simulation for $m = 10$ loci and retention probability $r = 0{\cdot}5$ are shown in Table 2. These results suggest that maximum ordering accuracy is obtained for $3{\cdot}0 < \lambda < 7{\cdot}0$, but that there is little difference in accuracy of ordering throughout the interval $3{\cdot}0 < \lambda < 7{\cdot}0$. The optimal intensity lies between the intensities $4{\cdot}61 = 0{\cdot}88 + 0{\cdot}373 \times 10$ and $8{\cdot}22 = 1{\cdot}33 + 0{\cdot}689 \times 10$ suggested by the minimum standard error and the minimum coefficient of variation criteria, respectively. Contrary to what we anticipated, the optimal intensity λ from the minimum standard error criterion is somewhat closer to the optimal simulation intensity λ. For retention probability $r = 0{\cdot}2$, the optimal intensity and the acceptable intensity range were essentially the same as for $r = 0{\cdot}5$; however, probabilities of correct ordering were substantially reduced, being in no case greater than 0·77. Results for four loci were qualitatively the same. The optimal intensity λ was intermediate between the values predicted by minimum standard error and coefficient of variation criteria, but closer to the former, and there was a broad range of intensities that gave essentially equal probabilities of accurate ordering.

In designing experiments, one can usually do better with a mixture of experiments than with any one simple experiment (Chernoff, 1979). Let us therefore consider the effect of two radiation doses. A fraction α of all clones could be exposed to a lower dose λ_1, and the remaining fraction

Table 3. *Minima of the average coefficient of variation (19) for two-dose designs*

Number of loci m	Minimum value of (19)	Optimal λ_1 for (19)	Optimal λ_2 for (19)	Optimal α for (19)
3	3·43	3·00	99·5	0·922
4	3·50	3·63	98·8	0·906
5	3·54	4·24	101·	0·894
6	3·57	4·85	104·	0·883
8	3·61	6·05	113·	0·866
10	3·63	7·22	123·	0·854
12	3·65	8·39	134·	0·844
14	3·66	9·55	145·	0·837
16	3·67	10·7	156·	0·830
18	3·68	11·9	168·	0·825
20	3·69	13·0	179·	0·821
30	3·71	18·7	238·	0·806
40	3·72	24·4	298·	0·797
50	3·72	30·1	358·	0·792

$1-\alpha$ could be exposed to a higher dose λ_2. For this two-dose experiment, the average information entry $J_{\delta\delta}$ over all H clones becomes

$$J_{\delta\delta} = \frac{\alpha\lambda_1^2}{e^{2\lambda_1\delta}-1} + \frac{(1-\alpha)\lambda_2^2}{e^{2\lambda_2\delta}-1}.$$

Minimizing the average coefficient of variation of $\hat{\delta}$ now amounts to minimizing

$$\int_0^1 \left[\frac{\alpha\lambda_1^2}{e^{2\lambda_1\delta}-1} + \frac{(1-\alpha)\lambda_2^2}{e^{2\lambda_2\delta}-1}\right]^{-\frac{1}{2}} \frac{m(1-\delta)^{m-1}}{\delta}\, d\delta, \tag{19}$$

relative to the admixture parameter α and the two intensities λ_1 and λ_2. Our results for this delicate optimization problem are displayed in Table 3. The entries in this table suggest that the low dose be about 5% less than that prescribed by the single-dose optimal design. The high dose should be more than an order of magnitude higher and should be applied to 10–20% of the clones. Both the low dose and the high dose are nearly linear in m. Unfortunately from a comparison of Tables 1 and 3, it appears that only modest gains of about 5% in ordering efficiency can be expected from two-dose radiation designs. Major gains in ordering efficiency from three or four-dose designs are unlikely. Perhaps sequential designs would yield better results, but rigorous mathematical proof of their superiority would be formidable.

