Statistical Science
1994, Vol. 9, No. 3, 355–366

Monte Carlo Likelihood in Genetic Mapping

E. A. Thompson

Abstract. Monte Carlo likelihood is becoming increasingly used where exact likelihood analysis is computationally infeasible. One area in which such likelihoods arise is that of genetic mapping, where, increasingly, researchers wish to extract additional information from limited trait data through the use of multiple genetic markers. In the genetic analysis context, Monte Carlo likelihood is most conveniently considered as a latent variable problem. Markov chain Monte Carlo provides a method of obtaining realisations of underlying latent variables simulated under a genetic model, conditional upon observed data. Hence a Monte Carlo estimate of the likelihood surface can be formed. Choice of the latent variables can be as critical as choice of sampler. In the case of very few individuals observed in each pedigree structure, such as occurs in homozygosity mapping and affected relative pair methods of genetic mapping, multilocus segregation indicators are defined and proposed as the latent variables of choice. An example of five Werner's syndrome pedigrees is given; these are a subset of the 21 pedigrees on which homozygosity mapping has recently confirmed the location of the Werner's syndrome gene on chromosome 8. However, multilocus computations on these pedigrees are impractical with standard methods of exact likelihood computation.

Key words and phrases: Importance sampling, latent variable framework, Monte Carlo likelihood, Markov chain Monte Carlo Metropolis–Hastings algorithms, segregation indicators and grandparental gene origins, homozygosity mapping and gene-identity-by-descent, linkage analysis, genetic mapping.

1. INTRODUCTION TO LINKAGE ANALYSIS

Monte Carlo likelihood is becoming increasingly used where exact likelihood analysis is computationally infeasible. One area in which such likelihoods arise is that of genetic mapping, where the location in the genome of genes influencing a given trait is to be inferred. With modern molecular genetics techniques, individuals can be typed for a wide variety of DNA markers of known location in the genome. These DNA markers can be chosen to be highly polymorphic; there are several different alleles (types of genes) that an individual may have. The genes at these DNA marker loci segregate in a Mendelian way (Mendel, 1866); each individual has two genes at the locus, one a copy of a randomly chosen one of the two in his father, and the other a copy of a randomly chosen one in his mother. Segregation of genes from different parents to a child, and from a parent to different children, are independent. These simple 50/50 probabilities underlie all of genetics, but in considering the joint segregation at several genetic loci, or the pattern of single-locus segregations on an extended family, computations can rapidly become very complex, principally because not all the relevant information can be observed.

Genetic loci A and B that index segments of DNA on the same chromosome are "linked"; the segregation of genes at the two loci is not independent. If the maternal gene at locus A in a father segregates to a child, it is more probable that the gene that segregates at the adjacent locus B is also the father's maternal gene. Similarly for the father's paternal gene, and similarly also for genes segregating from the mother. This dependence can be expressed through the "recombination fraction" $r(A, B)$ between the two loci. The probability that genes at loci A and B segregating from one parent to the child have different grandparental origins is $r(A, B)$. In fact, the value of a recombination fraction between two loci depends on

E. A. Thompson is Professor of Statistics and of Biostatistics, Department of Statistics, GN-22, University of Washington, Seattle, Washington 98195.

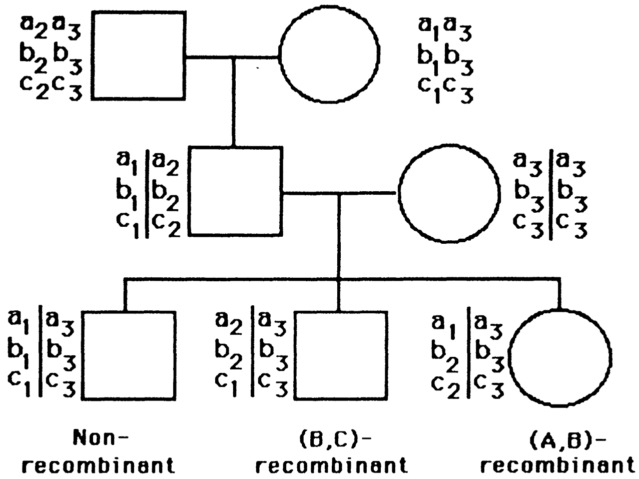

FIG. 1. *Segregation and linkage in a small family. There are three loci, A, B and C as shown in Figure 2a. The grandparents have genotypes at each locus as shown; their multilocus haplotypes are unknown. The multilocus haplotypes of the parents are known, the father's from the data on his parents, the mother's by default since she carries two identical alleles at each of the three loci. Each child receives an $a_3b_3c_3$ chromosome from its mother, regardless of recombination events. There are eight possible chromosomes from the father, the full list (and probabilities) being given in Table* 1.

numerous factors, most importantly on the sex of the parent. This fact can be incorporated into analyses, but, for simplicity, will be ignored in the current review. The biological phenomenon underlying recombination is a "crossover" between the two parental chromosomes in the formation of the offspring chromosome. There will be a recombination between loci A and B if there is an odd number of crossover events.

The genotype of an.individual at a locus is the pair of alleles carried at that locus. For example, consider an individual who is genotype a_1a_2 at locus A and b_1b_2 at locus B. There is no information in this notation as to which allele derives from which parent, nor which alleles are on the same chromosome. In fact, the individual may have two-locus genotype either a_1b_1/a_2b_2 or a_1b_2/a_2b_1, the pairs separated by "/" designating the alleles on the two chromosomes. The alternative arrangements are alternative *phases*; the alleles on a single chromosome (e.g., a_1b_1) constitute a *haplotype*. The term *multilocus genotype* will be used to refer to the unordered pair of haplotypes carried by an individual, that is, it includes a specification of phase. The set of single-locus genotypes (here a_1a_2 at locus A and b_1b_2 at locus B) can correspond to many different multilocus genotypes (here a_1b_1/a_2b_2 or a_1b_2/a_2b_1).

Consider now three genetic loci on a chromosome (Figure 1). A recombination has taken place between two loci, if the genes segregating to the offspring derive from different chromosomes in the parent (i.e., from different grandparents). The genetic (map) distance between two loci is the expected number of recombinations between them and hence is additive (Haldane, 1919). However, the data provide information only on recombination frequencies between loci (Fisher, 1922). This pattern is related to map distance, but also depends on the pattern of interference between the two chromosome segments (A, B) and (B, C). Interference is the name given to the biological phenomenon that a crossover at one point on a chromosome affects the chance that crossovers occur at other points in the vicinity. Under an assumption of no interference, recombination events in the two segments are independent, and the joint segregation probabilities at the three loci is shown in Table 1.. In practice, interference exists, particularly where the loci are close together and recombination fractions between them are small. However, the amounts of data required to estimate levels and patterns of interference seldom exist in human genetic studies. In genetic mapping, the objective is to detect linkage, to infer locus order and to place loci

TABLE 1
Haplotype probabilities for offspring in Figure 1

Paternal	Recombination		Grandparental origin indicators*			
haplotype	In (A, B)	In (B, C)	W_{pA}	W_{pB}	W_{pC}	Probability
$a_1b_1c_1$	no	no	0	0	0	$\frac{1}{2}(1-r_1)(1-r_2)$
$a_1b_1c_2$	no	yes	0	0	1	$\frac{1}{2}(1-r_1)r_2$
$a_1b_2c_1$	yes	yes	0	1	0	$\frac{1}{2}r_1r_2$
$a_1b_2c_2$	yes	no	0	1	1	$\frac{1}{2}r_1(1-r_2)$
$a_2b_1c_1$	yes	no	1	0	0	$\frac{1}{2}r_1(1-r_2)$
$a_2b_1c_2$	yes	yes	1	0	1	$\frac{1}{2}r_1r_2$
$a_2b_2c_1$	no	yes	1	1	0	$\frac{1}{2}(1-r_1)r_2$
$a_2b_2c_2$	no	no	1	1	1	$\frac{1}{2}(1-r_1)(1-r_2)$

*For a definition of segregation indicators **W**, we see Section 4.

on a chromosome by estimating recombination fractions between them. For such purposes, interference can safely be ignored.

