MCMC IN THE ANALYSIS OF GENETIC DATA ON PEDIGREES

Elizabeth A. Thompson
Department of Statistics, University of Washington
Box 354322, Seattle, WA 98195-4322, USA
Email: thompson@stat.washington.edu

This chapter provides a tutorial introduction to the use of MCMC in the analysis of data observed for multiple genetic loci on members of extended pedigrees in which there are many missing data. We introduce the specification of pedigrees and inheritance, and the structure of genetic models defining the dependence structure of data. We review exact computational algorithms which can provide a partial solution, and can be used to improve MCMC sampling of inheritance patterns. Realization of inheritance patterns can be used in several ways. Here, we focus on the estimation of multilocus linkage lod scores for the location of a locus affecting a disease trait relative to a known map of genetic marker loci.

Contents

1. Introduction

This chapter provides a tutorial introduction to the use of MCMC in the analysis of data observed for multiple genetic loci on members of extended pedigrees in which there are many missing data. In section 2, we introduce the specification of pedigrees and inheritance, and then in section 3 discuss structure of genetic models defining the dependence structure of data. In section 4, we review exact computational algorithms which can provide a partial solution, and can be used to improve MCMC sampling of inheritance patterns (Section 5). In sections 6 and 7 we show how realizations of inheritance patterns can be used in the Monte Carlo estimation of multilocus linkage lod scores and thence used to find the location of a genetic locus affecting a disease trait. Finally, in section 8 we provide a small illustrative example using simulated data.

This chapter is based on previously published material. For earlier work, readers may consult Thompson [21, 22, 23], in which many references to the previous literature may be found: only a few key references will be repeated here. More recent references will be given: one of these of particular relevance to the efficient MCMC estimation of lod scores is that of George & Thompson [6].

2. Pedigrees, inheritance, and genetic models

A pedigree is a specification of the genealogical relationships among a set of individuals. Each individual is given a unique identifier, and the two parents of each individual are specified. Individuals with unspecified parents are *founders*: the others are *non-founders*. Graphically, males are traditionally represented by squares, females by circles, while any individual of unknown sex may be represented by a diamond. In the graphical representation of a pedigree known as a *marriage node graph* individuals having common offspring are connected to a *marriage node*, and the marriage node is connected to each offspring. See the example in Figure 1. (Although this pedigree structure may appear contrived, it derives from a real study [9].) Each marriage node is connected upward to two parent individuals, and downward to at least one (and possibly many) offspring individuals. Each non-founder is connected upward to precisely one marriage node. A parent individual may be connected to multiple marriage nodes. The shading of individuals may represent affectation status for a particular trait, or other

specified information.

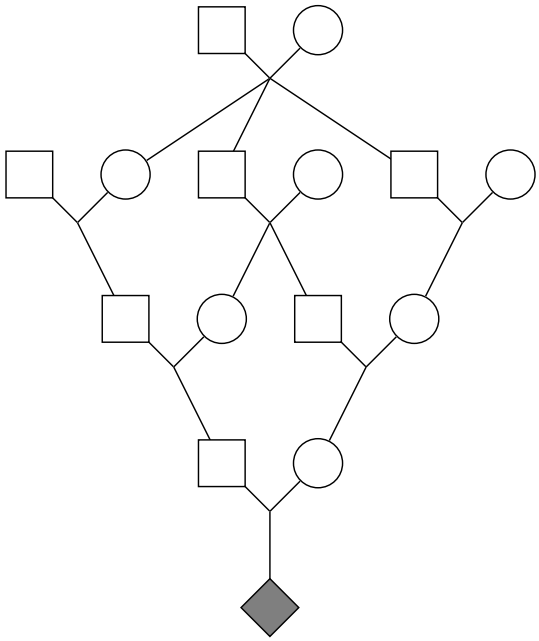

Fig. 1. An example pedigree structure deriving from a real study

Human individuals are diploid: every cell nucleus contains two haploid copies of the DNA of the human genome. One of these copies derives from the DNA in the individual's mother (the maternal genome), and the other from the DNA in the individual's father (the paternal genome). Note that *all* DNA is double-stranded. The double-stranded nature of DNA has nothing to do with the haploid (single genome copy) or diploid (two copy) genome content of a cell or organism. The biological process through which DNA in parent cells is copied and transmitted to offspring is known as *meiosis*, and *Mendel's First Law (1866)* specifies this transmission marginally, at any location in the genome. A genome location is known as a *locus*: the plural is *loci*. In modern terminology, Mendel's First Law states that the

copy transmitted from parent to offspring is a randomly chosen one of the two parental copies, and that all meioses, whether to different offspring of a single parent or in different parental individuals, are independent.

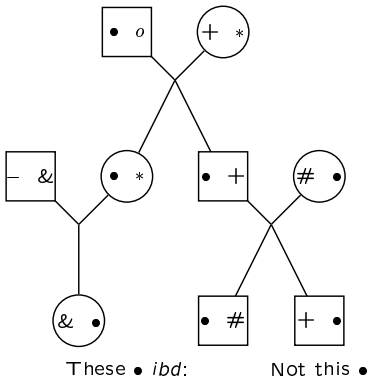

Fig. 2. Identity by descent results in observable similarity among individuals.

Segments of DNA in different genomes that are copies of the same genomic material in a recent common ancestor are said to be *identical by descent* (*ibd*) . Note that *ibd* is always defined relative to a founder population. In the analysis of data on a fixed set of pedigree structures, *ibd* is defined relative to the founders of the pedigrees. By definition, the genomes of founders are nowhere *ibd*. Identity by descent underlies all similarity among relatives that results from the effects of their DNA. The different possible allelic types of the DNA at a locus are known as the *alleles* at that locus. The unordered pair of allelic types that an individual carries at a locus is his *genotype* at that locus. The observable trait characteristics that may be controlled or affected by an individual's genotype at a locus is the individual's *phenotype*.

A small example of the transmission of genome at a single genetic locus is given in Figure 2. One pair of cousins share the "•" *ibd* from their grandparent. The sibling cousin also carries a • symbol at this locus, perhaps representing the same allele (the type of the DNA). However, it can be seen that this • is not *ibd* to the ones in his sibling and cousin, relative

to the specified pedigree. Of course, further investigation might reveal that the two founders carrying • symbols are related, and that these •'s are *ibd* relative to a larger pedigree: *ibd* is always defined relative to the specified pedigree. Given that I have a particular genetic characteristic or phenotype, for example blood type O, the probability my cousins have blood type O is increased, because with some probability they share DNA *ibd* with me at this locus.

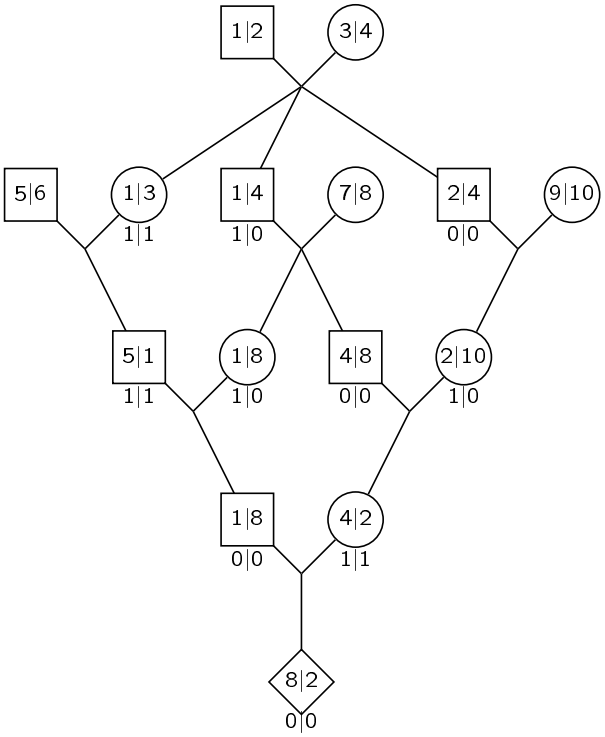

Fig. 3. The inheritance of genome at a specific locus

We can now specify more formally the inheritance of genome at any specific genome location or locus. We provide unique identifiers to each

of the two haploid genomes of every founder. We call these the *founder genome labels* or FGL . In the literature, these are often known as "founder alleles" or "founder genes", but these terms can become ambiguous. The inheritance of the FGL at a particular locus j is specified by binary meiosis indicators

$$S_{i,j} \quad = \quad 0 \text{ or } 1$$

as in meiosis i at locus j the maternal or paternal DNA (respectively) of the parent is transmitted to the offspring.

In Figure 3, an example realization of the paternal and maternal meiosis indicators of each non-founder individual are shown under each individual, with the paternal indicator on the left and maternal indicator on the right. The numbers 1 through 10 in the symbols representing founders are the FGL identifiers. It is easily seen that application of the binary indicators to the FGL enables the descent of FGL down the pedigree to be established. The resulting FGL present in non-founder individuals are also shown in Figure 3. These are the two numbers within the symbols representing each individual, again with the paternally derived FGL on the left and the maternally derived FGL on the right. It is seen that *ibd* at a locus is equivalent to presence of the same FGL at that locus.

We can now specify the inheritance of genome at any set of discrete loci indexed by j, $j = 1, \ldots, l$:

$$\begin{aligned} S_{i,j} = 0 \quad & \text{if DNA at meiosis } i \text{ locus } j \text{ is parent}'\text{s maternal DNA,} \\ = 1 \quad & \text{if DNA at meiosis } i \text{ locus } j \text{ is parent}'\text{s paternal DNA.} \end{aligned}$$

For convenience we define the two sets of vectors each of which makes up the array $\mathbf{S} = \{S_{i,j}\}$:

$$\begin{aligned} S_{\bullet,j} &= \{S_{i,j}; i = 1, \ldots, m\}, \quad j = 1, \ldots, l, \\ S_{i,\bullet} &= \{S_{i,j}; j = 1, \ldots, l\}, \quad i = 1, \ldots, m. \end{aligned}$$

where m is the number of meioses in the pedigree (twice the number of non-founders), and l the number of loci under consideration. In the literature, the vector $S_{\bullet,j}$ is known as the *inheritance vector* at locus j [14].

3. The structure of a genetic model

In order to derive an appropriate probability model for the array of latent meiosis indicators $\mathbf{S}$, we first outline the events in the biological process of meiosis. The DNA in each cell nucleus of an individual is packaged into 46 chromosomes, 23 of which derive from the DNA of the father, and 23 from

the mother. Only one pair differ between individuals of different sex (the sex chromosomes). The two members of each of the other 22 pairs carry essentially the same DNA, although of course at many locations along the chromosome there may be allelic differences. Prior to meiosis, each chromosomes duplicates, but the two parts remain connected at the *centromere*. The two chromosomes of a pair (the maternal and paternal ones in the parental cell nucleus) then become tightly aligned, and may exchange DNA. Through two successive meiotic divisions, the chromosomes separate, leading to four potential offspring gametes (Figure 4). Each gamete (sperm or egg) cell contains a full haploid genome, and may pass to an offspring whose observable genetic characteristics result from the combined diploid DNA of their maternal and paternal gametes.