The above pessimistic conclusions about two-dose designs are reinforced by parallel computations for the average standard deviation of $\hat{\delta}$. Minimizing this criterion with respect to the admixture parameter α and the two intensities λ_1 and λ_2 yields the same results as the single-dose designs. The parameter α is driven to 1, and the intensity λ_1 is driven to the values displayed in the rightmost column of Table 1.

APPLICATION

We applied the two Bayesian methods for computing posterior probabilities of locus order to the radiation hybrid data of Richard *et al.* (1991). These data for the proximal long arm of human chromosome 11 involve $m = 16$ markers typed on $H = 101$ hybrids. The data are nearly complete with each marker typed on at least 99 hybrids. The markers MTC, P11EH, HSTF1,

Table 4. *Best locus order compared by likelihood (L) and Posterior probability (P) ratios using the abbreviations from Table 5*

(Rank refers to the maximum likelihood rank. The symbol * indicates that the order was never visited by the Metropolis algorithm. Single underlines indicate block inversions with respect to the most likely locus order; double underlines indicate more complex rearrangements.)

Rank	Locus order	Maximum likelihood L_1/L_i	Metropolis P_1/P_i	Metropolis P_i	Approximate integral P_1/P_i	Approximate integral P_1
1	1 2 3 4 5 6 7 8 9 10 11 12	1	1	0·9899	1	0·9849
2	1 2 3 4 5 6 7 8 9 10 12 11	90	458	0·0022	138	0·0072
3	1 2 3 4 5 6 7 9 10 8 11 12	137	138	0·0072	151	0·0065
4	1 2 3 4 5 6 7 8 10 9 11 12	774	1650	0·0006	1253	0·0008
5	1 2 3 4 5 6 7 8 11 12 10 9	3184	*	0·0000	5533	0·0002
6	1 2 3 4 5 6 8 7 9 10 11 12	3796	5823	0·0002	5266	0·0002
7	1 2 3 4 5 6 7 8 12 11 10 9	5492	*	0·0000	9209	0·0001
8	1 2 3 5 4 6 7 8 9 10 11 12	12599	*	0·0000	15577	0·0001
9	1 2 3 4 5 6 7 9 10 8 12 11	21155	*	0·0000	34535	0·0000
10	1 2 3 4 5 6 7 8 10 9 12 11	24939	*	0·0000	64166	0·0000
11	1 2 3 4 5 6 8 7 9 10 12 11	339680	*	0·0000	724223	0·0000
12	1 2 3 5 4 6 7 8 9 10 12 11	1134400	*	0·0000	21411811	0·0000

Table 5. *Locus positions for the best locus order in the Richard* et al. *(1991) data*

Map lengths for the Metropolis and approximate integral methods have been normalized to 2·666 R_{9000}, the maximum likelihood map length.

Locus	Abbrev.	Cox *et al.* position	Max. likelihood position	Metropolis position	Approx. integral position
CINH	1	0·00	0·000	0·000	0·000
OSBP	2	0·50	0·415	0·395	0·374
CD5	3	0·73	0·622	0·608	0·582
PGA	4	0·87	0·749	0·747	0·722
FTH6L	5	1·01	0·876	0·887	0·866
COX8	6	1·42	1·258	1·252	1·235
PYGM	7	1·65	1·489	1·486	1·474
SEA	8	1·85	1·694	1·696	1·695
KRN1	9	2·12	1·968	1·967	1·967
MTC	10	2·26	2·108	2·120	2·123
GST3	11	2·63	2·517	2·505	2·503
PP1a	12	2·77	2·666	2·666	2·666

and INT2 were concordant in all hybrids. Since such markers cannot be ordered relative to one another, all but MTC were excluded from our analysis. Markers CD5 and CD20 also were always concordant; CD20 was excluded. Thus, we analysed data on $m = 12$ markers.

