Monte Carlo Estimation of Multilocus Autozygosity Probabilities.

Elizabeth Thompson*

Department of Statistics, GN-22,
University of Washington
Seattle, WA 98195

Abstract

The ability to sample latent variables using Markov chain Monte Carlo (MCMC) has had a major impact on computations relating to the genetic analysis of complex traits, or traits observed on complex pedigrees. One area in which exact likelihood computation is often infeasible is multilocus linkage mapping. One method of linkage analysis for rare recessive traits is homozygosity mapping where data on affected inbred individuals are analysed. Key to this method are the patterns of autozygosity in the individuals, and MCMC provides also a method for studying these patterns. Algorithms for the exact computation of autozygosity probabilities on an arbitrary pedigree very rapidly become computationally infeasible. However, an MCMC algorithm can provide accurate estimates in reasonable computing time, and these probabilities can then be used to map the genes responsible for disease.

1. Introduction

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 locations in the genome of genes influencing a given trait are to be inferred.

The elements of genetic models are straightforward: genes exist, genes segregate (are copied) from parents to offspring, and the types of genes carried by an individual influence observable trait characteristics. A locus is a specification of the position of a gene on a chromosome. With modern molecular genetic 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 many 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, $L_1, ..., L_k$ that index segments of DNA on the same chromosome are "linked"; the segregations of genes at two loci are not independent. If the maternal gene at locus L_h in a father segregates to a child, it is more probable that the gene that segregates at an adjacent locus, L_j, 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 fractions", $r_{h,j}$, between the two loci. Specifically, the probability that genes at loci L_h and L_j segregating from one parent to the child have different grandparental origins is $r_{h,j}$. In fact, the value of a recombination fraction between two loci depends on numerous factors, most importantly on the sex of the parent. This fact can be incorporated into analyses, but, for simplicity, is ignored in the current paper.

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 L_h and L_j if there is an odd number of crossover events. The genetic (map) distance between two loci is the expected number of crossovers 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

*Research supported in part by PHS grant GM-46255.

depends on the pattern of interference. 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 occur along the chromosome as a Poisson process rate 1, when the chromosome is measured in units of map distance. In practice, interference exists, particularly where the loci are close together and recombination fractions between them are small. However, the amount of data required to estimate levels and patterns of interference seldom exists in human genetic studies. In genetic mapping, the objective is to detect linkage, to infer locus order, and place loci 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 determined; 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. Third, even where single-locus marker genotypes are observable the haplotype information is not; that is, it is not known which alleles are on the same chromosome, having been received from the same parent. One set of single-locus genotypes (a specification of the unordered pair of alleles at each locus) can correspond to many different multilocus genotypes (a specification of the alleles on each chromosome, at each of the loci). 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. However, more markers, and marker loci with more alleles, and traits observable for 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.

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. 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).

In this paper, we consider one possible approach to the computations needed to map a rare recessive disease from data on affected inbred individuals. We consider only marker loci at which the types of the two genes carried by a observed individual are known, and a recessive disease for which it is known whether or not an observed individual carries two copies of the disease allele. The (multilocus) genotype G_i of individual i is a specification of the types of the genes on each of a pair of chromosomes of the individual. The phenotype Y_i of i is a specification of the observed trait characteristics determined by the underlying genotypes. We subsume all the parameters of the genetic model into the parameter vector θ, and use $P_\theta(\cdot)$ to denote probabilities under the model. The total set of genotypes on a pedigree is denoted $\mathbf{G}$, and of observed phenotypes $\mathbf{Y}$.

2. Monte Carlo likelihood

Monte Carlo estimates of integrals or expectations are not new, either in general (Hammersley and Handscomb, 1964) or in genetic linkage analysis (Thompson et al. 1978). However, Monte Carlo methods have only become widely used with the explosion in use of Markov chain Monte Carlo (MCMC) which permits simulation from distributions known only up to a normalising constant, and hence simulation from conditional distributions. The statistical problems involved in fitting genetic linkage models to trait data, $\mathbf{Y}$, on a set of related individuals may be viewed as latent variable or "missing data" problems. Were all the underlying genetic events observable, likelihood computation and parameter estimation would be trivial, but only trait data (phenotypes) of some individuals are observed. We denote the latent variables by $\mathbf{X}$.