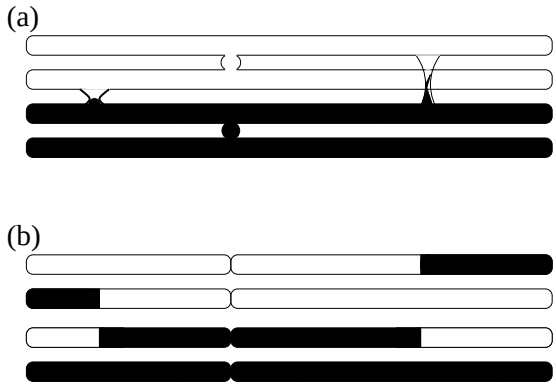

Fig. 4. Meiosis and four resulting potential offspring chromosomes

Each chromosome of the gamete cell consists of alternating segments of the two parental chromosomes. These segments are large, comprising on average about 10^8 base pairs (bp) of DNA. A location at which the DNA switches from the parent's maternal to paternal DNA, or from paternal to maternal, is known as a *crossover*. Between any two loci, the genetic distance d (in Morgans) is defined as the expected number of crossover events between them in an offspring gamete. Since, regardless of dependence, expectations are additive, this definition provides an additive measure of distance along the chromosome. Note that genetic distance is defined through the meiosis process, not in terms of a physical distance such as number of bp. The relationship between physical and genetic distance varies over the

genome, and depends on many factors. A key such factor is the sex of the parent in whom the meiosis occurs. Genetic distances are normally reported in centiMorgans (1 cM = 0.01 Morgans). As a rough average, 1 cM is about 1 megabase (10^6 bp).

In an offspring gamete resulting from a meiosis i, between any two loci j and j', a *recombination* is said to have occurred if the DNA at those locations derived from two different parental chromosomes: $S_{i,j} \neq S_{i,j'}$. The probability of this event is the recombination fraction ρ between the two loci. The value $\rho(d)$ of the recombination fraction at genetic distance d is the *map function*. Under almost all models of meiosis, and apparently in reality, $\rho(0) = 0,\ \rho'(0) = 1,\ \rho(d) \nearrow d$, and $\rho(\infty) = \frac{1}{2}$.

The above considerations are almost sufficient to define the probability model for $\mathbf{S}$. From Mendel's First Law we have that the vectors $S_{i,\bullet}$ are independent, and that $\Pr(S_{i,j} = 0) = \Pr(S_{i,j} = 1) = 1/2$. Now we have also $\Pr(S_{i,j-1} \neq S_{i,j}) = \rho_{j-1},\ \ j = 2, ..., l$, for all $i,\ i = 1, ..., m$. For notational convenience we assume the recombination fraction is the same for all meioses i, but in modeling real data it is important to allow at least for different values in male and in female meioses. Our model now determines the pairwise probability distribution for any two inheritance vectors:

$$\Pr(S_{\bullet,j} \mid S_{\bullet,j-1}) = \rho_{j-1}^{R_{j-1}}(1 - \rho_{j-1})^{m-R_{j-1}}, \tag{1}$$

where $R_{j-1} = (\#i : S_{i,j} \neq S_{i,j-1})$. To define the joint distribution of all the components of $\mathbf{S}$ an additional assumption is required. The simplest is to assume the absence of *genetic interference*. This assumption implies that crossovers arise as a Poisson process (rate 1 per Morgan), and hence that the occurrences of recombination in disjoint intervals of the chromosome are independent. In this case the inheritance vectors $S_{\bullet,j}$ are first-order Markov in j:

$$\Pr(\mathbf{S}) = P(S_{\bullet,1}) \prod_{2}^{l} \Pr(S_{\bullet,j} \mid S_{\bullet,j-1})$$

$$\text{or} \quad \Pr(S_{i,j} | \mathbf{S}_{-(i,j)}) = \Pr(S_{i,j} | S_{i,j-1}, S_{i,j+1}),$$

where $\mathbf{S}_{-(i,j)}$ denotes the set of all components of $\mathbf{S}$ except $S_{i,j}$.

We have specified a model for $\mathbf{S}$, but $\mathbf{S}$ is not observed. The data consist of the trait characteristics of individuals, which are determined by the allelic types of their DNA at the relevant genetic loci. The simplest possible model relating *ibd* to observable data at a single locus is that DNA segments that are *ibd* are of the same allelic type, while non-*ibd* DNA segments are of independent types. While this model ignores the possibility of mutation

within the pedigree, and of possible dependence at the population level among founders of a pedigree, it is an adequate model for most purposes. Use of more general models is possible, if desired. At locus j, we denote by $\mathcal{A}_j$ an allocation of allelic types to the distinct FGL. Our model assumes that the FGL g are independent in their types with, say, type probabilities $q_j(g)$, and, more specifically, that each FGL g has type k independently with some probability $q_{j,k}$. Then

$$\Pr(\mathcal{A}_j) = \prod_g q_j(g) \quad = \quad \prod_g q_{j,k}^{n_j(k)},$$

where $n_j(k)$ is number of FGL g with type k at locus j. We shall also assume independence of the allelic types of a FGL over loci j. Except in very small genetic isolates, this is an accurate assumption for loci for which $\rho > 0.005$. This is fortunate since this assumption is hard to generalize.

Thus we have now all the components of a genetic model, and the classes of parameters involved. The population model, with parameters such as $\mathbf{q} = \{q_{j,k}\}$, provides the probabilities for the latent $\mathcal{A}$, the allelic types of FGL at each j. The inheritance model, with parameters ρ, provides probabilities for the latent $\mathbf{S}$, the inheritance of FGL at j, jointly over j. The *genotype* of an individual at a particular locus is the unordered pair of allelic types of the DNA he carries at that locus. The (phased) multilocus genotype of an individual is the unordered pair of collections of allelic types in his maternal and paternal genomes. The ordered genotype of an individual is the ordered pair of allelic types, conventionally ordered (paternal, maternal). The ordered multilocus genotype is the set of ordered single-locus genotypes, and is the most detailed specification. We will refer to the set of ordered multilocus genotypes for all members of a pedigree structure as $\mathbf{G}$. $\mathbf{G}$ defines both phased multilocus genotypes and the set of genotypes at each locus. Jointly over loci, note that $\mathbf{G}$ is in turn determined by $(\mathbf{S}, \mathcal{A})$. At each locus j, the ordered genotypes $G_{\bullet,j}$ of pedigree members are determined by $S_{\bullet,j}$ and $\mathcal{A}_j$.

The final component of a genetic model is the part that connects the latent genotypes to observable data $\mathbf{Y}$. The penetrance model, with parameters β specifies the probability of trait data $\mathbf{Y}$ given the latent genotypes $\mathbf{G}$. For simplicity, we shall assume that our data $\mathbf{Y}$ can be partitioned into $\{Y_{\bullet,j}; j = 1, ..., l\}$, with $Y_{\bullet,j}$ depending only on $G_{\bullet,j}$. Each locus j may be a DNA marker locus, for which this will naturally be true, or may correspond to a trait. For a marker locus, parameters β may include a typing error model or other factors causing the recorded marker data on an in-

dividual potentially to differ from the true latent genotype. For a more general trait determined by locus j, the assumption is that the only locus in this genome region affecting the trait is the locus j. In this case, the penetrance probabilities $P_\beta(Y_{\bullet,j} \mid G_{\bullet,j})$ may in general depend on observable covariate information on individuals (age, gender, diet, ...) and also on other heritable effects contributed by genes elsewhere in the genome, but not linked to these l loci.

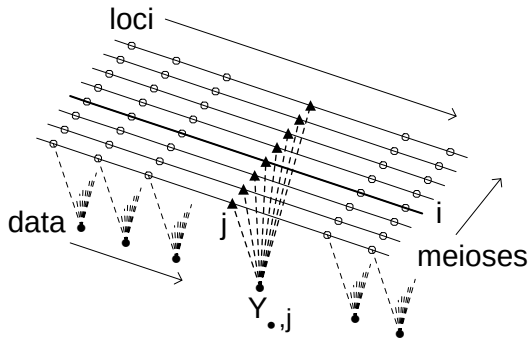

Fig. 5. The dependence structure of pedigree data

The complete set of parameters will be denoted $\xi = (\mathbf{q}, \rho, \beta)$, and the likelihood for the model may be written formally as

$$\begin{aligned} L(\xi) = P_\xi(\mathbf{Y}) &= \sum_{\mathbf{G}} P_\beta(\mathbf{Y} \mid \mathbf{G}) P_{(\mathbf{q},\rho)}(\mathbf{G}) \\ &= \sum_{(\mathbf{S},\mathcal{A})} P_\beta(\mathbf{Y} \mid \mathbf{G}(\mathbf{S},\mathcal{A})) P_\rho(\mathbf{S}) P_{\mathbf{q}}(\mathcal{A}). \end{aligned} \tag{2}$$

The dependence structure of the data $\mathbf{Y}$ in terms of the latent $\mathbf{S}$ is shown in Figure 5. For $\mathbf{S}$, the meioses i are independent, while loci j have first-order Markov dependence. At each locus j, the data $Y_{\bullet,j}$ are determined probabilistically by the latent inheritance pattern $S_{\bullet,j}$. In the representation of Figure 5, the pedigree structure is implicit in the labeling of the meioses. Additionally, the allelic types of the FGL $\mathcal{A}_j$ which also contribute to $G_{\bullet,j}(S_{\bullet,j}, \mathcal{A}_j)$ and hence to $Y_{\bullet,j}$ are omitted. In most contexts, the latent allelic types are nuisance variables which are integrated over (Section 4).

4. Exact computations on pedigrees: peeling algorithms

Before proceeding to MCMC, it is important to consider what parts of the computation may be achieved exactly. Where a partial exact computation is feasible, this may often be incorporated into a Monte Carlo sampling procedure to improve Monte Carlo performance. Additionally, partial exact computation may permit the use of Rao-Blackwellized estimators [5], improving efficiency in the use of sampled realizations. Summations such as those in equation (2), may, depending on the underlying dependence structure, be accomplished via a variety of *peeling* algorithms [2] in which the summation is performed sequentially over subsets of the variables. In the context of signal processing, time series, and hidden Markov models (HMMs) these methods date back to the 1960s and the work of Baum and colleagues [1]. A few years later, similar methods were developed for simple genetic models on pedigrees having a simple tree structure by Elston and colleagues [4]. The methods were generalized to arbitrarily complex pedigree structures and more complex models by Cannings and colleagues later in the 1970s [2, 3, 20], and 10 years later to general graphical structures by Lauritzen and Spiegelhalter [16].