To obtain a set of locus orders for Bayesian analysis, we use the stepwise algorithm described above and in more detail in Boehnke *et al.* 1991. With $k = 10^{20}$, we identified 12 locus orders with maximum likelihoods no more than 10^{10} times smaller than that of the best maximum likelihood order. The best maximum likelihood order coincides with the one arrived at by Richard *et al.* (1991) using the two-point method-of-moments ordering strategy of Cox *et al.* (1990). Table 4 presents the 12 best maximum-likelihood orders identified, together with their likelihood ratios relative to the best order. For example, the second ranked maximum likelihood order – obtained from the best ranked order by inverting loci GST3 and PP1a – had maximum likelihood 90 times smaller than that of the best order. Table 5 displays the estimated locus positions under the best maximum likelihood order.

To prime the Metropolis method, we started with a randomly chosen locus order and then ignored the results of the first 10^5 steps. We estimated locus order probabilities as sample proportions of their occurrences in the next 10^6 steps. Table 4 presents the posterior probabilities for the various orders as well as the ratios of these posterior probabilities relative to the posterior probability of the best order.

The best maximum likelihood order also was the most probable order under the Metropolis method with an estimated posterior probability of 99·0%. The next three most likely locus orders were the same as under maximum likelihood, although the relative ranks of the second and third best orders were reversed. These three orders together had an estimated total posterior probability of 1·0%. Only one other locus order had an estimated posterior probability greater 0. The fact that seven of the most likely locus orders were never visited suggests that the Metropolis algorithm may need to sample far more than 10^6 iterates if accurate posterior probabilities beyond the first few locus orders are desired. For the locus orders visited, posterior probability ratios were not strikingly different from the corresponding maximum likelihood ratios (Table 4). Carrying out the Metropolis analysis required about 16 h on our 486 33 MHz computer.

The results for the approximate integral method also are summarized in Tables 4 and 5. In Table 4 the ranks of the 12 best orders coincide with those under maximum likelihood. It is noteworthy that the posterior probability ratios are uniformly larger than the maximum likelihood ratios. Thus, the approximate integral method provides stronger support for the best order then does maximum likelihood. The most probable order has approximate posterior probability of 98·5%, and in contrast to the Metropolis analysis, no posterior probability is estimated as 0. The posterior map positions for both Bayesian methods are similar and agree well with the maximum likelihood map positions after normalization to the total maximum likelihood map length (Table 5). Agreement with the map given by the method-of-moments estimates (Cox *et al.* 1990) is less striking.

DISCUSSION

Radiation hybrid mapping as refined by Cox *et al.* (1990) is potentially one of the most powerful tools in the arsenal of gene mappers. It offers a level of resolution intermediate between that of linkage analysis and *in situ* hybridization on one hand and pulsed-field gel electrophoresis on the other. Radiation hybrid mapping is inherently a statistical technique. Avoiding the uncertainties of fragment generation and retention is impossible since these are at the heart of the technique. The logical response to these uncertainties is to bring modern statistical methods to bear on the problems of locus ordering and distance estimation. The current paper discusses in depth a variety of statistical methods. With the exception of the minimum breaks criterion for ordering, all of these methods depend on a precise model for fragment generation and retention in the individual clones. While the model discussed here and elsewhere (Boehnke *et al.* 1991; Bishop & Crockford, 1992; Boehnke, 1992; Chakravarti & Reefer, 1992; Green, 1992) is undoubtedly false in minor details, it does lead to reasonably robust conclusions and does provide a conceptual framework for examining subtle quantitative issues such as optimal design of radiation dose and posterior probabilities for order.

Minimum obligate breaks provides an attractive criterion for locus ordering (Boehnke *et al.* 1991; Bishop & Crockford 1992; Boehnke, 1992; Weeks *et al.* 1992). Its main virtue is conceptual and computational simplicity. Many of the concrete assumptions necessary for maximum likelihood and Bayesian methods are never explicitly invoked for the minimum obligate breaks criterion. Barrett (1992) has recently shown the minimum obligate breaks criterion to be statistically consistent as the number of fully typed clones tends to infinity. However, implementation of even this simple criterion encounters difficulties when the number of loci reaches about 10. The sheer number of locus orders rules out exhaustively evaluating all orders. Fortunately, short cut techniques such as branch and bound, stepwise locus ordering, and simulated annealing can identify the best orders without actually visiting all orders (Barker *et al.* 1987; Boehnke *et al.* 1991; Weeks *et al.* 1992). The real value of the minimum obligate breaks criterion is the ease with which it can be combined with these search techniques to identify good orders for further analysis. Other nonparametric methods are discussed by Falk (1991) and Weeks *et al.* (1992).