Now in mapping a genetic disease, marker types will be available for some individuals in a pedigree in which the disease is segregating. Disease or relevant quantitative trait data will be available also for some members of the pedigree. However, first, not all individuals will be observed; some will be unavailable, particularly ancestors. Second, the genes underlying the trait phenotypes may not be precisely clear; for example, for a recessive disease, two copies of the disease allele are needed to express the trait, but those who do not express it may have one copy of the disease allele or none; that is, there is no 1–1 correspondence between phenotype and genotype. Third, even where single-locus marker genotypes are observable, the haplotype or phase information is not. As commented above, one set of single-locus genotypes can correspond to many different multilocus genotypes. Thus in computing a likelihood, for a given locus order and set of recombination fractions, a huge sum over all the possible configurations of haplotypes is required. With the increasing availability of DNA markers, there is an increasing potential for mapping traits with more limited trait data or more complex modes of expression. For example, traits may be subject to environmental effects, may be age dependent or may result from genetic effects at several loci, which may or may not interact in producing a phenotypic effect and may or may not be linked. However, more markers, and marker loci with more alleles, and traits observable for perhaps a more limited subset of the pedigree members, all compound the computational difficulties, since the number of possible underlying configurations of genes on all the relevant members of the pedigree increases vastly.

Two further terms require definition at this point. The first is the *lod score* used in the detection of linkage between two loci, normally a trait locus and a marker locus. This is the log-likelihood ratio of the likelihood maximised over a recombination frequency, $0 \leq r \leq \frac{1}{2}$, to the likelihood under the "null hypothesis" of independent segregation, $r = \frac{1}{2}$. The second is the *location score*, which is the log-likelihood as a function of position of a presumed trait locus against a fixed map of genetic markers. For mapping a disease trait, the method of location scores is often used (Lathrop et al., 1984). The marker map is taken as known, although in practice there may be considerable uncertainty even as to the map order of marker loci. The disease locus is then mapped against this fixed background; the advantage of this approach is that there is only one varying parameter in the likelihood; the location of the disease locus determines all the recombination fractions. The null hypothesis presumes the same marker map, with the trait locus being unlinked. A *multipoint lod score* may refer to a location score, but more typically refers to a log-likelihood ratio between alternatives in which recombination fractions between all loci, and even locus order, are also allowed to vary (Figure 2b). For historical reasons, lod scores have typically been given using logs to base 10, while location scores have used logs to base *e*. While important in interpreting the applied literature, this difference obviously has no statistical relevance and will be ignored in this paper.

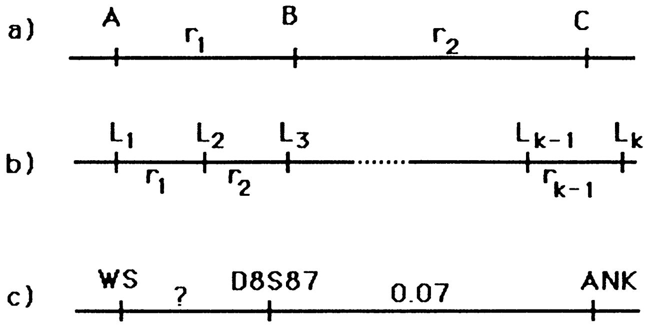

FIG. 2. *Chromosome maps, showing notation for recombination between adjacent markers.* (a) *the general case of three loci, discussed in Section* 1; (b) *the multilocus case;* (c) *the three loci used in the example of Section* 5.

There are many further aspects of linkage analysis and many alternative approaches to localising the genes responsible for a genetic disease. A much fuller description of standard statistical methods in linkage analysis may be found in the text by Ott (1991).

Thus, with the increasing desire to examine multiple markers, and markers with multiple alleles, a major limitation of linkage analysis has become the practical and theoretical bounds on the computational feasibility of likelihood evaluation. Many programs and program packages have been written; the best algorithms for exact likelihood evaluation are based on the method of Elston and Stewart (1971). With this algorithm, computing times increase exponentially with pedigree complexity, numbers of alleles and numbers of loci modelled. One of the most powerful and versatile packages is the LINKAGE program (Lathrop et al., 1984). In some quite standard recent applications, a single run of this program may take several months (Schellenberg et al., 1992a). For an ongoing study, with continuing data collection, this is not acceptable.

Potential solutions include improvement of the programs, improvement of the computational algorithms, or a radically different approach to likelihood assessment. Recently Cottingham, Idury and

Schäffer (1993) have shown that fairly standard computer science programming procedures can improve performance of the LINKAGE program by an order of magnitude. However, improved computers and programs cannot compete with the exponentially increasing demands due to increasing complexity of data available on traits and DNA markers.

Due to the infeasibility or impracticality of exact likelihood computation, approximate methods are often used. Whereas pairwise analyses of a disease gene with a single marker are often computationally feasible, but lacking in power, computation of multipoint location scores can be computationally infeasible. A compromise is provided by *interval mapping* (Lander and Green, 1987). If all markers are fully informative and typed on all individuals, then the likelihood of a given location of the disease locus depends only on the data at the two flanking markers. Alternatively, computations may be carried out using only the data on two (presumed) flanking markers, to simplify computation. Even this three-locus computation can be computationally very intensive, but has greatly increased power over separate pairwise analyses of the disease with each marker.

Curtis and Gurling (1993) have recently proposed an approximation method that extends the interval mapping idea, by accounting for information from more distant markers when closer ones are uninformative. For practical purposes, these methods may be excellent, but without some method of direct evaluation this is impossible to assess. One approach which provides such an assessment is Monte Carlo likelihood, in which exact computation of likelihood ratios is replaced by Monte Carlo estimation.

2. MONTE CARLO IN LINKAGE ANALYSIS

Monte Carlo estimates of integrals or expectations are not new, either in general (Hammersley and Handscomb, 1964) or in genetic linkage analysis. One of the earliest uses of Monte Carlo in linkage analysis was that of Thompson et al. (1978), who proposed use of an *elod*, or expected lod score, in assessing the potential information in a given pedigree structure for purposes of detecting linkage. In that case, data were simulated on small pedigrees, under some proposed linkage and trait parameter values, and the mean lod score (at the same parameter values) evaluated. Elods became quite widely used in the 1980's, but suffered from the defect that what was really required was an elod conditional on observed trait data. Ploughman and Boehnke (1989) and Ott (1989) separately resolved this problem, producing Monte Carlo methods for simulating marker data conditional on trait data, provided that posterior trait genotype probabilities could be computed for all members of the pedigree. For simple but perhaps large pedigrees this is often feasible, and thus it became possible to assess the potential power of a linkage study, knowing the trait data available for study.

The statistical problems involved in fitting genetic linkage models to trait data **Y** on a set of related individuals may be viewed as latent variable or "missing data" problems. Were the underlying multilocus genotypes (pair of haplotypes) of all individuals observable, likelihood computation and parameter estimation would be trivial, but only the trait data (phenotypes) and single-locus marker genotypes of some individuals are observed. We denote by **X** the underlying genotypes, recombination events and/or other unobserved indicators of the patterns of genes segregating in pedigrees. The observed trait and marker data will be denoted by **Y**, and, where necessary, we separate **Y** into its trait **T** and marker **M** components. The vector θ will denote the complete set of parameters underlying a genetic model, while r will denote a recombination fraction.