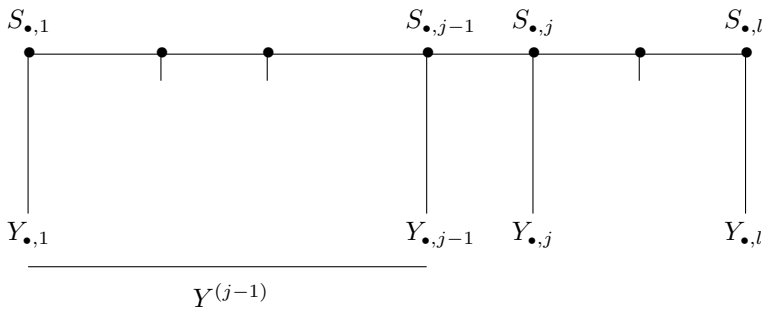

Fig. 6. Dependence structure of data along a chromosome

In the current context we have three relevant structures. The first is the linear structure along a chromosome shown in Figure 6. The second is the undirected structure relating to the assignment of allelic types to FGL, and the third is the directed graphical structure of a pedigree. We consider first the computation of likelihoods on small pedigrees, using the Baum

algorithm. Note that $\Pr(\mathbf{Y})$ may be written

$$\begin{aligned}\Pr(\mathbf{Y}) &= \sum_{\mathbf{S}} \Pr(\mathbf{Y} \mid \mathbf{S})\ \Pr(\mathbf{S}) \\ &= \sum_{\mathbf{S}} \left(\prod_{j=1}^{l} \Pr(Y_{\bullet,j} \mid S_{\bullet,j}) \right) \ \left(\Pr(S_{\bullet,1}) \prod_{2}^{l} \Pr(S_{\bullet,j} \mid S_{\bullet,j-1}) \right).\end{aligned}$$

We defer to below the computation of $\Pr(Y_{\bullet,j} \mid S_{\bullet,j})$, and define $Y^{(j)} = (Y_{\bullet,1}, \ldots, Y_{\bullet,j})$, the data up to and including locus j, and $R^*_j(s) = P(Y^{(j-1)}, S_{\bullet,j})$. Then $R^*_1(s) = \Pr(S_{\bullet,1} = s)$ and

$$\begin{aligned}R^*_{j+1}(s) &= \Pr(Y^{(j)},\ S_{\bullet,j+1} = s) \\ &= \sum_{s^*} \left[\Pr(S_{\bullet,j+1} = s \mid S_{\bullet,j} = s^*)\ \Pr(Y_{\bullet,j} \mid S_{\bullet,j} = s^*)\ R^*_j(s^*) \right],\end{aligned} \tag{3}$$

for $j = 1, 2, \ldots, l-1$, with

$$L \ = \ \Pr(\mathbf{Y}) = \sum_{s^*} \Pr(Y_{\bullet,l} \mid S_{\bullet,l} = s^*)\ R^*_l(s^*).$$

Since $S_{\bullet,j}$ can take 2^m values, where m is number of meioses, computation using equation (4) is limited to small pedigrees.

To facilitate discussion of computation on directed graphs, it is convenient to note an alternate form of equation (4) in which the computation is done in the reverse direction along the chromosome, but the transition probabilities are still used in the direction $\Pr(S_{\bullet,j+1} = s \mid S_{\bullet,j} = s^*)$. Now we define $R_j(s) = \Pr(Y_{\bullet,j+1}, ..., Y_{\bullet,l} \mid S_{\bullet,j} = s)$. Then

$$\begin{aligned}R_{j-1}(s) &= \Pr(Y_{\bullet,j}, ...Y_{\bullet,l} \mid S_{\bullet,j-1} = s) \\ &= \sum_{s^*} \left[\Pr(S_{\bullet,j} = s \mid S_{\bullet,j-1} = s^*)\ \Pr(Y_{\bullet,j} \mid S_{\bullet,j} = s^*)\ R_j(s^*) \right].\end{aligned} \tag{4}$$

Of course, only one of equations (4) and (5) is needed in order to compute the likelihood, and in any case the transition probabilities on a chromosome may be considered in either direction. However, even for this undirected linear case, both forms are useful, since then

$$\Pr(S_{\bullet,j} = s \mid \mathbf{Y}) \propto R^*_j(s)\ R_j(s)\ \Pr(Y_{\bullet,j} | S_{\bullet,j} = s). \tag{5}$$

Thus if computation of likelihoods $\Pr(\mathbf{Y})$ is feasible, so too is computation of the conditional probability of latent variables $S_{\bullet,j}$ given all the data $\mathbf{Y}$.

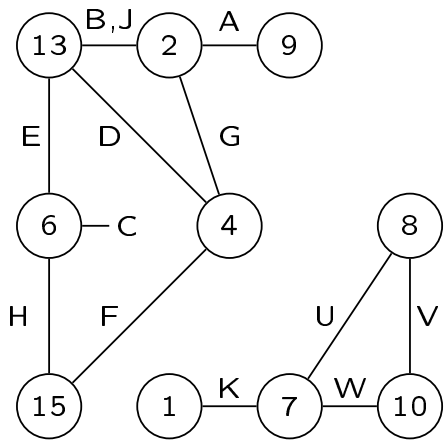

Fig. 7. Peeling the FGL graph to compute $\Pr(Y_{\bullet,j}|S_{\bullet,j})$

We consider now the deferred computation of $\Pr(Y_{\bullet,j}|S_{\bullet,j})$:

$$\begin{aligned}\Pr(Y_{\bullet,j}|S_{\bullet,j}) &= \sum_{\mathcal{A}_j} \Pr(Y_{\bullet,j}|\mathbf{G}(\mathbf{S_{\bullet,j}}, \mathcal{A}_\mathbf{j}))\Pr(\mathcal{A}_\mathbf{j}) \\ &= \sum_{\mathcal{A}_j} \left(\prod_n \Pr(Y_{n,j}|G_{n,j}(S_{\bullet,j}, \mathcal{A}_j))\right)\left(\prod_g q_j(g)\right), \qquad (6)\end{aligned}$$

where here $Y_{n,j}$ denotes the phenotype of an observed individual n at locus j, $G_{n,j}$ is the genotype of individual n at locus j, and g is an FGL. Again peeling is simply a reorganization of the joint summation over all $\mathcal{A}_j$ in order to perform the summation sequentially. We illustrate this with an example (Figure 7) taken from Thompson [24]. In the FGL graph, the nodes are FGL and each edge corresponds to at least one observed individual. The scenario corresponds to a pedigree in which there are presumably at least 8 founders, since the highest FGL label is 15 and each founder has two. At the locus j in question, there are observed individuals A, B, C, ..., U, V, W. It is supposed that under the specified inheritance pattern $S_{\bullet,j}$, individual A receives FGL 2 and 9, both B and J receive FGL 2 and 13, and so on. An edge joins the two FGL received by each observed individuals. The parents

of individual C must be related, since C received two copies of FGL 6.

First, for an FGL such as 3 or 11, not present in any observed individual under the $S_{\bullet,j}$ under consideration, any allelic type may be assigned, and the contribution of the FGL to sum in equation (6) is a factor $\sum_k q_{j,k} = 1$. These FGL can thus be ignored. Second, where the FGL falls into two or more disconnected components, as in Figure 7 the sum in equation (6) factorizes into the contributions from summation over allelic allocations to the FGL in each component. Thus each component may be considered separately, and the results multiplied. Thus we will consider the summation only on the larger component of Figure 7. Finally, FGL enter together into a term in the sum, only through the probability of a phenotype $Y_{n,j}$ of an individual given the allelic types assigned to the two FGL that he carries. Thus, in our example, equation (6) becomes

$$\begin{aligned}\Pr(Y_{\bullet,j}|S_{\bullet,j}) &= \sum_{\mathbf{g}} \left(\prod_n \Pr(Y_n|g_{n,1},g_{n,2}) \right) q(g_2)q(g_6)q(g_4)q(g_9)q(g_{13})q(g_{15}) \\ &= \sum_{g_6} q(g_6)\Pr(Y_C|g_6) \left(\sum_{g_{15}} q(g_{15})\ \Pr(Y_H|g_{15},g_6) \right. \\ &\left(\sum_{g_4} q(g_4)\Pr(Y_F|g_{15},g_4) \left(\sum_{g_{13}} q(g_{13})\Pr(Y_E|g_6,g_{13})\Pr(Y_D|g_4,g_{13}) \right. \right. \\ &\left(\sum_{g_2} q(g_2)\Pr(Y_B|g_2,g_{13})\Pr(Y_J|g_2,g_{13})\Pr(Y_G|g_2,g_4) \right. \\ &\left. \left. \left(\left(\sum_{g_9} q(g_9)\Pr(Y_A|g_2,g_9) \right) \right) \right) \right) \right) .\end{aligned}$$

Beginning with the final term, data Y_A is incorporated into the summation over the allelic types assigned to FGL 9, for each value of the allelic type of FGL 2. Then including data on B, J, and G, the summation over the allelic types assigned to FGL 2 is done for each assignment to FGL 4 and 13, and so on through the graph. Of course, summations may be done in any order: this example shows one sensible ordering.

Where there are many observed individuals in a pedigree, the FGL graph may become complex, and even the best sequential summation may involve consideration jointly of too many FGL for the method to be feasible. However, where data are relatively sparse, the FGL graph is often quite small even on a large and complex extended pedigree. There is one case where the summation of equation (6) is trivial, however much data or however large

and complex the pedigree. This is the case of a genotypic marker observed without the possibility of error. In this case, on each disjoint component of the FGL graph, there are 0, 1, or 2 allelic assignments consistent with the data. Consider, for example, the smaller component of Figure 7. Suppose we observe that each of K, U and W is of genotype ab at locus j, and suppose that at this locus the population allele frequency of allele a is q_a and of b is q_b. Considering first individual K, we see that there are two possibilities: $g(7) = a,\ g(1) = b$ or $g(7) = b,\ g(1) = a$. Including the information on U and W, these two possibilities remain and

$$\begin{aligned} & g(7) = a,\ g(1) = g(8) = g(10) = b; \quad \text{probability contribution } q_a q_b^3 \\ \text{or } & g(7) = b,\ g(1) = g(8) = g(10) = a; \quad \text{probability contribution } q_a^3 q_b, \end{aligned}$$

giving a total probability contribution $q_a q_b^3 + q_a^3 q_b$. Now suppose V is also observed. If V has genotype aa, $g(8) = g(10) = a$ and only the second alternative remains: the probability is then $q_a^3 q_b$. If V has genotype bb, $g(8) = g(10) = b$ and only the first alternative remains: the probability is then $q_a q_b^3$. If V is observed to have any other genotype, there is no feasible allelic assignment on this component of the FGL-graph, and the probability of these data on the pedigree, under this particular inheritance pattern $S_{\bullet,j}$, has probability 0.