Maximum likelihood and Bayesian methods have the advantage of providing distance estimates as well as criteria for ordering. These methods are naturally more model dependent and more computationally intensive than minimum obligate breaks. Maximum likelihood estimation is familiar to all geneticists involved in linkage analysis. Boehnke *et al.* (1991) show how maximum likelihood can be carried out with reasonable computational speed via an EM algorithm. In the current paper we have stressed the nature of the expected information matrix $\mathbf{J}$. (Explicit expressions for the entries of $\mathbf{J}$ are available in our section on maximum likelihood.) The inverse of $\mathbf{J}$ gives the asymptotic standard errors and correlations of the parameter estimates for a given order. When the fragment retention probability $r = \frac{1}{2}$ or when r is not estimated along with the breakage parameters, then the breakage parameter estimates are asymptotically uncorrelated. There is also maximal information on the breakage parameters when $r = \frac{1}{2}$.

Guerra *et al.* (1992) have emphasized the importance of computing posterior probabilities for locus orders. Posterior probabilities offer the most logically satisfying criterion for determining order. In the present paper we have adopted a different Bayesian prior from that of Guerra *et al.* (1992). We offer two techniques for computing posterior probabilities for locus orders under our prior. The Bayesian Metropolis algorithm is a generic technique for computing expectations with respect to posterior probabilities. Since this algorithm is not well known even in statistical circles, we have developed the necessary theory in detail. At this juncture the algorithm appears feasible and furnishes reasonable posterior probabilities. Computation times on a personal computer are in hours rather than in microseconds for minimum obligate breaks or seconds for maximum likelihood. With minor modifications the Bayesian Metropolis algorithm is applicable to locus ordering by linkage analysis of pedigree data. However, the longer computation times per likelihood in linkage analysis may render the method impractical for ordering large numbers of loci at this time.

Our second algorithm for computing posterior probabilities for locus orders is specific to the radiation hybrid problem. With no missing data, it is even faster than maximum likelihood. However, the presence of untyped loci in the hybrid clones drastically slows the algorithm. For the Cox *et al.* (1990) data, which have about 17% missing observations, computation times range from 8 h to 3 days per order on our 486 computer. The Metropolis algorithm with 10^6

steps takes about 17 h for these data. One possible source of incomplete data is the natural tendency of geneticists to add more loci over time. If some of the original clones are no longer available for typing, then there is no avoiding missing observations. Perhaps the substitution of typing by the polymerase chain reaction for typing by Southern blots will lead to more nearly complete typing. In the presence of nearly complete typing, our second algorithm will probably be the algorithm of choice for comparing candidate locus orders.

Our analysis of optimal radiation dose does suggest a simple rule of thumb. Given the range of retention probabilities $0{\cdot}2 \leqslant r \leqslant 0{\cdot}5$ typical of recent radiation hybrid experiments, it appears that the optimal expected number of breaks per hybrid between two adjacent loci should be about 0·55. This translates into a breakage probability between the loci of about 0·42. We base this rule partly on simulation evidence, and partly on minimizing the average coefficient of variation and the average standard error of the maximum likelihood estimate $\hat{\delta}$ of the distance δ between the loci. (The coefficient of variation is the standard error of $\hat{\delta}$ divided by δ.) At this time the rule should be considered tentative. Better theoretical criteria or more extensive simulations may suggest improved rules. We do wish to emphasize that a broad range of dose levels lead to essentially the same probability of correctly determining locus order.