The likelihood is

$$L(\theta) = P_\theta(\mathbf{Y}) = \sum_{\mathbf{X}} P_\theta(\mathbf{Y}, \mathbf{X}) = \sum_{\mathbf{X}} P_\theta(\mathbf{Y}|\mathbf{X}) P_\theta(\mathbf{X}) \quad (1)$$

Although the summation may be infeasible, we suppose that the latent variables, $\mathbf{X}$, are chosen in such a way that each term of the expression is easily computed.

Monte Carlo estimators of likelihood ratios can be based on

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

(Thompson and Guo, 1991), provided simulation from the appropriate distribution is possible. Suppose $\mathbf{X}(l)$,

$l = 1, ..., N$, are realisations from $P_{\theta_0}(\mathbf{X}|\mathbf{Y})$ then a Monte Carlo estimate of the likelihood ratio (2) is

$$\frac{1}{N}\sum_{l=1}^{N}\left(\frac{P_\theta(\mathbf{Y},\mathbf{X}(l))}{P_{\theta_0}(\mathbf{Y},\mathbf{X}(l))}\right) \tag{3}$$

From an importance sampling perspective, the estimator (3) is efficient; for values of θ_0 close to θ the sampling distribution mimics the shape of the integrand $P_\theta(\mathbf{Y},\mathbf{X})$ of (1). Further, equation (3), 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). At least for values of θ close to θ_0, a single simulation provides an estimate of the local likelihood surface.

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

$$P_{\theta_0}(\mathbf{X} \mid \mathbf{Y}) \;=\; P_{\theta_0}(\mathbf{X},\mathbf{Y})/P_{\theta_0}(\mathbf{Y}) \tag{4}$$

With well chosen latent variables $\mathbf{X}$, 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 often 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 $\mathbf{X}$ a "proposal distribution" $q(\cdot,\mathbf{X})$ is defined. Then, if the process is now at $\mathbf{X}$ the next value is generated as follows:

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

$$\lambda = \frac{q(\mathbf{X},\mathbf{X}^*)P_{\theta_0}(\mathbf{X}^*|\mathbf{Y})}{q(\mathbf{X}^*,\mathbf{X})P_{\theta_0}(\mathbf{X}|\mathbf{Y})} = \frac{q(\mathbf{X},\mathbf{X}^*)P_{\theta_0}(\mathbf{Y},\mathbf{X}^*)}{q(\mathbf{X}^*,\mathbf{X})P_{\theta_0}(\mathbf{Y},\mathbf{X})}$$

Note that λ can be computed without knowledge of $P_{\theta_0}(\mathbf{Y})$.

3. With probability $\lambda^* = min(1,\lambda)$ the process moves to $\mathbf{X}^*$ and with probability $(1-\lambda^*)$ it remains at $\mathbf{X}$.

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 $\mathbf{X}$.

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. 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. The space of Lange and Matthysse (1989) is even larger, including also indicators of the grandparental origins of genes. Although local updating methods are very slow, they are convenient for genetic analysis problems. If large changes in genotypic configuration are proposed, the Hastings ratio can be impossible to compute, and constraints in the feasible genotypic patterns on pedigrees mean that almost all proposals have zero probability.

There are various approaches to improving sampler performance in genetic problems. Lin (1993) made great progress towards increasing the practicality of MCMC methods in linkage analysis, using Metropolis-coupled samplers (Geyer, 1991), and a form of "heating" in the Metropolis-Hastings steps to improve mixing of the chain. Geyer and Thompson (1994) used simulated tempering (Marinari and Parisi, 1992) to make sampling feasible on a very large complex pedigree with many constraints. These strategies result in a sampler that can sample genotypes efficiently on a large pedigree. However, for several linked markers, the huge space of possible genotypic configurations that then arises may render the sampler ineffective.