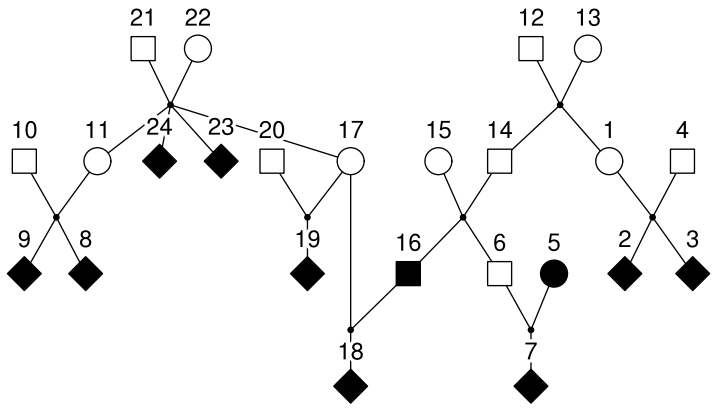

Fig. 8. A pedigree without loops

Finally in this section, we consider peeling on the directed graph representing a pedigree structure. Figure 8 shows a marriage node graph representation of a pedigree with no loops, but not a simple tree structure. The shaded individuals are assumed to have data, and the joint probability of all the observed data under a specified genetic model is to be computed. Conditional on genotypes of parents, data on each grandparent couple and on each offspring are all mutually independent. Thus the idea of pedigree peeling is to accumulate the probability sequentially over the pedigree, using these individual genotypes as the latent variables. In a pedigree without loops, the part of the pedigree on which the probability has been accumulated will either be connected to an individual B through his parents, and will be denoted $A(B)$ or through his descendants of spouses, in which case it will be denoted $D(B)$.

Analogously to the HMM case we define two R-functions for any individual B,

$$\begin{aligned} R_B(g) &= \Pr(Y_C,\ C \in D(B) \mid G_B = g) \\ R^*_B(g) &= \Pr(Y_C,\ C \in A(B),\ G_B = g). \end{aligned} \tag{7}$$

For the example of Figure 8 we see:

$$\begin{aligned} R_1(g) &= \Pr(Y_2, Y_3 \mid G_1 = g) \\ &= \sum_{g^*} \Pr(G_4 = g^*) \\ &\qquad \left(\sum_{g'} \Pr(Y_2|G_2 = g') \Pr(G_2 = g'|G_1 = g, G_4 = g^*) \right) \\ &\qquad \left(\sum_{g''} \Pr(Y_3|G_3 = g'') \Pr(G_3 = g''|G_1 = g, G_4 = g^*) \right), \end{aligned}$$

and then

$$\begin{aligned} R^*_{14}(g) &= \Pr(Y_2, Y_3, G_{14} = g) \\ &= \sum_{g'} \Pr(G_{12} = g') \\ &\qquad \left(\sum_{g''} \Pr(G_{13} = g'') \Pr(G_{14} = g \mid G_{12} = g', G_{13} = g'') \right. \\ &\qquad\qquad \left. \left(\sum_{g^*} \Pr(G_1 = g^* \mid G_{12} = g', G_{13} = g'') R_1(g^*) \right) \right). \end{aligned}$$

Thus, using the R and R^* functions we may accumulate the probability of observed data over the entire pedigree. We refer the reader to [21] for details. Because the pedigree is a directed graph, in which the genetic model specifies probabilities of offspring genotypes given those of their parents, both R and R^* functions will generally be used in working through a pedigree. In the two initial peeling steps shown here, a function $R_1(g)$ is used in accumulating up to individual 1 from $D(1) = \{2, 3, 4\}$ and a function R^*_{14} in accumulating down to individual 14 from $A(14) = \{12, 13, 1, 2, 3, 4\}$. However, it is only the interpretation of the function as a conditional or joint probability that is affected. The form of the sequential summation equation is the same whether one is peeling up or down: one simply inserts the appropriate founder genotype probabilities (for 4 and then for (12,13)), penetrances (for (2,3)), previously computed R-functions ($R_1(\cdot)$), and transmissions from parents to offspring (from (1,4) to (2.3) and then from (12,13) to (1,14)).

For the directed HMM, functions $R^*(\cdot)$ were used in peeling forward, and $R(\cdot)$ in peeling backward, and equation (5) shows how these may be combined to provide the probabilities $\Pr(S_{\bullet,j} \mid \mathbf{Y})$. An analogous result applies here. In peeling a pedigree in one order, from right to left in the example of Figure 8 one obtains R for some individuals, such as 1, and R^* for other individuals, such as 14. Reversing the peeling order, and working from left to right, one would obtain the function R for 14, and R^* for 1. Thus by working in both directions, and storing the functions computed, one has, for each individual B, the terms needed to compute

$$\begin{aligned}\Pr(G_B = g \mid \mathbf{Y}) &\propto \Pr(Y_C, C \in A(B), G_B = g)\ \Pr(Y_B \mid G_B = g) \\ &\qquad \Pr(Y_C, C \in D(B) \mid G_B = g) \\ &= R^*_B(g)\ \Pr(Y_B \mid G_B = g)\ R_B(g).\end{aligned}$$

The same procedures, both with regard to peeling and the computation of marginal genotype probabilities for each individual given the full data $\mathbf{Y}$ apply also to pedigrees with loops. The only difference is that the genotypes of several individuals may need to be considered jointly in peeling, as in peeling the FGL-graph, and that the resulting R-functions may be partially of type R^* and partially of type R. That is, they are probabilities of data on a peeled section of the pedigree, jointly with the genotypes of individuals whose parents have been peeled, conditional on the genotypes of individuals whose descendants have been peeled. Again this does not affect the form of the equations for the R-functions, only the interpretation of the resulting function. We refer to [21] and earlier literature cited therein for details.

Finally, we consider the form of genotypes that will be used. For peeling multiple loci over a pedigree, phased genotypes are necessary. If locus j has k_j alleles, there are $K = \prod_{j=1}^{l} k_j$ possible haplotypes, and $\frac{1}{2}K(K+1)$ possible phased genotypes for an individual. Since for each possible combination of genotypes of parents we consider the possible genotypes of each child, peeling complexity (even on a pedigree without loops) is of order K^6, and hence exponential in the number of loci. For convenience in combining with the meiosis patterns $\mathbf{S}$, we often prefer to use ordered rather than unordered genotypes. Thus there are K^2 rather than $\frac{1}{2}K(K+1)$ genotypes to consider for each individual. Extra store is then required (although still of order K^6), but of course the symmetries can be used to avoid extra computation. For complex pedigrees, more individuals must be considered jointly both in storage and in computation. Thus the pedigree peeling algorithm is linear in pedigree size, but exponential both in pedigree complexity and number of loci, and can be computationally challenging even for a single locus if the number of alleles is large.

5. MCMC on pedigree structures

We have seen how probabilities can be computed on small pedigrees for multiple loci using the Baum algorithm, and on extended pedigrees for a very few loci using pedigree peeling. However, when both the size of the pedigree (as measured by the number of meioses m) and number of loci (l) are large, exact computation is infeasible, and some form of Monte Carlo or approximation must be used. We note that the computation of $\Pr(Y_{\bullet,j} \mid S_{\bullet,j})$ by peeling the FGL-graph is limited neither by pedigree size nor number of linked loci, but may become computationally challenging if there are large numbers of FGL and large numbers of combinations of FGL possible in observed individuals. However, normally the FGL graph partitions into manageable components, and we will focus on MCMC methods for sampling $\mathbf{S}$ given data $\mathbf{Y}$ assuming $\Pr(Y_{\bullet,j} \mid S_{\bullet,j})$ readily computable. We note that even where exact computation is possible, peeling provides only probabilities $\Pr(S_{\bullet,j} = s \mid \mathbf{Y})$ for each j (equation (5)), or, at best, probabilities $\Pr(S_{\bullet,j} = s, S_{\bullet,j+1} \mid \mathbf{Y})$ for pairs of adjacent loci [21]. Monte Carlo will provide realizations from $\Pr(\mathbf{S} \mid \mathbf{Y})$, the full joint distribution of $\mathbf{S}$ given all the data $\mathbf{Y}$.

For the sampling of $\mathbf{S}$, the dependence structure of Figure 5 immediately suggests several possible block Gibbs samplers, each updating a subset S_u of $\mathbf{S} = \{S_{i,j}\}$ conditional on $\mathbf{Y}$ and on the rest of $\mathbf{S}$ (S_f). The first of these

is the locus-sampler or L-sampler [12, 10], in which each updating set S_u is $S_{\bullet,j}$ for some j. Now

$$\Pr(S_{\bullet,j}|\{S_{\bullet,j'}, j' \neq j\}, \mathbf{Y}) = \Pr(S_{\bullet,j}|S_{\bullet,j-1}, S_{\bullet,j+1}, Y_{\bullet,j}),$$

and resampling from this distribution requires computation of

$$\Pr(Y_{\bullet,j} \mid S_{\bullet,j-1}, S_{\bullet,j+1}) = \sum_{S_{\bullet,j}} \Pr(Y_{\bullet,j} \mid S_{\bullet,j}) \Pr(S_{\bullet,j} \mid S_{\bullet,j-1}, S_{\bullet,j+1}).$$

This is simply a single-locus pedigree-peeling computation in which the Mendelian transmission probabilities are replaced by the meiosis-specific values $\Pr(S_{i,j}|S_{i,j-1}, S_{i,j+1})$. Thus the L-sampler can be implemented on any pedigree on which single-locus peeling is feasible. Provided each inter-locus recombination fraction is strictly positive, the sampler is clearly irreducible. However, if the loci are tightly linked, mixing performance will be poor.

An alternative block-Gibbs sampler is the M-sampler [26], in which each updating set S_u is a subset of the meiosis indicators $S_{i,\bullet}$ for a set of meioses $i \in M^*$ over all loci. Computation and resampling from the probabilities

$$\Pr(\{S_{i,\bullet}; i \in M^*\} \mid \mathbf{Y}, \{S_{i',\bullet}; i' \notin M^*\})$$

requires peeling along the chromosome using the Baum algorithm, with a state space of size $2^{|M^*|}$. In the basic M-sampler [26], each meiosis is resampled separately ($|M^*| = 1$). Proposals for joint updating of several meiosis indicator vectors have been made [21, 19]: these can substantially improve performance. Unfortunately, unless $|M^*| = m$ which is infeasible, it is hard to show that the M-sampler is irreducible. Moreover, although it is not affected by tight linkage, since the meiosis indicators over all loci are updated jointly, it can perform poorly on extended pedigrees where there are many missing data.

Each of our currently implemented L- and M-samplers does a random scan of loci or meioses, respectively. That is, at each scan a random permutation of loci [meioses] is formed, and then the vectors $S_{\bullet,j}$ [$S_{i,\bullet}$] are updated from their full conditional distributions in the order specified by the permutation. The L-sampler and M-sampler have somewhat orthogonal performance characteristics, the L-sampler performing well on extended pedigrees under loose linkage and the M-sampler on small pedigrees under tight linkage. Of course, any valid MCMC samplers can be combined, and our LM-sampler, which combines L- and M-sampler [26] usually has much better mixing performance than either. In this case, before each scan, a decision is made to do an L-sample or M-sample scan with probabilities p

and $1-p$ respectively, independently of past history or current state of the system. Although the optimal p should depend on the linkage map, pedigree structure, and extent of missing data, we have found little difference in performance provided $0.2 \leq p \leq 0.8$, so typically we choose $p = 0.2, 0.5$ or 0.8, depending on which sampler, if either, is substantially more computationally intensive in the data set at hand.