Goss & Harris (1975) estimate that the expected number of breaks between two loci is proportional to $D^{1{\cdot}6}$, where D is dose. It is fairly straightforward to directly count the number of breaks B per human chromosome by *in situ* hybridization (Cox *et al.* 1990) for a given dose D. Thus, if we assume m loci reside in a chromosome region that is a fraction π of the total chromosome length, then the expected number of breaks between two adjacent loci of the group will be $\pi B/(m+1)$. To achieve 0·55 expected breaks between two adjacent loci, the dose D that results in B breaks per chromosome should be multiplied by $([0{\cdot}55\,(m+1)]/\pi B)^{0{\cdot}625}$.

It is tempting to subject the cells to more than one level of radiation. Our analysis of two-dose optimal designs suggests that this will not be extremely helpful. The average value of the coefficient of variation for the distance estimates is diminished by only about 5% when some cells are irradiated at a low dose and other cells at a high dose. Two-dose designs under the average standard deviation criterion give no improvement over the best single-dose designs. Possibly if irradiation is done in two rounds, with the dose for the second round chosen after analysis of the first round of typing, results would be more favourable. There may also be more definitive criteria for locus ordering that would show two-dose designs in a better light. However, it makes a certain amount of sense to use a single dose for mapping the majority of loci, leaving the order of closely spaced loci to be resolved by pulsed-field gel electrophoresis (Cox *et al.* 1990). Furthermore, there is some evidence that raising radiation dose decreases fragment retention probability (D. R. Cox, personal communication). Perhaps, the longer irradiation periods necessary for higher dose levels somehow interfere with the biological mechanisms for fragment retention. We have not taken these complications into account, but they certainly argue against using extremely high dose levels.

In summary, radiation hybrid mapping is destined to play a major role in improving the speed and resolution of human gene mapping. Forging the best statistical tools for the design and analysis of experiments is just as crucial to radiation hybrid mapping as it is to conventional pedigree methods of linkage analysis. We hope the partial solutions offered here will help our laboratory colleagues as well as stimulate our fellow statisticians to further research.

We thank Hermann Chernoff for his many suggestions on optimal design of radiation doses and David Cox for his help in understanding the biology underlying radiation hybrids and for sharing with us the Richard *et al.* (1991) data. Research supported in part by the University of California, Los Angeles; Harvard University; the University of Michigan; and USPHS Grants CA16042, HG00376, and HG00209.

REFERENCES

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

Barrett, J. H. (1992). Genetic mapping based on radiation hybrid data (submitted).

Bishop, D. T. & Crockford, G. P. (1992). Comparisons of radiation hybrid mapping and linkage mapping. *Genetic Analysis Workshop 7: Issues in Gene Mapping and Detection of Major Genes* (ed. J. W. MacCluer, A. Chakravarti, D. Cox, D. T. Bishop, S. J. Bale and M. H. Skolnick). *Cytogenetics and Cell Genetics*. Basel: Karger.

Boehnke, M. (1992). Radiation hybrid mapping by minimization of the number of obligate chromosome breaks. *Genetic Analysis Workshop 7: Issues in Gene Mapping and Detection of Major Genes* (ed. J. W. MacCluer, A. Chakravarti, D. Cox, D. T. Bishop, S. J. Bale and M. H. Skolnick). *Cytogenetics and Cell Genetics*. Basel: Karger.

Boehnke, M., Lange, K. & Cox, D. R. (1991). Statistical methods for multipoint radiation hybrid mapping. *Amer J. Hum. Genet.* (in press).

Burmeister, M., Kim, S., Price, E. R., de Lange, T., Tantravahi, U., Myers, R. M. & Cox, D. R. (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, J. E. (1992). A theory for radiation hybrid (Goss–Harris) mapping: application to proximal 21q markers. *Genetic Analysis Workshop 7: Issues in Gene Mapping and Detection of Major Genes* (ed. J. W. MacCluer, A. Chakravarti, D. Cox, D. T. Bishop, S. J. Bale and M. H. Skolnick). *Cytogenetics and Cell Genetics*. Basle: Karger.