An alternative approach is to consider alternative latent variables $\mathbf{X}$, to produce a smaller space more easily sampled by MCMC methods. 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. For the problem of mapping rare recessive traits from data on inbred affected individuals, it is possible to bypass the multilocus genotypes of unobserved individuals, and use only segregation indicators as the latent variables.

3. Homozygosity mapping.

In linkage analysis, due to uncertainties as to whether an unaffected individual carries a disease gene, the computational difficulties on extended pedigrees, and the costs of typing large numbers of individuals, there have been many approaches towards basing linkage analyses on a small number of observed (usually affected) individuals. The extreme case is *homozygosity mapping* in which a rare recessive is mapped using only marker and trait data on independent inbred 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, and thus receiving two copies of the disease gene identical-by-descent or *autozygous* 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. Unrelated inbred individuals will be homozygous at independent segments of the genome, but 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.

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. The key underlying events that determine the data on the affected inbred individual are the segregations that specify the ancestral genes that he receives. Let m and p index the maternal and paternal segregations to some individual. Let $S_{mj} = 0$ if the maternal allele received by the individual at locus j is of grandmaternal origin, and $S_{mj} = 1$ otherwise, and let S_{pj} be similarly defined for the paternal allele. Then, at any locus j,

$$P(S_{mj} = 0) \;=\; P(S_{mj} = 1) \;=\; P(S_{pj} = 0) \;=\; P(S_{pj} = 1) \;=\; \frac{1}{2}$$

and at two loci h and j

$$P(S_{mh} = S_{mj}) \;=\; P(S_{ph} = S_{pj}) \;=\; (1 - r_{h,j})$$

where $r_{h,j}$ is the recombination fraction between the two loci.

Then for a given segregation i, the recombination events are determined by segregation indicators S_{ij}, $j = 1, ..., k$, where S_{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 $\mathbf{S} = \{S_{ij}\}$ as the latent variables $\mathbf{X}$ in the Monte Carlo likelihood framework of section 2. Figure 1 shows the case of the offspring of a first-cousin marriage. At any locus, the offspring individual may receive genes autozygous from either of his parents' common grandparents; there are eight relevant segregation indicators that will specify the gene descents.

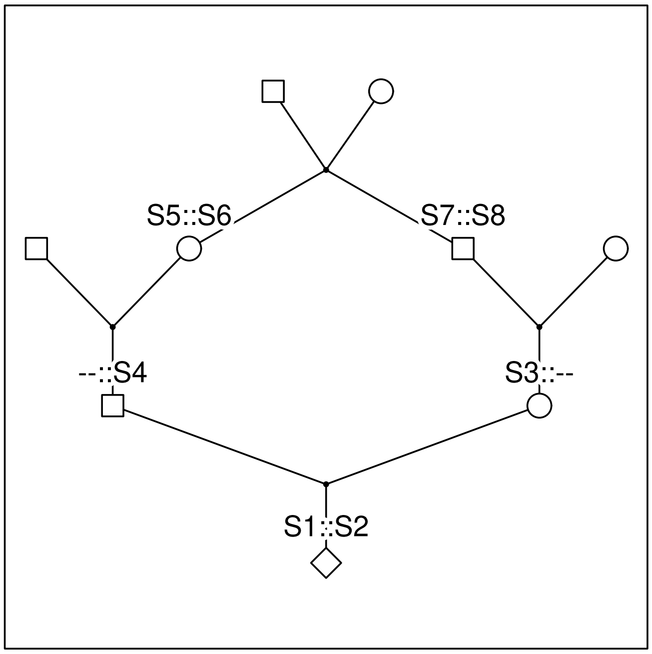

Figure 1: A first cousin marriage, showing segregation indicators.