Gibbs samplers, even block-Gibbs samplers, have a tendency to explore a space locally, and not make large changes in the latent variables. Metropolis-Hastings rejected restarts can be a way to make larger changes [27]. In the current context, sequential imputation [13] provides a possible proposal distribution for restarts as well as a way to obtain good initial starting configurations [6]. Realizations $S^*_{\bullet,j}$ of the inheritance vectors $S_{\bullet,j}$ are obtained sequentially, each one conditionally on the previously realized $S^*_{\bullet,j-1}$ and on $Y_{\bullet,j}$. This leads to the sequential imputation sampling distribution for data on extended pedigrees given by [11]:

$$P^*(\mathbf{S}^*) = \prod_{j=1}^{l} P_{\xi_0}(S_{\bullet,j} \mid S^{*(j-1)}, Y^{(j)}) \quad = \quad \frac{P_{\xi_0}(\mathbf{S}^*, \mathbf{Y})}{W_l(\mathbf{S}^*)}, \tag{8}$$

where $W_l(\mathbf{S}^*) = \prod_{j=1}^{l} w_j$ and

$$w_j = P_{\xi_0}(Y_{\bullet,j} \mid Y^{(j-1)}, S^{*(j-1)}) \quad = \quad P_{\xi_0}(Y_{\bullet,j} \mid S^*_{\bullet,j-1}).$$

Weights w_j and hence $W_l(\mathbf{S})$ can be computed: each predictive weight w_j is the conditional probability of data observations $Y_{\bullet,j}$ and is obtainable by single-locus peeling, with meiosis-specific transition probabilities determined by the previously realized $S^*_{\bullet,j-1}$.

If $P^*(\cdot)$ is used as a proposal distribution $q(\cdot; \mathbf{S})$, then the Metropolis-Hasting acceptance probability, for a proposal $\mathbf{S}^\dagger$ when the current configuration is $\mathbf{S}$, becomes $\max(1, h)$ where the Hastings ratio h is given by

$$\begin{aligned} h(\mathbf{S}^\dagger; \mathbf{S}) &= \frac{q(\mathbf{S}; \mathbf{S}^\dagger) P_\xi(\mathbf{S}^\dagger, \mathbf{Y})}{q(\mathbf{S}^\dagger; \mathbf{S}) P_\xi(\mathbf{S}, \mathbf{Y})} \\ &= \frac{P_\xi(\mathbf{S}, \mathbf{Y}) W_l(\mathbf{S}^\dagger) P_\xi(\mathbf{S}^\dagger, \mathbf{Y})}{W_l(\mathbf{S}) P_\xi(\mathbf{S}^\dagger, \mathbf{Y}) P_\xi(\mathbf{S}, \mathbf{Y})} \quad = \quad \frac{W_l(\mathbf{S}^\dagger)}{W_l(\mathbf{S})}. \end{aligned} \tag{9}$$

Thus the Hastings ratio is just the ratio of weights, which are easily computed, and for $\mathbf{S}^\dagger$ must be computed already in making the proposal. Although these Metropolis-Hastings proposals are easily incorporated, acceptance probabilities may be low. In preliminary examples, the procedure works well for up to about 5 loci, but for larger numbers of loci substantial changes in $\mathbf{S}$ proposed by sequential imputation are rarely accepted [6].

Another area in which Metropolis-Hastings proposals may be used is to allow for a general model of interference, I, while still using the HMM dependence structure which depends on the assumption of no interference of the Haldane model, H [21]. Suppose the interference model provides probabilities $P^{(I)}(S_{i,\bullet})$ for meiosis i, in place of Haldane model $P^{(H)}(S_{i,\bullet})$ we have used so far. Of course, under either model, the vectors $S_{i,\bullet}$ are independent over i, $i = 1, ..., m$. Suppose the current configuration is $\mathbf{S}$ and any block-Gibbs update of S_u, keeping fixed $S_f = \mathbf{S} \setminus S_u$, under the Haldane model is used as a Metropolis-Hastings proposal $\mathbf{S}^\dagger$. The Hastings ratio is

$$\begin{aligned} h(\mathbf{S}^\dagger; \mathbf{S}) &= \frac{P^{(I)}(\mathbf{S}^\dagger, \mathbf{Y})}{P^{(I)}(\mathbf{S}, \mathbf{Y})} \frac{P^{(H)}(S_u \mid S_f, \mathbf{Y})}{P^{(H)}(S_u^\dagger \mid S_f, \mathbf{Y})} \\ &= \frac{P^{(I)}(\mathbf{S}^\dagger, \mathbf{Y}) P^{(H)}(\mathbf{S}, \mathbf{Y})}{P^{(I)}(\mathbf{S}, \mathbf{Y}) P^{(H)}(\mathbf{S}^\dagger, \mathbf{Y})} \\ &= \frac{P(\mathbf{Y}|\mathbf{S}^\dagger) P^{(I)}(\mathbf{S}^\dagger) P(\mathbf{Y}|\mathbf{S}) P^{(H)}(\mathbf{S})}{P(\mathbf{Y}|\mathbf{S}) P^{(I)}(\mathbf{S}) P(\mathbf{Y}|\mathbf{S}^\dagger) P^{(H)}(\mathbf{S}^\dagger)} \\ &= \frac{P^{(H)}(\mathbf{S})}{P^{(I)}(\mathbf{S})} \frac{P^{(I)}(\mathbf{S}^\dagger)}{P^{(H)}(\mathbf{S}^\dagger)} \\ &= \prod_{k=1}^{m} \frac{P^{(H)}(S_{k,\bullet})}{P^{(I)}(S_{k,\bullet})} \frac{P^{(I)}(S_{k,\bullet}^\dagger)}{P^{(H)}(S_{k,\bullet}^\dagger)}. \end{aligned}$$

In the case of the M-sampler, this is particularly straightforward, since only one or a few meioses $i \in M^*$ are updated, and the product reduces to

$$\prod_{k \in M^*} \frac{P^{(H)}(S_{k,\bullet})}{P^{(I)}(S_{k,\bullet})} \frac{P^{(I)}(S_{k,\bullet}^\dagger)}{P^{(H)}(S_{k,\bullet}^\dagger)}.$$

For moderate numbers of loci $l \leq 14$ the ratios of the probabilities, under interference (I) and Haldane (H) models, of the 2^{l-1} vectors of recombination and non-recombination indicators (equation (1)) may be computed once and pre-stored. For larger numbers of loci, an interference model permitting rapid computation of probabilities $P^{(I)}(\cdot)$ is necessary. While this Metropolis-Hastings algorithm is easily implemented, and performs well, it is, of course, also possible to sample entirely under the Haldane model and reweight realizations $\mathbf{S}$ with weights $P^{(I)}(\mathbf{S})/P^{(H)}(\mathbf{S})$. Which procedure is more computationally effective will depend on how close are probabilities of configurations $\mathbf{S}$ under the two models.

6. Genetic mapping and the location lod score

In this section we introduce the framework and notation for likelihood-based inference for the presence, linked to a set of genetic marker loci, of a genetic locus affecting a trait, and for estimation of the location of this trait locus relative to the set of marker loci. Specifically, what is used is the log-likelihood curve or *lod score* for the location of a locus underlying a trait of interest.

The human genome consists of 3×10^9 base pairs (bp) of DNA. There are now many known DNA variants that can be typed in individuals and whose genomic locations are known. These DNA variants of known genomic locations are known as *genetic markers*, and the objective is to determine the locations of DNA variation underlying a trait relative to the known marker positions. Current DNA markers are broadly of two types. There are microsatellite marker loci. At each of these loci there are many potential alleles that chromosomes may carry. However, in a typical study only several hundred marker loci spread across the genome will be typed. Thus the spacing of these markers is of order 10^7bp. The alternative are SNP markers: each of these typically has only two alleles, but many more exist. There may be as many as 3×10^6 SNP variants in the human genome: potentially, one could type a marker every 1000bp. For the purposes of linkage detection and initial localization of trait loci, microsatellite markers are more readily obtainable and more easily analyzed.

While the probability model for $S_{i,\bullet}$ is defined in terms of recombination fractions, in mapping it is convenient to represent the locations of markers and trait loci on an additive scale. Genetic distance d (in Morgans) between two loci defines this additive metric, and is the expected number of crossovers between the two loci in an offspring chromosome (Section 3). Recall that the assumption of no genetic interference is equivalent to the assumption that crossovers arise as a Poisson process of rate 1 (per Morgan). In this case, the number of crossovers $W(d)$ has a Poisson distribution with mean d. Further, there is a recombination between two loci if $W(d)$ is odd. This gives rise to the Haldane map function

$$\rho(d) = (1/2)(1 - \exp(-2d)).$$

Other meiosis models give rise to other map functions. The key thing is the model: the map function just puts loci onto a linear location map.

While traits of biological or medical importance may be affected by DNA at many loci and by environmental factors, and with complex interactions, simple Mendelian genetics applies well to DNA markers. Thus we assume

a known genetic marker model, including the marker map. That is, we have l genetic markers at known locations λ_i in the genome, and known allele frequencies $\mathbf{q_i}$, $i = 1, ..., l$. The marker model is parametrized by $\Lambda_M = \{\lambda_i, \mathbf{q_i}\}$. For the purposes of lod score estimation, it is also necessary to assume a trait model, parametrized by β, specifying how the trait is determined by underlying genes. The linkage analysis objective is then, given data on the trait phenotypes and marker genotypes for some of the members of some number of pedigree structures, to estimate the location γ of a locus (if any) affecting the trait, in this marker region of the genome. The trait model may incorporate the effects of observable environmental covariates, and even other genetic effects of genes unlinked to these markers, but the question at issue concerns only the existence of linkage and the location γ.

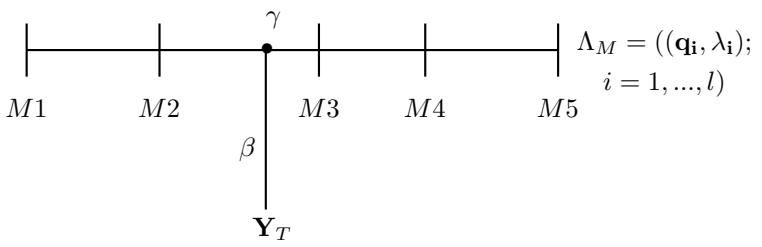

Fig. 9. Defining the location lod score

The data consist of both trait data and marker data, $\mathbf{Y} = (\mathbf{Y}_M, Y_T)$, and the full model is now indexed by parameter $\xi = (\beta, \gamma, \Lambda_M)$. The model is shown schematically in Figure 9. The trait locus location γ is the parameter of interest: $\gamma = \infty$ implies absence of linkage of the trait to these markers. The statistical approach taken is then to compute a likelihood and hence a *location lod score*:

$$lod(\gamma) = \log_{10}\left(\frac{\Pr(\mathbf{Y}; \Lambda_M, \beta, \gamma)}{\Pr(\mathbf{Y}; \Lambda_M, \beta, \gamma = \infty)}\right). \tag{10}$$

Note that the lod score is simply a log-likelihood difference, although traditionally in this area logs to base 10 are used rather than natural logarithms.