The likelihood is

$$\begin{aligned} L(\theta) &= P_\theta(\mathbf{Y}) = \sum_{\mathbf{X}} P_\theta(\mathbf{Y},\mathbf{X}) \\ &= \sum_{\mathbf{X}} P_\theta(\mathbf{Y} \mid \mathbf{X}) P_\theta(\mathbf{X}). \end{aligned} \tag{1}$$

Although the summation may be infeasible, we suppose that the latent variables **X** are chosen in such a way that each term of the expression is easily computed. Then equation (1) may also be written

$$P_\theta(\mathbf{Y}) = \sum_{\mathbf{X}} P_\theta(\mathbf{Y} \mid \mathbf{X}) P_\theta(\mathbf{X}) = \mathbb{E}_\theta P_\theta(\mathbf{Y} \mid \mathbf{X}),$$

where the expectation is over **X**-values, with probabilities according to the "prior" $P_\theta(\mathbf{X})$. This expression of the likelihood as an expectation was noted by Ott (1979), who also noted as a personal communication from Lange that

$$P_\theta(\mathbf{Y}) = \mathbb{E}_{\theta_0}\left(P_\theta(\mathbf{Y} \mid \mathbf{X}) \frac{P_\theta(\mathbf{X})}{P_{\theta_0}(\mathbf{X})} \right),$$

the implication of this importance sampling formula being that realisations could be simulated at θ_0, which might be easier or more effective than simulating at θ. The problem with these early Monte Carlo likelihoods, as with the early Monte Carlo elod estimates, is that sampling of genotypes **X** is not conditioned on the data **Y**. Thus, on a large pedigree, the vast majority of realisations **X** provide minute (or even zero) likelihood contributions.

This situation was changed dramatically by the explosion in use of Markov chain Monte Carlo (MCMC) methods, for these provide for simulating from

$$P_\theta(\mathbf{X} \mid \mathbf{Y}) = \frac{P_\theta(\mathbf{X}, \mathbf{Y})}{P_\theta(\mathbf{Y})}$$

(see Section 3). These methods include the Gibbs sampler (Geman and Geman, 1984) and the Metropolis algorithm (Metropolis et al., 1953). Sheehan, Possolo and Thompson (1989) used a Gibbs sampler to sample from the posterior distribution of genotypes **X** given phenotypes **Y** on a pedigree, for a single diallelic Mendelian locus. Lange and Matthyse (1989) used a Metropolis algorithm to estimate the probability distribution of potential lod scores on a pedigree, conditional upon trait data. For proof of the irreducibility of their Markov chain, they required a diallelic trait locus and no marker phenotypes, and thus addressed very similar design assessment questions to those of Ploughman and Boehnke (1989). However, the MCMC approach is potentially far more widely applicable, allowing simulation of underlying genes conditional on a variety of partial phenotypic information.

Monte Carlo estimates of likelihoods soon followed. Thompson and Guo (1991) used the form

$$(2) \qquad \frac{L(\theta)}{L(\theta_0)} = \frac{P_\theta(\mathbf{Y})}{P_{\theta_0}(\mathbf{Y})} = \mathbb{E}_{\theta_0}\left(\frac{P_\theta(\mathbf{X}, \mathbf{Y})}{P_{\theta_0}(\mathbf{X}, \mathbf{Y})} \middle| \mathbf{Y} \right),$$

which will be pursued further in this paper. Lange and Sobel (1991) used a slightly different form, estimating likelihoods, rather than only likelihood ratios. The likelihood for linkage at a given recombination fraction is proportional to

$$(3) \qquad P_\theta(\mathbf{T} \mid \vec{\mathbf{M}}) = \mathbb{E}_\theta(P_\theta(\mathbf{T} \mid \mathbf{X}) \mid \mathbf{M}).$$

In general terms, there are two main differences between equations (2) and (3). The latter provides a likelihood at each particular hypothesis θ, and hence the likelihood ratio relative to an hypothesis of no linkage, for which the likelihood can be evaluated exactly. The former, through simulation at a given θ_0, provides a likelihood ratio approximant, as a function of θ, in the sense of Geyer and Thompson (1992). For values of θ close to θ_0, equation (2) provides a functional form for the local likelihood surface. The question of obtaining a likelihood surface approximant over a wide range of hypotheses will be addressed below.

A different Monte Carlo approach was taken by Kong et al. (1992). This uses the form of the likelihood (2), but generates samples by simulating successively over loci (sequential imputation); see Figure 2b. Sampling of genotypes at locus k is conditional upon the phenotypes at loci $1, \ldots, k$ and previously realised genotypes at loci $1, \ldots, k-1$; the method requires exact computation of probabilities at only a single locus at a time.

3. METROPOLIS–HASTINGS ALGORITHMS AND MONTE CARLO LIKELIHOOD

In Monte Carlo approaches to complex problems with many latent variables, the key is simulation conditional upon data; that is, in the notation of the previous section, from

$$(4) \qquad P_{\theta_0}(\mathbf{X} \mid \mathbf{Y}) = \frac{P_{\theta_0}(\mathbf{X}, \mathbf{Y})}{P_{\theta_0}(\mathbf{Y})}.$$

With well-chosen latent variables **X**, the numerator of this expression is readily evaluated, but the denominator is

$$L(\theta_0) = P_{\theta_0}(\mathbf{Y}) = \sum_{\mathbf{X}} P_{\theta_0}(\mathbf{X}, \mathbf{Y})$$

and this summation is often infeasible. The denominator is, in fact, precisely the likelihood whose exact evaluation is impossible, necessitating the Monte Carlo estimation.

Metropolis–Hastings algorithms are Markov chain Monte Carlo methods designed to meet this need, providing realisations (approximately) from a distribution known up to a normalising constant (Hastings, 1970). For each **X** a "proposal distribution" $q(\cdot, \mathbf{X})$ is defined. Then, if the process is now at **X**, the next value is generated as follows:

1. Generate $\mathbf{X}^*$ from the proposal distribution $q(\cdot, \mathbf{X})$.
2. Compute the Hastings ratio

$$h = \frac{q(\mathbf{X}, \mathbf{X}^*) P_{\theta_0}(\mathbf{X}^* \mid \mathbf{Y})}{q(\mathbf{X}^*, \mathbf{X}) P_{\theta_0}(\mathbf{X} \mid \mathbf{Y})} = \frac{q(\mathbf{X}, \mathbf{X}^*) P_{\theta_0}(\mathbf{Y}, \mathbf{X}^*)}{q(\mathbf{X}^*, \mathbf{X}) P_{\theta_0}(\mathbf{Y}, \mathbf{X})}.$$

   Note that h can be computed without knowledge of $P_{\theta_0}(\mathbf{Y})$.
3. With probability $h^* = \min(1, h)$ the process moves to $\mathbf{X}^*$ and with probability $(1 - h^*)$ it remains at **X**.

The distribution (4) is an equilibrium distribution of the Markov chain just defined. Provided $q(\cdot, \cdot)$ is chosen so that the chain is ergodic, running the chain provides (after a sufficient number of steps for convergence) realisations from the distribution (4).