Table 1: Example of segregation array, for the pedigree of figure 1

Segreg.:	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
Locus								
L_1	0	1	1	1	0	1	0	0
L_2	0	1	1	0	0	1	0	0
L_3	0	1	1	0	0	1	0	1
L_4	0	1	0	0	0	1	0	1

Table 1 shows four successive patterns of values for the eight segregation indicators of figure 1, such as might arise along a chromosome segment, or at four loci. In the first pattern, the paternal gene of the offspring individual derives from his grandmother ($S_1 = 0$), and is the paternal gene of this grandmother ($S_4 = 1$), and is in fact the great-grandfather's maternal gene ($S_5 = 0$). Likewise the final individual's maternal gene is this same maternal gene in his great-grandfather; the individual is autozygous for this gene. By locus 2, S_4 has become 0; the final individual's paternal gene is now the paternal ($S_6 = 1$) gene of his great-grandmother. By locus 3,

S_8 has become 1; this leaves the genes in the final individual unchanged, since $S_3 = 1$, so the grandfather's maternal gene is not transmitted. However, by locus 4, S_3 becomes 0; now the final individual is autozygous for the paternal gene in his great-grandmother.

Consider now Table 1 as illustrating a possible value of the state $\mathbf{S}$ at four loci being used in a genetic analysis. The prior probabilities of $\mathbf{S}$ are straightforward. However, for implementation of a Metropolis algorithm, relative values of $P_{\theta_0}(\mathbf{Y}, \mathbf{S})$ are required, or $P_{\theta_0}(\mathbf{Y} \mid \mathbf{S})$. The binary indicators, $\mathbf{S} = \{S_{ij}\}$, of grandparental origins of genes in each given offspring individual, at each locus readily determine the multilocus autozygosity patterns in the observed individual. This is done simply by following the descent paths of genes, as in the example described above; an efficient algorithm is easily implemented to update these descent paths, and hence the resulting autozygosity pattern, when a S_{ij} changes. For a single observed individual, the autozygosity pattern is k binary indicators, specifying whether or not the S_{ij} result in the individual having two genes autozygous at locus j, $j = 1, ..., k$. The probability of a genotype homozygous for an allele with frequency q is q^2 or q, as the individual is not/is autozygous at the locus. The probability of a heterozygous genotype is 0 if the individual is autozygous at the locus, and is $2q_1q_2$ otherwise, where q_1 and q_2 are the two allele frequencies.

Table 2: Probability ratios of segregation indicators S_{ij}

$S_{i,j-1}$	$S_{i,j+1}$	*probability ratio**
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	$r_{j-1}r_j/(1-r_{j-1})(1-r_j)$

* : $P(S_{ij} = 1 \mid \mathbf{S}_{-(ij)})/P(S_{ij} = 0 \mid \mathbf{S}_{-(ij)})$.
r_h is the recombination frequency between L_h and L_{h+1}.
Note also:
$P(S_{ij} = 1 \mid \mathbf{S}_{-(ij)}) = P(S_{ij} = 1 \mid S_{i,j-1}, S_{i,j+1})$
$P(S_{ij} = 0 \mid \mathbf{S}_{-(ij)}) = P(S_{ij} = 0 \mid S_{i,j-1}, S_{i,j+1})$
$\mathbf{S}_{-(ij)}$ denotes all elements of $\mathbf{S}$ other than S_{ij}.

The space of $\mathbf{S}$-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 $\mathbf{S}$ depends only on the indicators at adjacent loci for the same segregation (Table 2). For example suppose the current $\mathbf{S}$ were that of Table 1, and the proposal was to change $S_{4,2}$ from its current value 0 to 1. This would eliminate a recombination between loci 1 and 2 ($S_{4,1} = 1$) giving probability ratio

$$(1 - r_{1,2})/r_{1,2},$$

but create one between loci 2 and 3 ($S_{4,3} = 0$) giving another factor

$$r_{2,3}/(1 - r_{2,3}).$$

This recombination ratio is then weighted by the appropriate conditional probability of phenotypic observations $P_{\theta_0}(\mathbf{Y} \mid \mathbf{S})$, for current and proposed $\mathbf{S}$-values. This sampler is clearly irreducible: if a given pattern of autozygosity in the observed individual is compatible with the data, then so also is any pattern with fewer loci at which the affected individual is autozygous and hence homozygous.