More importantly, note that the models in numerator and denominator differ only in γ. The likelihood of a particular location γ is compared to the likelihood of no linkage ($\gamma = \infty$), under the *same* trait model (β) and marker model (Λ_M).

We have seen that if a pedigree has too large a number of meioses m, or joint analysis of a data at a large number $l+1$ of loci is desired, then exact computation of likelihoods in infeasible. Thus we must now consider how MCMC realizations $\mathbf{S}^{(\tau)}$ $\tau = 1, ..., N$ sampled conditionally on marker and/or trait data can be used to provide a Monte Carlo estimate of the relevant likelihoods, and hence of the location lod score curve (10).

7. Monte Carlo likelihood on pedigrees

Monte Carlo estimates expectations, and we have the general formula

$$L(\xi) = P_\xi(\mathbf{Y}) \;=\; \sum_{\mathbf{S}} P_\xi(\mathbf{S}, \mathbf{Y}) \;=\; \mathrm{E}_{P^*}(P_\xi(\mathbf{S}, \mathbf{Y})/P^*(\mathbf{S})), \quad (11)$$

where P^* is any sampling distribution for $\mathbf{S}$ whose support includes that of $P_\xi(\mathbf{S} \mid \mathbf{Y})$. That is, $P^*(\mathbf{S}) > 0$ if $P_\xi(\mathbf{S},\ \mathbf{Y}) > 0$. If N realizations $\mathbf{S}^{(\tau)}$, $\tau = 1, ..., N$ are made from $P^*(\cdot)$ then $N^{-1}\sum_{\tau=1}^{N} P_\xi(\mathbf{S}^{(\tau)}, \mathbf{Y})/P^*(\mathbf{S}^{(\tau)}))$ is an unbiased Monte Carlo estimator of the expectation (11). Of course, the properties of this estimator, other than unbiasedness, will depend on the joint distribution of the $\mathbf{S}^{(\tau)}$. Using MCMC, the $\mathbf{S}^{(\tau)}$ will normally be (possibly subsampled) successive realizations from an ergodic Markov chain.

The simplest possible sampling distribution is $P^*(\mathbf{S}) = P_\xi(\mathbf{S})$ leading to the expression

$$L(\xi) \;=\; \mathrm{E}_\xi(P_\xi(\mathbf{Y} \mid \mathbf{S})).$$

However this form is generally not useful. Few realizations from $P_\xi(\mathbf{S})$ will even give positive probabilities $P_\xi(\mathbf{Y} \mid \mathbf{S})$. From equation (11), in accordance with importance sampling principles, what is needed for effective Monte Carlo estimation of $L(\xi)$ is a sampling distribution $P^*(\mathbf{S})$ close to proportional to the numerator $P_\xi(\mathbf{S}, \mathbf{Y})$. That is, $P^*(\cdot)$ should be close to $P_\xi(\cdot|\mathbf{Y})$.

Sequential imputation [13] is one attempt to find such a distribution giving rise to the sampling distribution of equation (8). Now

$$L(\xi_0) \;=\; P_{\xi_0}(\mathbf{Y}) = \mathrm{E}_{P^*}\left(\frac{P_\xi(\mathbf{S},\ \mathbf{Y})}{P^*(\mathbf{S})}\right) \;=\; \mathrm{E}_{P^*}(W_l(\mathbf{S}^*)).$$

Given N realizations $\mathbf{S}^{(\tau)}$ the estimate of $L(\xi_0)$ is $N^{-1}\sum_\tau W_l(\mathbf{S}^{(\tau)})$. For moderate numbers of not too informative markers, sequential imputation

can perform well [11, 18], just as when used as a Metropolis-Hastings proposal distribution for MCMC (equation (9)). However, for large numbers of loci with multiple alleles, the sequential imputation sampling distribution can differ widely from the target distribution $P_\xi(\mathbf{S}|\mathbf{Y})$ and a very large Monte Carlo sample size N would be required to achieve reasonable estimates.

Another attempt to obtain a good sampling distribution was proposed by [25]. Since we want $P^*(\mathbf{S})$ close to $P_\xi(\mathbf{S}, \mathbf{Y})$ we choose $P^*(\mathbf{S}) = P_{\xi_0}(\mathbf{S} \mid \mathbf{Y})$, where $\xi_0 \approx \xi$ and sample from this distribution using MCMC. Then

$$\begin{aligned} P_\xi(\mathbf{Y}) = \sum_{\mathbf{S}} P_\xi(\mathbf{Y}, \mathbf{S}) &= \sum_{\mathbf{S}} \frac{P_\xi(\mathbf{Y}, \mathbf{S})}{P_{\xi_0}(\mathbf{S} \mid \mathbf{Y})} P_{\xi_0}(\mathbf{S} \mid \mathbf{Y}) \\ &= \mathrm{E}_{\xi_0}\left(\frac{P_\xi(\mathbf{Y},\ \mathbf{S})}{P_{\xi_0}(\mathbf{S} \mid \mathbf{Y})} \mid \mathbf{Y} \right) \\ &= P_{\xi_0}(\mathbf{Y})\ \mathrm{E}_{\xi_0}\left(\frac{P_\xi(\mathbf{Y}, \mathbf{S})}{P_{\xi_0}(\mathbf{Y}, \mathbf{S})} \mid \mathbf{Y} \right). \end{aligned}$$

Thus we have

$$\frac{L(\xi)}{L(\xi_0)} = \frac{P_\xi(\mathbf{Y})}{P_{\xi_0}(\mathbf{Y})} = \mathrm{E}_{\xi_0}\left(\frac{P_\xi(\mathbf{Y}, \mathbf{S})}{P_{\xi_0}(\mathbf{Y}, \mathbf{S})} \mid \mathbf{Y} \right). \tag{12}$$

If $\mathbf{S}^{(\tau)}$, $\tau = 1, ..., N$, are realized from $P_{\xi_0}(\cdot|\mathbf{Y})$ then the likelihood ratio can be estimated by

$$\frac{1}{N} \sum_{\tau=1}^{N} \left(\frac{P_\xi(\mathbf{Y}, \mathbf{S}^{(\tau)})}{P_{\xi_0}(\mathbf{Y}, \mathbf{S}^{(\tau)})} \right).$$

The form for linkage lod score that follows directly from equation (12) is

$$\frac{L(\beta,\ \gamma_1,\ \Lambda_M)}{L(\beta,\ \gamma_0,\ \Lambda_M)} = \mathrm{E}_{\xi_0}\left(\frac{P_{\xi_1}(\mathbf{Y}_T, \mathbf{Y}_M, \mathbf{S}_T, \mathbf{S}_M)}{P_{\xi_0}(\mathbf{Y}_T, \mathbf{Y}_M, \mathbf{S}_T, \mathbf{S}_M)} \mid \mathbf{Y}_T, \mathbf{Y}_M \right),$$

for two hypothesized trait locus positions γ_1 and γ_0. Now $P_\xi(\mathbf{Y}, \mathbf{S}) = P_\beta(Y_T|S_T) P_{\Lambda_M}(\mathbf{Y}_M, \mathbf{S}_M) P_\gamma(S_T|\mathbf{S}_M)$, so the likelihood ratio reduces to

$$\frac{L(\beta,\ \gamma_1,\ \Lambda_M)}{L(\beta,\ \gamma_0,\ \Lambda_M)} = \mathrm{E}_{\xi_0}\left(\frac{P_{\gamma_1}(\mathbf{S}_T \mid \mathbf{S}_M)}{P_{\gamma_0}(\mathbf{S}_T \mid \mathbf{S}_M)} \mid \mathbf{Y}_T, \mathbf{Y}_M \right). \tag{13}$$

This provides a very simple estimator. Consider for example two hypothesized trait locations γ_0 and γ_1 within the same marker interval from marker j to j'. Then, for each meiosis i, we score whether or not there is recombination between the trait and marker j, and independently between

the trait and marker j', using the recombination fractions appropriate to the two hypothesized locations γ_0 and γ_1:

$$\frac{P_{\gamma_1}(\mathbf{S}_T \mid \mathbf{S}_M)}{P_{\gamma_0}(\mathbf{S}_T \mid \mathbf{S}_M)} = \prod_{i=1}^{m} \left[\left(\frac{\rho_{1j}}{\rho_{0j}}\right)^{|S_{i,T}-S_{i,j}|} \left(\frac{1-\rho_{1j}}{1-\rho_{0j}}\right)^{1-|S_{i,T}-S_{i,j}|} \right.$$
$$\left. \left(\frac{\rho_{1j'}}{\rho_{0j'}}\right)^{|S_{i,T}-S_{i,j'}|} \left(\frac{1-\rho_{1j'}}{1-\rho_{0j'}}\right)^{1-|S_{i,T}-S_{i,j'}|} \right],$$

where ρ_{0j} is the recombination fraction between trait and marker j under hypothesized trait location γ_0, and the other recombination fractions have the analogous interpretations.

This likelihood ratio estimator only works well when $\gamma_1 \approx \gamma_0$, but of course local likelihood-ratio estimates may be multiplied to accumulate a likelihood ratio between more distant hypotheses. Although this will require MCMC to be performed at numerous points, the procedure works very well when likelihood surfaces are smooth. This is the procedure implemented in our MORGAN package (www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml) in our first MCMC lod score estimation program lm_lods [21]. However, it does not work well for estimating multipoint location lod scores with highly informative markers, because the likelihoods are not smooth across markers, and because the distributions $P_\gamma(\mathbf{S}_T \mid \mathbf{S}_M)$ change abruptly as γ crosses a marker location.