The algorithm of Metropolis et al. (1953) is a special case; if $q(\mathbf{X}^*, \mathbf{X}) = q(\mathbf{X}, \mathbf{X}^*)$ the Hastings ratio reduces to the odds ratio of the proposal state $\mathbf{X}^*$ versus the current state **X**. The Gibbs sampler (Geman and Geman, 1984) is also a special case, in which the proposal distribution is the conditional distribution for changing one element of **X** conditional on current values of the others, in which case the Hastings ratio reduces to 1 and there is no rejection step. However, the fact of no rejection step is not necessarily advantageous; the Gibbs sampler can make only small changes in $\mathbf{X}$.

In the genetic context, the latent variables $\mathbf{X}$ have normally been taken to be the underlying multilocus genotypes (the pairs of haplotypes) carried by each individual in the pedigree, sometimes with some additional variables. This makes for easy evaluation of $P_{\theta_0}(\mathbf{X}, \mathbf{Y})$ but not for easy sampling of the large space of possible $\mathbf{X}$-values. Single-site updating MCMC algorithms such as the Gibbs sampler are in some ways ideally suited to pedigree analysis; the very local dependence pattern of transmission of genes from parents to offspring makes local conditional distributions easy to sample from. However, such local updating methods can be very slow to sample the space of underlying latent variables $\mathbf{X}$ effectively, and for multiallelic loci the Gibbs sampler need not be irreducible. Conversely, if large changes in $\mathbf{X}$ are proposed, the Hastings ratio can be impossible to compute, and the many constraints in the feasible genotypic patterns on pedigrees mean that almost all proposals have zero probability.

There are various approaches to resolving this problem. For a single multiallelic locus, Sheehan and Thomas (1993) use importance sampling, with 0/1 weights, running a Gibbs sampler for a modified genetic model, for which the Gibbs sampler is irreducible. Sobel and Lange (1993) use multiple Metropolis steps and rejection sampling on the larger latent-variable space of Lange and Matthysse (1989). Lin (1993) made great progress toward increasing the practicality of MCMC methods in multilocus linkage analysis, using Metropolis-coupled samplers (Geyer, 1991a), model modifications in the auxiliary samplers and a form of "heating" in the Metropolis–Hastings steps, to ensure irreducibility (Lin, Thompson and Wijsman, 1993) and to improve mixing of the chain. These strategies result in a sampler that can sample multiallelic genotypes efficiently on a large pedigree. In principle, this method extends to arbitrary numbers of linked markers, but the huge space of possible genotypic configurations that then arises may render the sampler ineffective.

Returning now to equation (2), the likelihood ratio is (Thompson and Guo, 1991)

$$\frac{L(\theta)}{L(\theta_0)} = \frac{P_\theta(\mathbf{Y})}{P_{\theta_0}(\mathbf{Y})} = \mathrm{E}_{\theta_0}\left(\frac{P_\theta(\mathbf{Y},\mathbf{X})}{P_{\theta_0}(\mathbf{Y},\mathbf{X})} \,\middle|\, \mathbf{Y}\right).$$

The MCMC sampler produces dependent realisations $\mathbf{X}(l)$, $l = 1, \ldots, N$ (approximately), from $P_{\theta_0}(\mathbf{X} \mid \mathbf{Y})$ which may be used in a Monte Carlo estimator of the likelihood ratio (2):

$$(5) \qquad \frac{1}{N}\sum_{l=1}^{N}\left(\frac{P_\theta(\mathbf{Y},\mathbf{X}(l))}{P_{\theta_0}(\mathbf{Y},\mathbf{X}(l))}\right).$$

In genetic mapping examples, often there will be data on many independent pedigrees, and it is inefficient not to use this known independence in estimating the likelihood ratio. Let $(\mathbf{Y}^{(i)}, \mathbf{X}^{(i)})$ denote the data variables $\mathbf{Y}^{(i)}$ and latent variables $\mathbf{X}^{(i)}$ on pedigree i, $i = 1, \ldots, k$. Then

$$(6) \qquad \begin{aligned}\frac{L(\theta)}{L(\theta_0)} &= \prod_{i=1}^{k}\frac{L_i(\theta)}{L_i(\theta_0)} = \prod_{i=1}^{k}\frac{P_\theta(\mathbf{Y}^{(i)})}{P_{\theta_0}(\mathbf{Y}^{(i)})}\\ &= \prod_{i=1}^{k}\mathrm{E}_{\theta_0}\left(\frac{P_\theta(\mathbf{Y}^{(i)},\mathbf{X}^{(i)})}{P_{\theta_0}(\mathbf{Y}^{(i)},\mathbf{X}^{(i)})} \,\middle|\, \mathbf{Y}^{(i)}\right).\end{aligned}$$

Thus estimation is most efficiently done for each pedigree, the estimate of $L_i(\theta)/L_i(\theta_0)$ being

$$(7) \qquad \frac{1}{N_i}\sum_{l=1}^{N_i}\left(\frac{P_\theta(\mathbf{Y}^{(i)},\mathbf{X}^{(i)}(l))}{P_{\theta_0}(\mathbf{Y}^{(i)},\mathbf{X}^{(i)}(l))}\right),$$

where N_i is the number of Markov chain Monte Carlo realisations on pedigree i. This separate estimation has practical advantages beyond computational efficiency. As first noted by Fisher (1936), heterogeneity in the likelihoods provided by different pedigrees may indicate genetic heterogeneity, although variation should not be overinterpreted. As discussed in the example of Section 5, there may also be other aspects of the probabilities $P_{\theta_0}(\mathbf{X} \mid \mathbf{Y})$ that are of interest on a pedigree-by-pedigree basis.

These estimators (5) and (7) work well if θ is close to θ_0, but, in comparing alternative genetic models, it is rarely only the local characteristics of the likelihood surface that are of interest. To overcome this problem, chains may be run at a set of parameter combinations, $\theta_0, \theta_1, \theta_2, \ldots, \theta_K$, spanning the range between the two hypotheses of interest, θ_0 and θ_K. An importance sampling approach allows one to use all the samples in a combined estimate of log-likelihood differences along the chain $\theta_0, \theta_1, \theta_2, \ldots, \theta_K$ (Geyer, 1991b). An intuitive way to view this approach is as follows. Assume N_j realisations are taken from chain $P_{\theta_j}(\cdot \mid \mathbf{Y})$. Rather than retaining the chain parameter value corresponding to each realisation, the collection of realisations are "pooled", and the pooled sample is regarded as a sample size $\Sigma_j N_j$ from the weighted average of the distributions indexed by $\theta_0, \ldots, \theta_K$, that is,

$$\frac{1}{\sum_j N_j}\sum_{j=0}^{K}\frac{N_j P_{\theta_j}(\mathbf{Y},\mathbf{X})}{L(\theta_j)}.$$

Then the unknown $L(\theta_j)$ are estimated by considering the version of (5) when sampling is from this "mixture distribution." This reduces to solving the equations

$$L(\theta_j) = \sum_{\mathbf{X}^*} \left(\frac{P_{\theta_j}(\mathbf{Y}, \mathbf{X}^*)}{\sum_{l=0}^{K} N_l P_{\theta_l}(\mathbf{Y}, \mathbf{X}^*)/L(\theta_l)} \right) \quad \text{for } j = 0, \ldots, K, \tag{8}$$

where the summation is over the total combined sample of realisations $\mathbf{X}^*$. If this procedure is implemented, every realisation contributes to the estimate of $L(\theta_j)$ for all j in accordance with the appropriate importance sampling weights (Geyer, 1991b).