Chernoff, H. (1979). *Sequential Analysis and Optimal Design*, revised edition. CMBS Series in Applied Mathematics. Philadelphia: SIAM.

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

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc.* B**39**, 1–22.

Falk, C. T. (1991). A simple method for ordering loci using data from radiation hybrids. *Genomics* **9**, 120–123.

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.* **6**, 721–741.

Goss, S. J. & Harris, H. (1975). New method for mapping genes in human chromosomes. *Nature* **255**, 680–684.

Green, P. (1992). Construction and comparison of chromosome 21 radiation hybrid and linkage maps using CRI-MAP. *Genetic Analysis Workshop 7: Issues in Gene Mapping and Detection of Major Genes* (ed. J. W. MacCluer, A. Chakravarti, D. Cox, D. T. Bishop, S. J. Bale and M. H. Skolnick). *Cytogenetics and Cell Genetics*. Basel: Karger.

Guerra, R., McPeek, M. S., Speed, T. P. & Stewart, P. M. (1992). A Bayesian analysis for mapping from radiation hybrid data. *Genetic Analysis Workshop 7: Issues in Gene Mapping and Detection of Major Genes* (ed. J. W. MacCluer, A. Chakravarti, D. Cox, D. T. Bishop, S. J. Bale and M. H. Skolnick). *Cytogenetics and Cell Genetics*. Basel: Karger.

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

Kalos, M. H. & Whitlock, P. A. (1986). *Monte Carlo Methods*, vol. 1. New York: Wiley.

Karlin, S. & Taylor, H. (1975). *A First Course in Stochastic Processes*, 2nd edn. New York: Academic.

Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* **220**, 671–680.

Meng, X.-L. & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *JASA* **86**, 899–909.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092.

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1989). *Numerical Recipes. The Art of Scientific Computing*. Cambridge: Cambridge University Press.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd edn. New York: Wiley.

READ, C. B. (1988). Spacings. In *Encyclopedia of Statistical Sciences*, vol. **8** (ed. S. Kotz and N. L. Johnson), pp. 566–569. New York: Wiley.

REINGOLD, E. M., NIEVERGELT, J. & DEO, N. (1977). *Combinatorial Algorithms: Theory and Practice*. Englewood Cliffs, New Jersey: Prentice-Hall.

RICHARD, C. W., WITHERS, D. A., MEEKER, T. C., MAURER, S., EVANS, G. A., MYERS, R. M. & COX, D. R. (1991). A radiation hybrid map of the proximal long arm of human chromosome 11 containing the multiple endocrine neoplasia type 1 (MEN-1) and bcl-1 disease loci. *Amer. J. Hum. Genet.* (in press).

ROSENBLATT, M. (1971). *Markov Processes: Structure and Asymptotic Behavior*. New York: Springer.

RUBIN, D. B. (1988). Using the SIR algorithm to simulate posterior distributions. *Bayesian Statistics 3* (ed. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith). Oxford: Oxford University Press.

SHORACK, G. R. & WELLNER, J. A. (1986). *Empirical Processes with Applications in Statistics*. New York: Wiley.

TANNER, M. & WONG, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Stat. Assoc.* **82**, 528–550.

WEEKS, D. E. & LANGE, K. (1989). Trials, tribulations, and triumphs of the EM algorithm in pedigree analysis. *IMA J. Math. Appl. Biol. and Med.* **6**, 209–232.

WEEKS, D. E., LEHNER, T. & OTT, J. (1992). Preliminary ranking procedures for multilocus ordering based on radiation hybrid data. *Genetic Analysis Workshop 7: Issues in Gene Mapping and Detection of Major Genes* (ed. J. W. MacCluer, A. Chakravarti, D. Cox, D. T. Bishop, S. J. Bale and M. H. Skolnick). *Cytogenetics and Cell Genetics*. Basel: Karger.