Werner's syndrome (WS) 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., 1992) in 21 small pedigrees of Japanese and Caucasian origin. The frequency of the disease allele is assumed to be 0.004. A Monte Carlo linkage likelihood analysis of a subset of five of these pedigrees is given by Thompson (1994); here we use just two of the pedigrees for purposes of illustration. Two markers were of significance in the published linkage reports: $D8S87$ and ANK. Originally ANK and $D8S87$ were thought to be flanking markers, but the likely order is now thought to be (WS, $D8S87$, ANK). For the purposes of illustration only, we take the recombination fractions between WS and $D8S87$ and between $D8S87$ and ANK each to be 0.1; this is probably larger than the true values, but of the correct order of magnitude. Data and information on these markers were provided by Dr. Ellen Wijsman (personal communication).

4. Autozygosity probabilities

In fact, for a single affected inbred individual, the data $\mathbf{Y}$ at a position h on a chromosome depend on $\mathbf{S}(h)$ only through $Z(h)$, the autozygosity (I) or non-autozygosity (N) in the inbred affected individual. Over multiple loci, or along the chromosome continuum, these patterns of autozygosity are themselves of interest. Although, for a very rare recessive trait, the posterior probability of autozygosity at the disease locus is very high, the probability that all of a set of unrelated affected inbred individuals are autozygous may be low. Further, the way in which such posterior probabilities are influenced by data on linked markers is non-trivial, for the patterns of autozygosity along a chromosome segment follow no simple process. Specifically, even in the absence of interference, the process is not Markov, since it is an

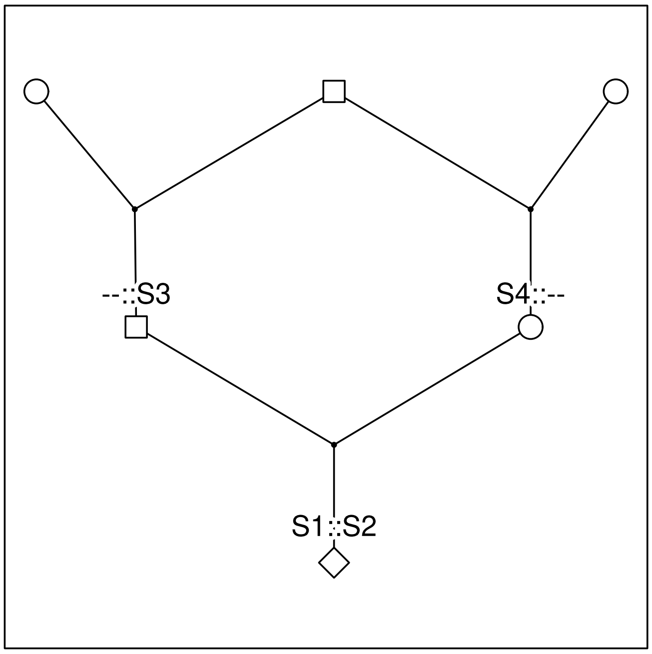

Figure 2: A half-sib marriage example.

aggregate process and shows the clumping phenomenon typical of such processes (Aldous, 1989; Blossey, 1993).