An alternative approach was provided by Lange and Sobel [15]. They write the likelihood in the form

$$\begin{aligned} L(\beta, \gamma, \Lambda_M) &= P_{\beta,\gamma,\Lambda_M}(\mathbf{Y}_M, \mathbf{Y}_T) \;\propto\; P_{\beta,\gamma,\Lambda_M}(\mathbf{Y}_T \mid \mathbf{Y}_M) \\ &= \sum_{\mathbf{S}_M} P_{\beta,\gamma}(\mathbf{Y}_T \mid \mathbf{S}_M) P_{\Lambda_M}(\mathbf{S}_M \mid \mathbf{Y}_M) \\ &= \mathrm{E}_{\Lambda_M}(P_{\beta,\gamma}(\mathbf{Y}_T \mid \mathbf{S}_M) \mid \mathbf{Y}_M). \end{aligned} \tag{14}$$

This provides an MCMC estimator based on sampling realizations of $\mathbf{S}_M$, $\mathbf{S}_M^{(\tau)}$, $\tau = 1, ..., N$, given $\mathbf{Y}_M$. For each realization $\mathbf{S}_M^{(\tau)}$, $P_{\beta,\gamma}(Y_T|\mathbf{S}_M^{(\tau)})$ is computed for each γ (and for each β) of interest. The MCMC here is quite efficient in that it need be done once only for the fixed marker data and marker model. Also, note that S_T are never even realized, and that the estimator integrates over S_T

$$P_{\beta,\gamma}(\mathbf{Y}_T \mid \mathbf{S}_M^{(\tau)})) = \sum_{S_T} P_\beta(\mathbf{Y}_T \mid S_T) P_\gamma(S_T \mid \mathbf{S}_M)$$

in a form of Rao-Blackwellization [5]. The computation is again accomplished by single-locus peeling, with meiosis-specific transition probabilities determined by the realized inheritance vectors at neighboring markers. Since Y_T is not used in the Monte Carlo, this estimator can perform quite poorly when the trait data provide information on inheritance patterns and $P(S_M|Y_M)$ differs substantially from $P(S_M|Y_M, Y_T)$ [21]. Also, although the sampling procedure is much simpler than for the likelihood ratio estimator (13), multiple peeling operations given each realized $\mathbf{S}_M^{(\tau)}$) are required to implement the estimate. This is computationally intensive on complex pedigrees. However, for simple pedigrees, and where there are substantial marker data on the pedigree, relative to trait data, the method works well. We have implemented the estimator (14) in our MORGAN package under the name lm_markers, so named because we use our LM-sampler for the MCMC, and because sampling is based on the marker data and model only.

Other authors ([10] for example) have developed fully Bayesian MCMC approaches to the problem of linkage detection and estimation. These approaches permit the use of more complex trait models, which are sampled over, with priors being placed on parameters β. In the current notation samples are obtained from the posterior distribution $\pi_{\Lambda_M}(\beta, \gamma, \mathbf{S} \mid \mathbf{Y})$. Of course, this does not produce a lod score, and from a likelihood perspective there are at least three main problems. First, the typically multidimensional parameter β is confounded with locations γ in the posterior distribution: for a likelihood we wish to compare alternative γ under a fixed β. Second, γ is typically treated as a continuous variable, with values binned in order to present posterior probabilities, whereas likelihood is a pointwise function of γ. Third, in sampling posterior probabilities, low-probability areas are not of interest, but in estimating a likelihood ratio relative to the trait locus being unlinked, we require good sampling both of the unlinked and linked locations. This can be hard if either there is a strong positive linkage signal, or a strong negative linkage signal.

In [6] we have developed an approach that retains some of the advantages of the Bayesian method in sampling over trait locations, but which avoids the above three problems. First we fix $\theta = (\Lambda_M, \beta)$, so that our full model is now $\xi = (\theta, \gamma)$. Next note that, for any prior distribution $\pi(\gamma)$, for the single parameter γ

$$\pi_\theta(\gamma|\mathbf{Y}) \propto P_\theta(\mathbf{Y};\gamma)\ \pi(\gamma) \quad \text{so} \quad L(\gamma)\ \propto\ \pi_\theta(\gamma|\mathbf{Y})/\pi(\gamma).$$

Thus a likelihood may be regained from the posterior. To estimate $L(\gamma)$ at

a set of discrete locations, it is only necessary that the prior distribution $\pi(\gamma)$ has support consisting of precisely that set of points. Further, since $\pi(\gamma)$ is arbitrary, it is chosen to improve the Monte Carlo estimate of the likelihood. In this sense it is a *pseudo-prior* [8]. We would like to choose this pseudo-prior so that the posterior distribution is approximately a discrete uniform over the set of positions γ.

Thus our sampling procedure implemented in our lm_bayes program is as follows:
(1) To update $(\mathbf{S}_M, S_T)$, use the block-Gibbs LM-sampler as before.
(2) To update γ, use a Metropolis-Hastings proposal γ^*, with integrated acceptance probability depending only on $\mathbf{S}_M$ (not on S_T).
(3) Update S_T given $(\gamma, \mathbf{S}_M)$, using the new γ^* if it was accepted.

(Steps (2) and (3) are equivalent to a joint update of $(\gamma, \mathbf{S}_T)$.)
Additionally, sequential imputation is used both to provide a starting configuration and also for Metropolis-Hastings rejected restarts, as described in section 5. To choose the prior, we use either estimates from another analysis, perhaps using each marker separately, or a uniform prior, to obtain a preliminary estimate of the posterior, and an order-of-magnitude estimate of the likelihood. Then the prior is readjusted, to be the inverse of this preliminary likelihood estimate, in order that in the main run sampling is approximately uniform across values of γ.

Suppose now we have MCMC realizations $(\gamma^{(\tau)}, \mathbf{S}^{(\tau)})$ from the posterior given $\mathbf{Y} = (\mathbf{Y}_M, Y_T)$, $\tau = 1, ..., N$. A crude estimator of the likelihood is then

$$\widehat{L(\gamma)}_1 = N^{-1} \sum_{\tau=1}^{N} I(\gamma^{(\tau)} = \gamma)/\pi(\gamma),$$

but a better estimator is obtainable by Rao-Blackwellization:

$$\widehat{L(\gamma)}_2 = N^{-1} \sum_{\tau=1}^{N} g(\mathbf{S}_M^{(\tau)}, \gamma),$$

where

$$g(\mathbf{S}_M, \gamma) = E_{\pi_\theta} \left(\frac{I(\gamma)}{\pi(\gamma)} \middle| \mathbf{S}_M, \mathbf{Y} \right).$$

Note than the crude estimator is function of only of the realized $\gamma^{(\tau)}$, while the improved estimator is a function only of the realized $\mathbf{S}_M^{(\tau)}$.

Now we may compute this Rao-Blackwellized estimator:

$$\begin{aligned} g(\mathbf{S}_M, \gamma) = E_{\pi_\theta}\left(\frac{I(\gamma)}{\pi(\gamma)}\middle| \mathbf{S}_M, \mathbf{Y}\right) &= \frac{P_\theta(\gamma, \mid \mathbf{S}_M, \mathbf{Y}_M, Y_T)}{\pi(\gamma)} \\ &= \frac{P_\theta(Y_T \mid \mathbf{S}_M, \mathbf{Y}_M, \gamma) P_\theta(\mathbf{S}_M, \mathbf{Y}_M)\pi(\gamma)}{\pi(\gamma)\sum_{\gamma^*} P_\theta(Y_T \mid \mathbf{S}_M, \mathbf{Y}_M, \gamma^*) P_\theta(\mathbf{S}_M, \mathbf{Y}_M)\pi(\gamma^*)} \\ &= \frac{P_\theta(Y_T \mid \mathbf{S}_M, \gamma)}{\sum_{\gamma^*} P_\theta(Y_T \mid \mathbf{S}_M, \gamma^*)\pi(\gamma^*)}. \end{aligned} \tag{15}$$

We can see some close similarities between the estimator based on (15) and that of equation (14). In both cases, for each realized $\mathbf{S}_M^{(\tau)}$, $P_\xi(Y_T|\mathbf{S}_M^{(\tau)})$ is computed for the given values of Λ_M and β and for each γ of interest, using the same integration over S_T given the realized $\mathbf{S}_M^{(\tau)}$). The major difference is in the sampling, where instead of sampling only $\mathbf{S}_M$ given only $\mathbf{Y}_M$, sampling is of $(\mathbf{S}_M, \gamma)$ given $(\mathbf{Y}_M, Y_T)$ at given β. The sampling of γ provides for better mixing, as in the Bayesian approaches, while conditioning on the trait data Y_T in sampling provides for a sampling distribution closer to the ideal target.

8. An illustrative example

We present here a small example using simulated data. For a more extensive study of performance on simulated data see [6]. For a study of real data, using tightly linked marker loci, and a variety of extended pedigree structures, including complex pedigrees, see [7]. As in the example here, even where exact computations are feasible, accurate Monte Carlo estimates of the lod score can be obtained with far less CPU time [6, 7].

Data were simulated on a simple tree-structure pedigree of 52 individuals over 5 generations (Figure 10). Inheritance patterns at 10 marker loci, equally spaced at 10 cM distances, and at a trait locus at the mid-point between markers 5 and 6, were simulated. Each marker locus was assumed to have only four alleles, with population frequencies 0.4, 0.3, 0.2, and 0.1. The trait locus had two alleles, each with frequency 0.5. The 32 individuals shaded dark in Figure 10 were assumed fully observed for marker and trait information. The simple pedigree structure and limited number of alleles at each marker were chosen to facilitate comparisons with exact computations. The program VITESSE [17] can compute exact lod scores on this pedigree using no more than four markers. For the purposes of illustration here, we use only markers 1, 4, 6 and 10 (M1, M4, M6, M10). The choice of which individuals were observed was made to give an overall proportion

(60%) typical of real data on extended pedigrees, with unobserved individuals predominantly in the earlier generations. However, the choice was made to have missing data on many recent parent individuals, making this slightly more challenging for the MCMC methods.

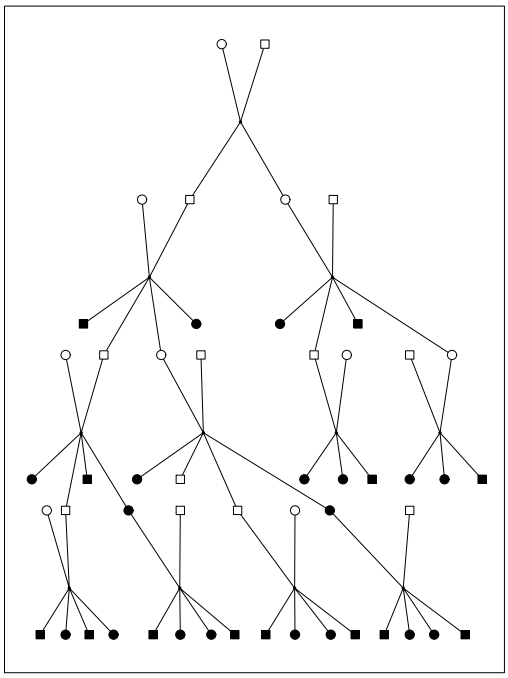

Fig. 10. Pedigree with 52 members. The 32 shaded individuals are assumed observed for simulated trait and marker data.

The trait locus was used to define three different traits. First, and most simply, is was assumed that the genotype at the trait locus was known for the 32 observed individuals: the genotypic trait. Next a quantitative trait was created by assuming the three trait genotypes gave rise to observations with mean 90.0, 100.0 and 110.0 respectively, with each observation having an independent additive residual with variance 25.0. Finally, the quantitative trait was dichotomized, with individuals with quantitative trait values over 98.0 being denoted "affected", and the remainder of the 32 observed individuals "unaffected". We refer to this affected/unaffected classification as the (dichotomized) phenotypic trait. In the analysis model we used the

simulation values for all the trait and marker model parameters. For the phenotypic trait, we used the approximate empirical values 0.05, 0.6, 0.95 for the probability that an individual of each of the three trait genotypes would be observed as "affected".