APPENDIX

The text relegated two tasks to the appendix. First, we need to check that the posterior density $p(\gamma \mid \text{obs})$ is invariant under the specific transition scheme outlined for the Bayesian Metropolis algorithm. Recall that the problem is that the proposal density $t(\gamma^* \mid \gamma)$ is not a legitimate density with respect to Lebesgue measure μ on $\Gamma = [0,1]^m$. In fact, unless we choose to invert all loci simultaneously, the proposal density lives on a subspace of $[0,1]^m$. Let $Z \subset \{1, \ldots, m\}$ define the loci we choose to invert and resample. We can project any point $\gamma \in \Gamma$ onto the point γ_Z corresponding to the coordinates of the loci in Z. We can likewise index the proposal density $t_Z(\gamma_Z^* \mid \gamma_Z)$ and transition density $q_Z(\gamma_Z^* \mid \gamma_Z)$ by Z and indicate their dependence on the relevant coordinates. This notation makes it clear that the proposal density is a density with respect to the product measure

$$\mu_Z = \prod_{z \in Z} \mu_z,$$

where the μ_z are copies of 1-dimensional Lebesgue measure. (Our semirigorous arguments will omit technical details of measurability.)

To prove invariance we again integrate the detailed balance relation (9) against a bounded, continuous function $g(\gamma^*)$. However, now instead of integrating with respect to the product measure $\mu(\gamma^*) \times \mu(\gamma)$, we integrate with respect to a more complicated product measure. Let Y be the complement of the set Z. If the projected point γ_Y and the measure μ_Y are defined in the obvious manner, then $\mu(\gamma) = \mu_Z(\gamma_Z) \times \mu_Y(\gamma_Y)$. Also let ω_{γ_Y} be the unit point measure at γ_Y. With this notation, we integrate with respect to

$$\omega_{\gamma_Y}(\gamma_Y^*) \times \mu_Z(\gamma_Z^*) \times \mu_Z(\gamma_Z) \times \mu(\gamma_Y).$$

The first factor of this product measure simply forces $\gamma_Y^* = \gamma_Y$. Integration of $g(\gamma^*)$ against the left hand side of the detailed balance equation (9) gives

$$\int\int\int\int g(\gamma^*)\, q_Z(\gamma_Z^* \mid \gamma_Z)\, d\omega_{\gamma_Y}(\gamma_Y^*)\, d\mu_Z(\gamma_Z^*)\, p(\gamma \mid \text{obs})\, d\mu_Z(\gamma_Z)\, d\mu_Y(\gamma_Y).$$

Let a_Z be the probability that we choose the set Z of loci to invert and resample. Multiplying the above quadruple integral by a_Z and summing on Z produces

$$\sum_Z a_Z \int\int\int g(\gamma^*)\, q_Z(\gamma_Z^* \mid \gamma_Z)\, d\omega_{\gamma_Y}(\gamma_Y^*)\, d\mu_Z(\gamma_Z^*)\, p(\gamma \mid \text{obs})\, d\mu(\gamma). \tag{20}$$

Let
$$c_Z(\gamma) = 1 - \int\int q_Z(\gamma_Z^* \mid \gamma_Z)\, d\omega_{\gamma_Y}(\gamma_Y^*)\, d\mu_Z(\gamma_Z^*)$$

be the probability of rejecting the proposed move from the current point γ conditional on selecting the set Z of loci to invert and resample. Adding

$$\sum_Z a_Z \int g(\gamma)\, c_Z(\gamma)\, p(\gamma \mid \text{obs})\, d\mu(\gamma) \tag{21}$$

to (20) yields the equality $E[g(\mathbf{U}_{k+1})] = E(E[g(\mathbf{U}_{k+1}) \mid \mathbf{U}_k])$ in disguised form, where $\mathbf{U}_k$ is the current state of the Markov chain and $\mathbf{U}_{k+1}$ is the next state.