Good choice of K and of θ_i, $i = 1, \ldots, K$, is unfortunately often a matter of trial and error, although posterior assessment is straightforward. For the method to have worked, the realisations X^* obtained under each θ_j must have nonnegligible probability under at least one other θ_l. The matrix of mean posterior probabilities that realisations under each model θ_j derive from each other model θ_l is a useful simple diagnostic. More technically, near-degeneracy of the Hessian required in the evaluation of the estimated covariance matrix of the Monte Carlo likelihood estimates (Geyer, 1991b) indicates an inadequate choice of the set of models.

Another question is choice of the latent variables $\mathbf{X}$. The most straightforward implementation of the Monte Carlo likelihood in linkage analysis is to use the multilocus genotypes of individuals as the latent variables $\mathbf{X}$. Alternatively, one could expand the space further and use both the genotypes and also indicators of the grandparental origins of genes (Lange and Matthysse, 1989). However, this results in a huge space of latent variables, which is difficult to sample effectively using MCMC methods.

Some methods of evading this problem have been mentioned above, but an alternative approach is to limit the space of latent variables $\mathbf{X}$. Note that the requirements on $\mathbf{X}$ are only that $P_\theta(\mathbf{Y}, \mathbf{X})$ should be very quickly computable. Now $P_\theta(\mathbf{Y}, \mathbf{X})$ is normally computed as $P_\theta(\mathbf{Y} \mid \mathbf{X}) P_\theta(\mathbf{X})$. Thus any $\mathbf{X}$ for which these two factors can be readily computed will suffice. In some cases, exact integration of some of the latent variables is possible; Thompson (1994b) discusses the example of a mixed model for a quantitative genetic trait in this context. In the context of linkage analysis, the approach of Kong et al. (1992) is also directed toward partial exact computation that will reduce Monte Carlo variance.

In this paper, we propose an alternative reduction of the space of multilocus genotypes of individuals, that has proven effective in linkage analysis examples where only a few individuals are observed on each pedigree.

4. HOMOZYGOSITY MAPPING; ALTERNATIVE LATENT VARIABLES

Due to uncertainties as to whether an unaffected individual carries a disease gene, the computational difficulties of linkage analysis on extended pedigrees and the costs of typing large numbers of individuals, there have been many approaches toward basing linkage analyses on a small number of observed (usually affected) individuals.

It was first pointed out by Smith (1953) that individuals affected with rare recessive diseases provide information for linkage analysis, even without any marker or phenotype data on other relatives. For a recessive disease, affected individuals are *homozygous* at the disease locus; that is, they carry two copies of the same allele. For a rare disease, many affected individuals are so through being the offspring of consanguineous marriages, thus receiving two copies of the disease gene identical-by-descent (IBD) from a recent common ancestor of the two parents. In this case, the affected individual is likely to be homozygous also at closely linked markers, and this homozygosity provides evidence for linkage. The evidence lies in an excess of homozygosity around the disease locus, over that expected. Inbred individuals have higher than average levels of homozygosity due to gene-identity-by-descent (GIBD), but unrelated inbred individuals will be homozygous at independent segments of the genome. The shared affected status of the individuals will cause shared homozygosity in the neighbourhood of the disease locus. The scope of "homozygosity mapping," which is simply linkage analysis using data only on unrelated inbred affected individuals, was extended by Lander and Botstein (1987). With a dense map of highly polymorphic DNA markers, a small number of affected individuals can provide substantial information for mapping a recessive disease gene.

A second rather similar idea underlies methods of linkage analysis based on affected pairs (or small sets) of relatives (Weeks and Lange, 1988; Bishop and Williamson, 1990). It has long been recognised that the majority of information for mapping a disease gene lies in the marker types of affected individuals. For some diseases, moreover, onset may be delayed or uncertain, so that it is not known whether unaffected individuals carry the disease gene. By restricting attention only to individuals known to have the disease, more efficient designs and robust analyses of linkage can be performed. Both in the case of homozygosity mapping (in pedigrees where there may be several affected individuals) and in affected relative set methods, phenotype and marker data are available on only a few individuals. However, there may be numerous other individuals defining the pedigree relationships. Computation of a multilocus linkage likelihood on the pedigree may be impossible, due to the large amount of missing data on a potentially complex pedigree. Thus we now propose an alternative definition of latent variables $\mathbf{X}$ (Section 2) that will permit Monte Carlo estimation of the likelihoods.

Linkage analysis is the analysis of cosegregation of genes at different loci, from parents to offspring. If two loci are tightly linked, there is a high probability that if the individual receives a grandmaternal [grandpaternal] allele from his mother at one locus, he will do so also at the adjacent one, and similarly for the gene received from his father. Let $W_{ml} = 0$ if the maternal allele received by the child at locus l is of grandmaternal origin, and $W_{ml} = 1$ otherwise, and let W_{pl} be similarly defined for the paternal allele. Then, at any locus l,

$$\begin{aligned} P(W_{ml} = 0) = P(W_{ml} = 1) &= P(W_{pl} = 0) \\ &= P(W_{pl} = 1) = \tfrac{1}{2}, \end{aligned} \tag{9}$$

and at two adjacent loci l_1 and l_2

$$\begin{aligned} P(W_{ml_1} = W_{ml_2}) &= P(W_{pl_1} = W_{pl_2}) \\ &= \big(1 - r(l_1, l_2)\big), \end{aligned} \tag{10}$$

where $r = r(l_1, l_2)$, $0 \le r \le \frac{1}{2}$, is the recombination fraction between the two loci.

Then for a given segregation i, the recombination events are determined by segregation indicators W_{ij}, $j = 1, \dots, L$, where W_{ij} is 0 or 1 as the origin of the segregating gene at locus j is grandmaternal or grandpaternal, respectively. That is, we shall take the indicators $\mathbf{W} = \{W_{ij}\}$ as the latent variables $\mathbf{X}$ in the Monte Carlo likelihood framework of Sections 2 and 3.

The prior probabilities of $\mathbf{W}$ are straightforward. However, for implementation of a Metropolis algorithm, recall that relative values of $P_{\theta_0}(\mathbf{W}, \mathbf{Y})$ are required, or $P_{\theta_0}(\mathbf{Y} \mid \mathbf{W})$. The binary indicators $\mathbf{W} = \{W_{ij}\}$ of grandparental origins of genes in each given offspring individual, at each locus readily determine the multilocus gene-identity-by-descent (GIBD) patterns in the observed individuals. This is done simply by following the descent paths of genes from the founders to the observed individuals. An efficient algorithm is easily implemented to update these descent paths, and hence the resulting GIBD pattern in observed individuals, when a W_{ij} is flipped. For the simplest case of homozygosity mapping of one individual in each pedigree, the GIBD pattern is L binary indicators, specifying whether or not the W_{ij} result in the individual having two genes IBD at locus j, $j = 1, \dots, L$. The probability of a genotype homozygous for an allele with frequency q is q^2 or q, as the individual is not or is IBD at the locus. The probability of a heterozygous genotype is 0 if the individual is IBD, and is $2q_1q_2$ otherwise, where q_1 and q_2 are the two allele frequencies. In general, the number of possible GIBD patterns in a set of observed individuals can be large, but it is only the pattern at each locus separately that affects the phenotype probabilities. The conditional probability of phenotypic observations for several individuals is also more complex than for a single observed individual, but can be easily determined provided only a very few individuals are observed on each pedigree (Thompson, 1974).