Consider first the prior, disregarding data $\mathbf{Y}$. The smallest non-trivial example consists of the offspring of a half-sib mating (figure 2). This is also the largest example for which the space of $\mathbf{S}$-values can be drawn readily (figure 3). As one moves along the chromosome, the process $\mathbf{S}(h)$ performs a random walk at rate n on the vertices of the n-dimensional hypercube (Donnelly, 1983). Here, $n = 4$ and, without loss of generality, the two vertices positioned as shown in figure 3 are those which result in autozygosity of the inbred offspring individual: $\mathbf{Z}(h) = I$ if $S_1(h) = S_2(h) = 1$ and $S_3(h) = S_4(h)$. Overall, $P(Z(h) = I) = 2/16 = 0.125$. When $Z(h) = I$, the next jump of the random walk will require $Z(h) = N$; when $Z(h) = N$, the next jump results in $Z(h) = I$ with overall probability $(2 \times \frac{1}{2} + 4 \times \frac{1}{4})/14 = 1/7$. However, although by symmetry $Z(h) = I$ is a renewal point of the process, when the process leaves $Z(h) = I$, the probability that the next jump will result in a return to $Z(h) = I$ is 3/8. The overall probability, $P(Z(h) = I)$ can easily be computed on even a complex pedigree; it is simply the inbreeding coefficient of the individual. For two loci, at given recombination fraction, the probability of autozygosity at both loci can be computed by the algorithm of Thompson (1988), again even on a complex pedigree. However, due to the non-Markov pattern of autozygosity along the chromosome, these marginal and

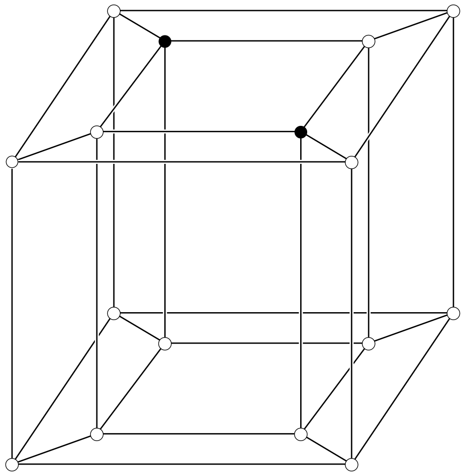

Figure 3: Random walk structure, corresponding to figure 2.

Table 3: Prior autozygosity probabilities for cousin marriage.

state $\mathbf{Z}$			$True^{(1)}$	$Markov^{(2)}$
N	N	N	0.8825	0.8811
N	N	I	0.0264	0.0277
N	I	N	0.0131	0.0131
N	I	I	0.0155	0.0155
I	N	N	0.0264	0.0277
I	N	I	0.0022	0.0009
I	I	N	0.0155	0.0155
I	I	I	0.0184	0.0183

(1) Results from 10^9 MCMC steps and 10^8 i.i.d realisations are almost identical to 10^{-4}.
(2) Results from assuming (incorrectly) a first-order Markov chain for autozygosity at successive loci.

pairwise probabilities do not suffice.

Table 3 shows the autozygosity probabilities for three loci, with recombination fraction 0.1 between each pair of adjacent loci, for the case of a first-cousin marriage (figure 1). For this small problem, exact results could have been obtained, but in fact these are Monte Carlo results, obtained both by 10^9 Metropolis steps of MCMC, and also by 10^8 independent realisations from the prior. For this problem, these two simulations give comparable accuracy (to $\pm 10^{-4}$) in comparable computing time (about 8 hours on a DEC3100). Also shown are the probabilities that would be given by a first-order Markov process with the same pairwise and marginal probabilities. First, it can be seen that $Z = I$ is a renewal point; where the central locus has $Z = I$ the "Markov" results agree with the correct results. Second, the major effect, in terms of relative error, is in the case $\mathbf{Z} = (I, N, I)$; there is a clumping of states $Z = I$ in the jump chain. Alternatively viewed, there is an increased probability of small regions of non-autozygosity (and hence likely heterozygosity at a highly polymorphic marker) within regions of autozygosity (and hence homozygosity). In this example, the sequence (I, N, I) has probability 2.5 times larger than a "Markov" view would predict.

Table 4: Prior autozygosity probabilities for pedigree of figure 4.

state $\mathbf{Z}$			$True^{(1)}$	$Markov^{(2)}$
N	N	N	0.7901	0.7889
N	N	I	0.0478	0.0493
N	I	N	0.0257	0.0251
N	I	I	0.0271	0.0273
I	N	N	0.0478	0.0493
I	N	I	0.0050	0.0031
I	I	N	0.0271	0.0273
I	I	I	0.0295	0.0297

(1) Results from 10^9 MCMC steps and 10^8 i.i.d realisations are almost identical to 10^{-4}.
(2) Results from assuming (incorrectly) a first-order Markov chain for autozygosity at successive loci.