Now choose a particular Z and integrate $g(\gamma^*)$ against the right hand side of the detailed balance equation (9). Taking into account that

$$\omega_{\gamma_Y}(\gamma_Y^*) \times \mu_Y(\gamma_Y) = \omega_{\gamma_Y^*}(\gamma_Y) \times \mu_Y(\gamma_Y^*),$$

integration against the right hand side gives

$$\int\int\int\int g(\gamma^*)\, q_Z(\gamma_Z \mid \gamma_Z^*)\, d\omega_{\gamma_Z^*}(\gamma_Y)\, d\mu_Z(\gamma_Z)\, p(\gamma^* \mid \text{obs})\, d\mu_Z(\gamma_Z^*)\, d\mu_Y(\gamma_Y^*).$$

Multiplying this second quadruple integral by a_Z and summing on Z in turn yields

$$\sum_Z a_Z \int\int\int g(\gamma^*)\, q_Z(\gamma_Z \mid \gamma_Z^*)\, d\omega_{\mathbf{g}_Y^*}(\gamma_Y)\, d\mu_Z(\gamma_Z)\, p(\gamma_Z^* \mid \text{obs})\, d\mu(\gamma^*).$$

If we add (21) to this, the final result for the right hand side of the detailed balance equation is

$$\int g(\gamma)\, p(\gamma \mid \text{obs})\, d\mu(\gamma).$$

Once again we have demonstrated that $E[g(\mathbf{U}_{k+1})] = E[g(\mathbf{U}_k)]$, and this proves invariance.

Our second task is to check that the Bayesian Metropolis algorithm is ergodic. As discussed in the text, it suffices to prove the mixing condition $\Pr^{(n_\gamma)}(\gamma, A) > 0$ for all sets A with $\mu(A) > 0$, for μ-almost all γ, and for some positive integer n_γ. In general, the conforming points γ will depend on A and the integer n_γ will depend on both γ and A. First we observe that there is some permutation σ of $\{1, \ldots, m\}$ such that $A \cap \{\gamma^* : \gamma^*_{\sigma(1)} < \ldots < \gamma^*_{\sigma(m)}\}$ has positive μ-measure. Without loss of generality, we can assume that σ is the identity permutation and that $A \subset \{\gamma^* : \gamma_1^* < \ldots < \gamma_m^*\}$. Now the likelihood satisfies $L(\text{obs} \mid \gamma) > 0$ provided no two coordinates of γ are equal. Equality of two coordinates implies two loci coincide, and this condition is incompatible with the data if there is at least one obligate break between the two loci. Thus, μ-almost all points proposed from the current point will be acceptable with positive probability. The ambiguous situation of some coordinates being equal is almost never reached, and it is sensible not to start the chain in it.

To verify ergodicity we first argue that, starting at any γ, it is possible to bring the loci into the reverse order $m, \ldots, 1$ in a finite number of steps. (Our reason for choosing the reverse order will be clear in a moment.) In fact, it is obvious that the order of the loci implied by the coordinates of γ can be rearranged to the order $m, \ldots, 1$ by a finite number of pairwise inversions of adjacent loci. This is just a consequence of the fact that any permutation can be written as a composition of pairwise transpositions. Since the probability of choosing to invert a given adjacent pair is positive, and there always is a positive probability of accepting a proposed rearrangement of a pair once it is chosen, there is a positive probability of success for all steps of the required sequence of pairwise inversions.

Once the reverse order $m, \ldots, 1$ has been achieved, one additional step of the chain will land us in the set A with positive probability; we merely need to invert all loci simultaneously. This particular inversion produces the order compatible with A. By the nature of selecting order statistics from $[0, 1]$, the transition density for resampling the m loci is

$$m! \min\left[1, \frac{L(\text{obs} \mid \gamma^*)}{L(\text{obs} \mid \gamma^\#)}\right],$$

where γ^* is the destination point and $\gamma^\#$ is the initial point. Integrating this positive density over A with respect to μ gives a positive probability of landing in A. This proves ergodicity.