We have seen three methods for MCMC estimation of location lod scores, each of which is implemented in the package MORGAN (www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml). The likelihood-ratio method (equation (12)) is implemented in our older MORGAN program lm_lods, the Lange-Sobel estimator (equation (14)) is implemented in lm_markers, and our new pseudo-Bayes estimator in lm_bayes. The prefix "lm" on each program indicates that the MCMC in each case uses the LM-sampler. For the genotypic and phenotypic traits, location lod scores were estimated using all three programs. As yet, a quantitative trait is implemented only for the program lm_markers. Lod scores were estimated at approximately equally spaced locations in the map; 8 locations between M1 and M4, 5 between M4 and M6, and 11 between M6 and M10. Additionally, lod scores were estimated at 5 locations at each end of the map, two quite close to each end marker (M1 and M10) and the others ranging out to about 110 cM from the markers. Together with the unlinked location, this provides 35 locations at which likelihoods or likelihood ratios are to be estimated.

For each program, an L-sampler probability of 0.2 was used, and there were 150 scans of burn-in. All runs were quite short: for lm_markers, 3000 MCMC scans were used, for lm_bayes the preliminary run was 1500 scans, and the main run 3000 scans, and for lm_lods only 300 scans were used at each of the 35 evaluation points. For these short runs, results were obtained in about 1 minute of CPU for each of the three programs, for the genotypic trait, and in 3, 5 and 8.5 minutes respectively for lm_markers, lm_bayes, and lm_lods, for the dichotomized phenotypic trait. For the quantitative trait, lm_markers took 2 minutes. In a study of real data, substantially longer runs would be used. Location lod scores were also computed at 16 positions within the marker map including the four marker positions using VITESSE [17]. For each of the three traits, these runs took of the order of several hours CPU on a comparable computer (Joe Rothstein: pers. comm.). The computed VITESSE lods scores and the MCMC estimates for each of the three traits are shown in Figure 11.

For the genotypic trait, we see the lod score is very well estimated by lm_markers and even better by lm_bayes, except right at the markers. In fact we do not attempt to estimate at marker locations: our closest positions

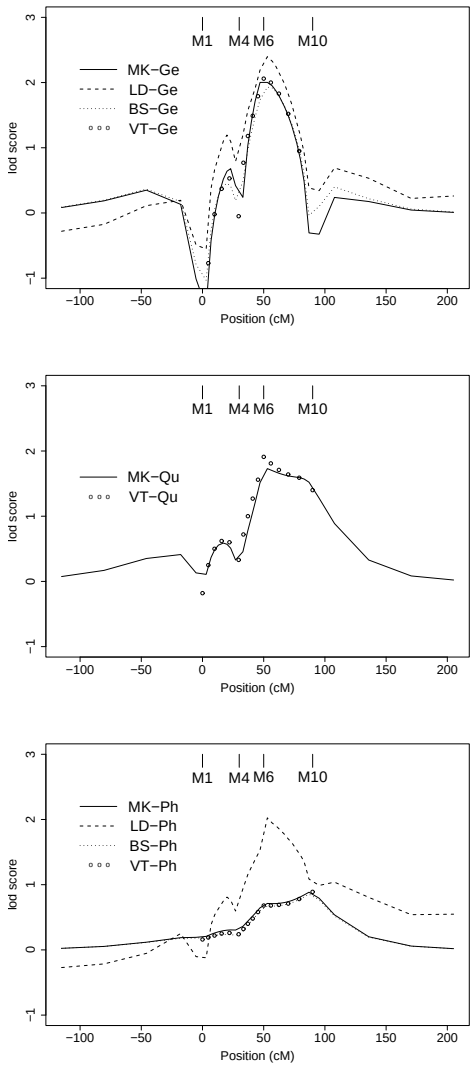

Fig. 11. Location lod scores for example pedigree, providing comparisons of MCMC-based estimates using lm_markers (MK), lm_lods (LD), and lm_bayes (BS) with exact results using VITESSE (VT). The top graph compares results for the genotypic trait (Ge), the center for the quantitative trait (Qu), and the bottom for the dichotomized phenotypic trait (Ph).

are 3cM from each marker. Moreover, at M1 and at M10 the true lod score is $-\infty$, which clearly cannot be estimated by MCMC. The lm_lods program provides a less accurate but still adequate estimate. Note in particular that the shape within each marker interval is well estimated, but that this approach has difficulties in estimating across marker boundaries (Section 7). For both genotypic and quantitative trait, the lod score is apparently maximized at M6, which is not surprising given the true trait location 5cM to the left of M6. The quantitative trait provides less information for linkage, but not much less: the main difference is that lod scores at marker locations are no longer $-\infty$, and indeed the lod score remains high at M10. Again, lm_markers provides an accurate estimate of the location lod score, given the fact that it is based on only 3,000 MCMC scans. For the phenotypic trait, there is little information for linkage, and in fact the maximum lod score of under 1.0 is at M10. Again, the program lm_markers does an excellent job, and lm_bayes an even better one. The lm_lods curve shown is a poor estimate, and the lm_lods MCMC has clearly remained stuck in a part of the space corresponding to the (now) unobserved trait genotypes. Other lm_lods runs (not shown) provided better results, but the results varied widely over runs. Reliable estimation using the lm_lods estimator requires far more MCMC.

9. Conclusion

On large pedigrees with data at multiple linked loci, and with substantial amounts of missing data, exact computation of probabilities and likelihoods is infeasible. Although feasible in principle, sampling of latent inheritance patterns given genetic data remains a challenging MCMC problem for these problems. Likelihood and lod score estimators can be based on latent inheritance patterns realized using MCMC, but it is important to have good estimators as well as good samplers. Lod scores based on multiple markers provide additional information on gene localization: this improved estimation is important for localizing the genes of complex traits. With good MCMC samplers and good estimators, real-time MCMC estimation of multipoint lod scores for a trait locus position is both feasible and practical. Even when exact computation is feasible, MCMC can provide an accurate result with far less computational effort.

Acknowledgment

This chapter was supported in part by NIH grant GM-46255. I am grateful to Joe Rothstein for simulating the marker and trait data on the 52-member pedigree used in section 8, and for the exact lod score computations on these data using VITESSE.

References

1. Baum LE, Petrie T. Statistical inference for probabilistic functions of finite state Markov chains. Annals of Mathematical Statistics 1966;37:1554–1563.
2. Cannings C, Thompson EA, Skolnick MH. Probability functions on complex pedigrees. Advances of Applied Probability 1978;10:26–61.
3. Cannings C, Thompson EA, Skolnick MH. Pedigree analysis of complex models. In *Current Developments in Anthropological Genetics* (edited by J Mielke, M Crawford). Plenum Press, New York, 1980; pp. 251–298.
4. Elston RC, Stewart J. A general model for the analysis of pedigree data. Human Heredity 1971;21:523–542.
5. Gelfand AE, Smith AFM. Sampling based approaches to calculating marginal densities. Journal of the American Statistical Association 1990;46:193–227.
6. George AW, Thompson EA. Multipoint linkage analyses for disease mapping in extended pedigrees: A Markov chain Monte Carlo approach. Statistical Science 2003;18:515–531.
7. George AW, Wijsman EM, Thompson EA. MCMC multilocus lod scores: Application of a new approach. Submitted 2004;.
8. Geyer CJ, Thompson EA. Annealing Markov chain Monte Carlo with applications to ancestral inference. Journal of the American Statistical Association 1995;90:909–920.
9. Goddard KA, Yu CE, Oshima J, Miki T, Nakura J, Piussan C, Martin GM, et al. Toward localization of the Werner syndrome gene by linkage disequilibrium and ancestral haplotyping: lessons learned from analysis of 35 chromosome 8p11.1-21.1 markers. American Journal of Human Genetics 1996; 58:1286–1302.
10. Heath SC. Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. American Journal of Human Genetics 1997;61:748–760.
11. Irwin M, Cox N, Kong A. Sequential imputation for multilocus linkage analysis. Proceedings of the National Academy of Sciences (USA) 1994;91:11684–11688.
12. Kong A. Analysis of pedigree data using methods combining peeling and Gibbs sampling. In *Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (edited by EM Keramidas, SM Kaufman). Interface Foundation of North America, Fairfax Station, VA, 1991; pp. 379–385.
13. Kong A, Liu J, Wong WH. Sequential imputations and Bayesian missing data problems. Journal of the American Statistical Association 1994;89:278–288.
14. Lander ES, Green P. Construction of multilocus genetic linkage maps in hu-

mans. Proceedings of the National Academy of Sciences (USA) 1987;84:2363–2367.
15. Lange K, Sobel E. A random walk method for computing genetic location scores. American Journal of Human Genetics 1991;49:1320–1334.
16. Lauritzen SL, Spiegelhalter DJ. Local computations with probabilities on graphical structures and their application to expert systems. Journal of the Royal Statistical Society, B 1988;50:157–224.
17. O'Connell JR, Weeks DE. The algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. Nature Genetics 1995;11:402–408.
18. Skrivanek Z, Lin S, Irwin ME. Linkage analysis with sequential imputation. Genetic Epidemiology 2003;25:25–35.
19. Thomas A, Gutin A, Abkevich V. Multilocus linkage analysis by blocked Gibbs sampling. Statistics and Computing 2000;10:259–269.
20. Thompson EA. Ancestral inference II: The founders of Tristan da Cunha. Annals of Human Genetics 1978;42:239–253.
21. Thompson EA. *Statistical Inferences from Genetic Data on Pedigrees*, vol. 6 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, Beachwood, OH, 2000.
22. Thompson EA. Monte Carlo methods on Genetic Structures. In *Complex Stochastic Systems* (edited by OE Barndorff-Nielsen, DR Cox, C Klüppelberg), Séminaire Européen de Statistique. Chapman and Hall, London, UK, 2000; pp. 179–222.
23. Thompson EA. Chapter 30: Linkage analysis. In *Handbook of Statistical Genetics, 2 nd ed.*. Wiley, Chichester, UK, 2003; pp. 893–918.
24. Thompson EA. Information from data on pedigree structures. In *Science of Modeling: Proceedings of AIC 2003*. Research Memorandum of the Institute of Statistical Mathematics, Tokyo, Japan, 2003; pp. 307–316.
25. Thompson EA, Guo SW. Evaluation of likelihood ratios for complex genetic models. I.M.A. Journal of Mathematics Applied in Medicine and Biology 1991;8:149–169.
26. Thompson EA, Heath SC. Estimation of conditional multilocus gene identity among relatives. In *Statistics in Molecular Biology and Genetics: Selected Proceedings of a 1997 Joint AMS-IMS-SIAM Summer Conference on Statistics in Molecular Biology* (edited by F Seillier-Moiseiwitsch), IMS Lecture Note–Monograph Series Volume 33. Institute of Mathematical Statistics, Hayward, CA, 1999; pp. 95–113.
27. Tierney L. Markov chains for exploring posterior distributions. Annals of Statistics 1994;22:1701–1728.