Another example is given in Table 4. Many of the pedigrees in the Werner's syndrome data set are first cousin marriages. The more complex pedigree (figure 4) was first ascertained as a first cousin marriage, but later it was discovered that each parent of the affected proband was also the offspring of a first cousin marriage, as shown. Although this is a small pedigree, exact linkage likelihood computations become infeasible with the standard methods with more than three loci, due to the pedigree complexity. The final offspring individual can be autozygous for a gene in any of the three original founders marked. Again, the "true" results in Table 4 are Monte Carlo results (both 10^8 independent samples and 10^9 MCMC steps, agreeing to 4 decimal places). The "Markov" assumption again underestimates most severely the probability of $\mathbf{Z} = (I, N, I)$. However, note also that now there is no renewal when $Z = I$; the lack of symmetry of the three relevant founder ancestors destroys this property, even though numerically the discrepancies are small.

Generally, for just three loci, only the low-probability state $\mathbf{Z} = (I, N, I)$ shows substantial departure from the

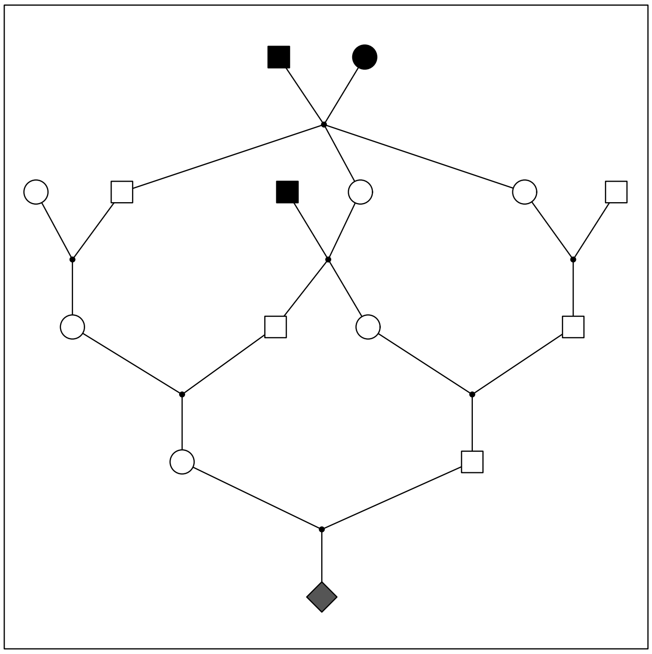

Figure 4: A more complex pedigree.

first-order Markov probability values. However, with data, this state may have high posterior probability. One of the first cousin marriages (figure 1) for the Werner's syndrome (WS) data illustrates this. The data consist of homozygosity (affected) at the WS locus (allele frequency 0.004), heterozygosity at the marker locus $D8S87$ (for two alleles, each frequency 0.5) and homozygosity at the ANK marker. The allele at ANK has population frequency 0.44, so homozygosity is not strong evidence of autozygosity, but the example will serve.

Table 5: Posterior autozygosity probabilities for cousin marriage.

state $\mathbf{Z}$			$MCMC^{(1)}$	$ratio^{(2)}$
N	N	N	0.1001	0.1134
N	N	I	0.0068	0.2576
I	N	N	0.7491	28.3750
I	N	I	0.1440	65.4545

(1) Results 10^9 MCMC steps, agree with *prior* $\times$ *likelihood* to within standard error.
(2) Ratio of posterior to prior probability (see text).