TABLE 2
Probability ratios of segregation indicators W_{ij}: $P(W_{ij} = 1 \mid \mathbf{W}_{-(ij)}) = P(W_{ij} = 1 \mid W_{i,j-1}, W_{i,j+1})$; $P(W_{ij} = 0 \mid \mathbf{W}_{-(ij)}) = P(W_{ij} = 0 \mid W_{i,j-1}, W_{i,j+1})$; $\mathbf{W}_{-(ij)}$ denotes all elements of $\mathbf{W}$ other than W_{ij}

$W_{i,j-1}$	$W_{i,j+1}$	$P(W_{ij} = 1 \mid \mathbf{W}_{-(ij)})/P(W_{ij} = 0 \mid \mathbf{W}_{-(ij)})$
1	1	$(1 - r_{j-1})(1 - r_j)/r_{j-1}r_j$
1	0	$(1 - r_{j-1})r_j/r_{j-1}(1 - r_j)$
0	1	$r_{j-1}(1 - r_j)/(1 - r_{j-1})r_j$
0	0	$r_{j-1}r_j/(1 - r_{j-1})(1 - r_j)$

The space of $\mathbf{W}$-values is also easy to sample from. The simplest algorithm uses a Metropolis proposals to change the grandparental origin of the gene at a random locus in a random segregation. The probability ratio for the proposed change in $\mathbf{W}$ depends only on the indicators at adjacent loci for the same segregation (Table 2), this then being weighted by the appropriate conditional probability of phenotypic observations $P_{\theta_0}(\mathbf{Y} \mid \mathbf{W})$. This sampler is clearly irreducible: if a given pattern of GIBD in the observed individuals is compatible with the data, then so also is any pattern with fewer genes constrained to be IBD and hence of the same allelic type.

In some situations, it may be possible to reduce the space of latent variables still further. For example, for homozygosity mapping for single affected inbred individuals, the latent variables could be taken simply as GIBD indicators in the affected individual. This then provides a very small space of $\mathbf{X}$-values, and easy computation of $P_\theta(\mathbf{Y} \mid \mathbf{X})$. Computation of $P_\theta(\mathbf{X})$ is harder, but could be achieved, for example, using the recursive algorithms of Thompson (1988). (In fact, this recursive algorithm was used to provide exact GIBD pattern probabilities to provide a deterministic check of some of the Monte Carlo results of the following section.) However, this extreme reduction of the space of $\mathbf{X}$-values decreases the applicability of the framework. Using $\mathbf{W}$, the indicators of grandparental origins in all segregations, as the latent variables provides a framework that can be extended to cases of several affected individuals in an inbred pedigree or, more generally, to small sets of observed relatives in an arbitrary pedigree.

5. HOMOZYGOSITY MAPPING FOR WERNER'S SYNDROME

Werner's syndrome is a very rare recessive genetic disease of premature aging. It has recently been mapped to chromosome 8 using outbred affected relatives (Goto et al., 1992), and this linkage has been confirmed by analysis of a set of inbred affected individuals (Schellenberg et al., 1992b) in 21 small pedigrees of Japanese and Caucasian origin. We exemplify the above methods with an analysis of a subset of the latter data; a set of five pedigrees each with one affected individual. Three of the affected individuals are the offspring of first-cousin marriages, one is the result of more complex inbreeding (Figure 3), and the last is the offspring of a marriage between first cousins once removed. The inbreeding coefficients (the prior probability of IBD genes at each locus) of the five affected individuals are 0.0625, 0.0625, 0.0625, 0.10938 and 0.03125. The five pedigrees chosen for analysis are of Caucasian origin, and the frequency of the disease allele is assumed to be 0.004. Although for any given affected individual the posterior probability of gene-identity-by-descent is high [approximately $f/(f + 0.004)$ for an affected individual with inbreeding coefficient f], the probability that all of the 35 pedigrees now sampled contain affected individuals due to gene-identity-by-descent is small. Thus there are affected individuals who are heterozygous at markers across the region in which the gene is inferred to be.

Two markers were of significance in the published linkage reports. These are D8S87 and ANK: the recombination fraction between them is about 0.07. At D8S87 there are five alleles, with frequencies ranging from 0.1 to 0.4 in Caucasians. At ANK there are two high-frequency alleles (0.44 and 0.50) and three rarer marker alleles. Among the five pedigrees chosen for analysis there is one affected individual heterozygous at D8S87 and one heterozygous at ANK. Originally ANK and D8S87 were thought to be flanking markers, but the likely order is now thought to be (WS, D8S87, ANK) with recombination between WS and D8S87 somewhere between 0.01 and 0.05. Data on and information on these markers were provided by Dr. Ellen Wijsman (1993).

Markov chain Monte Carlo runs to estimate likelihood ratios between alternative recombination values were set up as described in the previous sections. Segregation indicators W_{ij} were assigned for each locus, for each segregation in each pedigree. A random W_{ij} was selected for proposed change from 0 to 1, or 1 to 0. The consequences of the proposed change were determined in terms of the founder genes present in the affected individual. Note that not all founder genes must be traced; for the pedigree of Figure 3, at any given locus, only the six genes of the three founders shown can provide two identical-by-descent copies to the observed descendant. The probability of the data $\mathbf{Y}$ under the proposed change in $\mathbf{W}$ is then also easily determined, and the Metropolis proposal is accepted or rejected. At each step, (whether the change was accepted or not) complete-data likelihood ratios (6) were accumulated, for the pedigree in question, at a preselected set of alternative θ-values. Sampling was performed on the entire set of segregation indicators on all pedigrees; larger pedigrees thus receive an appropriately larger proportion of the Metropolis proposals.

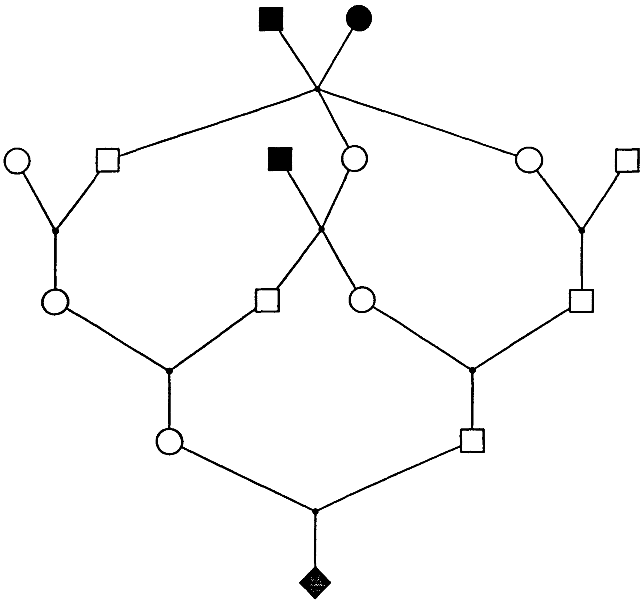

FIG. 3. *Complex pedigree for Werner's syndrome analysis: the final individual in the pedigree is affected; marker data are available only on this individual. The three founder individuals marked can contribute genes to both parents of the affected individual; the individual can receive two homologous genes that are identical-by-descent from any of these three founders.*

Table 3 shows the five recombination values used for sampling and likelihood ratio evaluation in seven runs each of 100,000,000 Metropolis steps. Average Metropolis acceptance probabilities ranged from 0.07 (for θ_1) to 0.82 (for θ_5). The log-likelihoods (summed over the five pedigrees) estimated from each of the runs are shown in Figure 4 (broken lines). Each of these seven runs took only 2.5 hours on a DEC3100 workstation (or 15 minutes on a DEC Alpha 300). Additionally, two longer runs were made at the parameters θ_0 of Table 3; these are the values currently

TABLE 3
Models used for Monte Carlo runs and likelihood estimation

	$r_{\mathrm{WS,\,S87}}$	$r_{\mathrm{S87,\,ANK}}$	Run length	Runs
θ_1	0.01	0.1	10^8	2
θ_2	0.05	0.1	10^8	1
θ_3	0.05	0.2	10^8	1
θ_4	0.1	0.2		0
θ_5	0.5	0.5	10^8	3
θ_0	0.04	0.07	10^9	2

believed. Likelihood ratios at $\theta_1, \ldots, \theta_5$ relative to θ_0 were estimated and are shown as dotted lines in Figure 4. These runs, each of 1,000,000,000 Metropolis steps each took about 23.5 hours on a DEC3100 workstation. Additionally, Figure 4 shows a combined estimate (solid line) from the seven shorter runs. This estimate is produced by a method analogous to the mixture formula of Geyer (1991b) treating the surface from each run as a single realisation. The combined estimate of $\nu_j = \log_e L(\theta_j)$, $j = 1, \ldots, 5$, is given as the solution of

$$\exp(-\nu_j) = \sum_s \left(\frac{L(\theta_j; s)}{\sum_t L(\theta_{l(t)}; s) \exp(\nu_{l(t)})} \right) \quad \text{for } j = 0, \ldots, 5, \tag{11}$$

where s and t index the separate runs, $L(\theta_j; s)$ is the (relative) likelihood estimate at θ_j provided by run s and $\theta_{l(t)}$ is the simulation parameter value used in run t.

Figure 4 shows that, although there is variability in the log-likelihood, particularly relative to the more distant θ_5, even these five pedigrees provide evidence for linkage, despite the fact that two of the affected individuals are heterozygous at one of the two markers. The log-likelihood for linkage [(0.04, 0.07) or (0.05, 0.1) relative to (0.5,0.5)] is about 9.0; using logs to base 10, this would be a "lod score" of about 3.9. Generally, the likelihood estimate tends to be biased upward at the simulation value used for each run, but in these long runs this effect is not large.

Each run was started at a configuration **W** in which none of the affected individuals has two genes identical by descent at any of the loci. This is convenient in that the configuration is then automatically compatible with the phenotypic data; it will tend to bias the results against linkage if the runs are too short. As with all Markov chain Monte Carlo, it is difficult to know whether the runs are long enough, but the symmetries of the example do provide some diagnostics. The gene descent configuration of the sampled individual can be scored by the founder genes present, not simply by whether the genes at a given locus are identical by descent. By symmetry the states corresponding to the four genes of the founder couple (Figure 3) should occur equally frequently. With tight linkage, very long runs may be required for all possible combinations of founder genes at the different loci to be realised in the affected offspring individual; some have very small probability. However, those having substantial probabilities, which are known by symmetry to be equal, are realised with approximately equal frequency even in much shorter runs.

Expectations other than likelihood ratios can, of course, be estimated from the same Monte Carlo

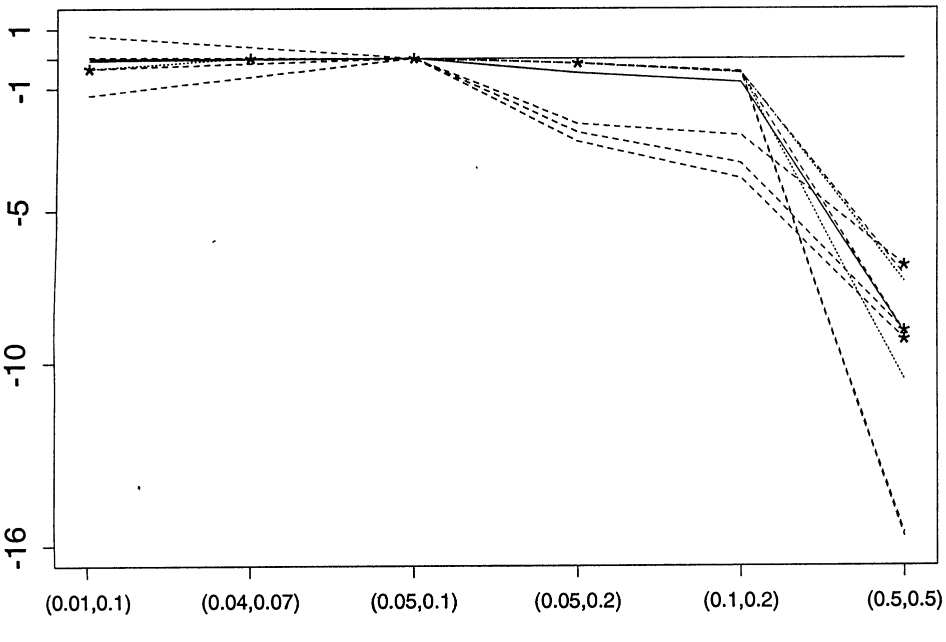

FIG. 4. *Log-likelihood for linkage parameters estimated by Markov chain Monte Carlo; for further details of the curves, see text.*

TABLE 4
Posterior gene identity by descent probabilities; based on run of 1,000,000,000 *total Metropolis realisations; run at recombination fractions* (0.04, 0.07)

Pedigree	Metropolis steps	Locus 1 WS	Locus 2 D8S87	Locus 3 ANK
1	156,256,383	0.793300	0.000000	0.126347
2	156,220,489	0.990610	0.960410	0.783383
3	156,254,166	0.937433	0.725601	0.000000
4	312,515,364	0.984823	0.909241	0.707956
5	218,753,598	0.956829	0.911231	0.662862

runs. In the context of homozygosity mapping, one expectation of interest is the posterior probability of gene identity by descent of the affected individuals at each of the loci. These can be estimated very simply by scoring each Metropolis step. Table 4 shows the estimates resulting from one of the two runs of 1,000,000,000 steps. Note that the first individual is heterozygous at D8S87 and the third at ANK, and that pedigrees 4 and 5 have more segregations and hence accumulate more Metropolis steps. Tight linkage to a locus at which the individual is heterozygous decreases the probability of gene identity by descent, and homozygosity for a very rare allele (e.g., the WS disease allele) gives much higher posterior probability of gene identity than does homozygosity at a common marker allele such as those at ANK.

For comparison with exact results, an additional set of runs were made, fixing $r_{\mathrm{S87,ANK}}$ to be 0.07 and simulating at $r_{\mathrm{WS,S87}} = 0.1$, and computing likelihood ratios relative to

$$r_{\mathrm{WS,S87}} = 0.01, 0.04, 0.05, 0.2, 0.5.$$

The log-likelihood results from two runs, each of 1,000,000,000 Metropolis steps agreed to at least three significant figures. Except for log-likelihoods relative to $r_{\mathrm{WS,S87}} = 0.5$, where there were small differences, the results also agreed with exact computations obtained both from the LINKAGE package (Wijsman, 1993) and using the algorithm of Thompson (1988). (Earlier findings, that results for the complex pedigree of Figure 3 were less accurate, were found to be due to an inadequately documented limitation in using the *makeped* preprocessing program of the LINKAGE package on pedigrees with multiple loops.) Moreover, these were results from single simulation values $r_{\mathrm{WS,S87}} = 0.1$, not close to $r_{\mathrm{WS,S87}} = 0.5$. With more intermediate values, or with better mixing samplers (e.g., Geyer and Thompson, 1994), this difference could be easily eliminated. It is also worth noting that a short check run of only 1,000,000 steps (taking 80 seconds) provided qualitatively similar results, correct to one significant figure in most cases; again it was only for log-likelihoods relative to $r_{\mathrm{WS,S87}} = 0.5$ where there were bigger discrepancies. Also, the short run gave almost exactly the same Metropolis acceptance rates for the proposals on each pedigree (to within 0.001) as for the long runs, and very similar (to within 0.01) posterior probabilities of GIBD at each locus for each pedigree. Reliable qualitative results can be obtained quite quickly.