Table 5 shows the posterior probabilities of the four relevant autozygosity states; states autozygous at the $D8S87$ locus are eliminated by the data, and so not listed. As expected, the states with autozygosity at the WS locus have much increased probability *a posteriori*;

the WS disease allele has population frequency only 0.004. Note in particular that the state with lowest prior probability now has a probability 0.1440, 65 times higher than before. Of course, the posterior probabilities could also be obtained by multiplying the prior state probabilities by the likelihoods (and this was done as a check). Prior state probabilities can be efficiently obtained by i.i.d Monte Carlo, but conditional probabilities can only be sampled via MCMC. However, even in this simple example, the standard error of the MCMC estimate for the state INI is smaller, for an equal amount of computing time, due to the 65-fold factor between prior and posterior. When, as here, the range of the ratios of posterior to prior is 3 orders of magnitude, sampling from the prior, and using importance sampling to reweight to the posterior, is far less efficient than sampling from the posterior, even though the latter requires use of MCMC.

5. Discussion

Monte Carlo estimation provides an approach when exact likelihood and probability computation is 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 likelihood estimates via a choice of latent variables. In this paper, we have focussed on one particular choice – the use of segregartion indicators. This seems to have promise in cases where a very few individuals are observed on each of a number of possibly large pedigrees, the individuals being observed for a number of DNA markers. A particular case is homozygosity mapping, where the key is the posterior pattern of autozygosity (gene identity by descent) in affected inbred individuals.

MCMC is used to sample from posterior distributions, but this does not require a Bayesian analysis. Realisations from the distribution of latent variables, conditional on the data, but at prespecified parameter values, can be used to provide efficient Monte Carlo estimates of a likelihood surface. Moreover, while multilocus genotypes are key unobservables in genetic analysis, it may not always be efficient to consider these the latent variables in a Monte Carlo analysis; segregation indicators that specify the passage of genes segregating in a pedigree are more fundamental even than genotypes, and, provided the relevant probabilities of observed data given the latent variables can be easily computed, the genotypes of individuals can be bypassed.

Autozygosity patterns at multiple linked loci become of increasing relevance as multilocus linkage analyses are performed. The random walk framework of Donnelly (1983), and the Posson clumping heuristic of Aldous (1989) together make study of the prior probability distribution of patterns more feasible (Blossey, 1993). However, in order to assess autozygosity in the light of data, or to use realisations from the posterior distribution of autozygosity consitional on data in a likelihood analysis, MCMC provides the most efficient computational approach in many cases. Posterior probabilities of autozygosity patterns are more efficiently estimated by MCMC, than by reweighting prior probabilities estimated by i.i.d Monte Carlo.

Acknowledgement

I am grateful to Ellen Wijsman, for discussion on homozygosity mapping and Werner's syndrome.

References

Aldous, D. J. (1989). *Probability approximations via the Poisson clumping heuristic* Springer; Berlin.

Blossey, H. (1993). *The Poisson clumping heuristic and the survival of genome in small pedigrees* Ph.D. Thesis, University of Washington.

Donnelly, K. P. (1983). The probability that related individuals share some section of the genome identical by descent. *Theoretical Population Biology* **23** 34–64.

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

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

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with Discussion). *Journal of the Royal Statistical Society (B)* **54** 657–699.

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

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 & Co., London.

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

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.

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

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

Marinari, E. and Parisi G. (1992). Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.* **19** 451–458.

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. *Journal of Chemical Physics* **21** 1087–1092.

Ott, J. (1991). *Analysis of Human Genetic Linkage.* 2 nd. edition. The Johns Hopkins University Press; Baltimore, MD.

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

Smith, C. A. B. (1953). Detection of linkage in human genetics. *Journal of the Royal Statistical Society (B)* **15** 153–192.

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.), pp. 247–253. Academic Press, New York.

Thompson, E. A. (1988). Two-locus and three-locus gene identity by descent in pedigrees. *IMA Journal of Mathematics Applied in Medicine & Biology* **5** 261–280.

Thompson E. A and Guo, S.-W. (1991). Monte Carlo evaluation of likelihood ratios. *IMA Journal of Mathematics Applied in Medicine & Biology* **8** 149-169.

Thompson, E. A. (1994). Monte Carlo likelihood in genetic mapping. *Statistical Science:* in press.