6. DISCUSSION

Genetic linkage mapping using highly polymorphic DNA markers is becoming increasingly used to localise the genes responsible for a wide variety of human genetic disease. As the human marker map becomes clearer and more detailed, the challenge to use it to locate disease genes increases. The computational problems involved in multipoint linkage likelihoods are immense, particularly when there is much missing data on the pedigree. Many statistical problems remain in the area of human genetic mapping.

Monte Carlo likelihood provides an approach when exact likelihood computation in infeasible, particularly in problems of complex dependent highly structured data, such as arise in genetic analysis. There are many ways to set up the Markov chain Monte Carlo that provides estimates of likelihood ratios. In this paper we have focussed on one particular formulation that seems to have promise in cases where a very few individuals are observed on each of a number of possibly large pedigrees, the individuals possibly being observed for a number of DNA markers. However, no one framework will provide a universal solution to the Monte Carlo estimation of likelihoods arising in the genetic mapping and analysis of complex traits. The example of the rare recessive Werner's syndrome pedigrees is intended as an illustration of what is possible in the area of Monte Carlo likelihood, not as the unique solution.

ACKNOWLEDGMENTS

I am grateful to Charles Geyer for discussions on Monte Carlo likelihood functions, and to Ellen Wijsman for discussions of homozygosity mapping and details of the Werner's syndrome study. This research was supported in part by NIH Grant GM-046255.

REFERENCES

BISHOP, D. T. and WILLIAMSON, J. (1990). The power of identity-by-state methods for linkage analysis. *American Journal of Human Genetics* **46** 254–265.

COTTINGHAM, R. W., IDURY, R. M. and SCHÄFFER, A. A. (1993). Faster sequential genetic linkage computations. *American Journal of Human Genetics* **53** 252–263.

Curtis, D. and Gurling, H. (1993). A procedure for combining two-point lod scores into a summary multipoint map. *Human Heredity* **43** 173–185.

Elston, R. C. and Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity* **21** 523–542.

Fisher, R. A. (1922). The systematic location of genes by means of crossover observations. *American Naturalist* **56** 406–411.

Fisher, R. A. (1936). Heterogeneity of linkage data for Friedreich's ataxia and the spontaneous antigens. *Annals of Eugenics* **7** 17–21.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** 721–741.

Geyer, C. J. (1991a). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (E. Kerimidas, ed.) 156–163. Interface Foundation of North America, Fairfax Station, VA.

Geyer, C. J. (1991b). Reweighting Monte Carlo mixtures. Technical Report 568, School of Statistics, Univ. Minnesota.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Statist. Soc. Ser. B* **54** 657–699.

Geyer, C. J. and Thompson, E. A. (1994). Annealing Markov chain Monte Carlo with applications to pedigree analysis. Unpublished manuscript.

Goto, M., Rubenstein, M., Weber, J., Woods, K. and Drayna, D. (1992). Genetic linkage of Werner's syndrome to five markers on chromosome 8. *Nature* **355** 735–738.

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

Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo Methods.* Methuen, London.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.

Kong, K., Irwin, M., Cox, N. and Frigge, M. (1992). Multiloci problems and the method of sequential imputation. Technical Report 351, Dept. Statistics, Univ. Chicago.

Lander, E. S. and Botstein, D. (1987). Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* **236** 1567–1570.

Lander, E. S. and Green, P. (1987). Construction of multilocus linkage maps in humans. *Proc. Nat. Acad. Sci. U.S.A.* **84** 2363–2367.

Lange, K. and Matthysse, S. (1989). Simulation of pedigree genotypes by random walks. *American Journal of Human Genetics* **45** 959–970.

Lange, K. and Sobel, E. (1991). A random walk method for computing genetic location scores. *American Journal of Human Genetics* **49** 1320–1334.

Lathrop, G. M., Lalouel, J.-M., Julier, C. and Ott, J. (1984). Strategies for multilocus linkage analysis in humans. *Proc. Nat. Acad. Sci. U.S.A.* **81** 3443–3446.

Lin, S. (1993). Markov chain Monte Carlo estimates of probabilities on complex structures. Ph.D. dissertation, Univ. Washington.

Lin, S., Thompson, E. A. and Wijsman, E. M. (1993). Achieving irreducibility of the Markov chain Monte Carlo method applied to pedigree data. *IMA J. Math. Appl. Med. Biol.* **10** 1–17.

Mendel, G. (1866). *Experiments in Plant Hybridisation.* (Mendel's original paper in English translation, with a commentary by R. A. Fisher. Oliver and Boyd, Edinburgh, 1965.)

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

Ott, J. (1979). Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigrees. *American Journal of Human Genetics* **31** 161–175.

Ott, J. (1989). Computer simulation methods in linkage analysis. *Proc. Nat. Acad. Sci. U.S.A.* **86** 4175–4178.

Ott, J. (1991). *Analysis of Human Genetic Linkage*, 2nd ed. Johns Hopkins Univ. Press.

Ploughman, L. M. and Boehnke, M. (1989). Estimating the power of a proposed linkage study for a complex genetic trait. *American Journal of Human Genetics* **44** 543–551.

Schellenberg, G. D., Bird, T. D., Wijsman, E. M., Orr, H. T., Anderson, L., Nemens, E., White, J. A., Bonneycastle, L., Weber, J. L., Alonso, M. E., Potter, H., Heston, L. L. and Martin, G. M. (1992a). Genetic linkage evidence for a familial Alzheimer's disease locus on chromosome 14. *Science* **258** 668–671.

Schellenberg, G. D., Martin, G. M., Wijsman, E. M., Nakura, J., Miki, T. and Ogihara, T. (1992b). Homozygosity mapping and Werner's syndrome. *The Lancet* **339** 1002.

Sheehan, N. A., Possolo, A. and Thompson, E. A. (1989). Image processing procedures applied to the estimation of genotypes on pedigrees. *American Journal of Human Genetics* **45** (Suppl.) A248.

Sheehan, N. A. and Thomas, A. W. (1993). On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. *Biometrics* **49** 163–175.

Smith, C. A. B. (1953). Detection of linkage in human genetics. *J. Roy. Statist. Soc. Ser. B* **15** 153–192.

Sobel, E. and Lange, K. (1993). Metropolis sampling in pedigree analysis. Unpublished manuscript.

Thompson, E. A. (1974). Gene identities and multiple relationships. *Biometrics* **30** 667–680.

Thompson, E. A. (1988). Two-locus and three-locus gene identity by descent in pedigrees. *IMA J. Math. Appl. Med. Biol.* **5** 261–280.

Thompson, E. A. (1994a). Monte Carlo likelihood in the genetic analysis of complex traits. *Phil. Trans. Roy. Soc. London Ser. B* **344** 345–351.

Thompson, E. A. (1994b). Monte Carlo likelihood in genetic analysis. In *Probability, Statistics, Optimization: A tribute to Peter Whittle* (F. P. Kelly, ed.) 281–293. Wiley, New York.

Thompson, E. A. and Guo, S. W. (1991). Evaluation of likelihood ratios for complex genetic models. *IMA J. Math. Appl. Med. Biol.* **8** 149–169.

Thompson, E. A., Kravitz, K., Hill, J. and Skolnick, M. H. (1978). Linkage and the power of a pedigree structure. In *Genetic Epidemiology* (N. E. Morton, ed.) 247–253. Academic, New York.

Weeks, D. E. and Lange, K. (1988). The affected-pedigree-member-methods of linkage analysis. *American Journal of Human Genetics* **42** 315–326.

Wijsman, E. (1993). Personal communication.