International Statistical Review (2002), **70**, 2, 241–267, Printed in The Netherlands

Bayesian MCMC Mapping of Quantitative Trait Loci in a Half-sib Design: a Graphical Model Perspective

N.A. Sheehan[1]*, B. Gulbrandtsen[2], M.S. Lund[2] and D.A. Sorensen[2]

[1]*Department of Epidemiology and Public Health, University of Leicester, Leics LE1 6TP, UK*
[2]*Department of Animal Breeding and Genetics, Danish Institute of Agricultural Sciences, PB50, DK-8820 Tjele, Denmark*

Summary

Graphical models provide a powerful and flexible approach to the analysis of complex problems in genetics. While task-specific software may be extremely efficient for any particular analysis, it is often difficult to adapt to new computational challenges. By viewing these genetic applications in a more general framework, many problems can be handled by essentially the same software. This is advantageous in an area where fast methodological development is essential. Once a method has been fully developed and tested, problem-specific software may then be required. The aim of this paper is to illustrate the potential use of a graphical model approach to genetic analyses by taking a very simple and well-understood problem by way of example.

Key words: Pedigree analysis; Peeling; Conditional distribution; Bayesian networks; Graphical models; Markov chain Monte Carlo; Mixing: Block updating.

1 Introduction

Probability and likelihood computations, relevant to applications in several areas such as genetic counselling, selective animal breeding, inference on the genetic nature of a disease, analysis of surviving genes in an endangered species and linkage analysis, are essential in any analysis of genetic data on groups of related individuals or pedigrees. An exact method for computing probabilities on pedigrees in which at least one of every parent pair is a founder was proposed by Elston & Stewart (1971), extended by Lange & Elston (1975) and finally generalised by Cannings, Thompson & Skolnick (1978) to include arbitrarily complex pedigrees and genetic models. This method has become known in the statistical genetics literature as *peeling* and is essentially the same method as is described ten years later in the expert systems literature (Lauritzen & Spiegelhalter, 1988) for the calculation of posterior probabilities on general Bayesian networks. Because of the enormous storage requirements of the method, peeling fails in practice either when the pedigree has too many interconnecting *loops* which are typically caused by inbreeding relationships or multiple inter-marital relationships, or when the genetic model under consideration is too complex. In all of these genetic applications, particularly that of linkage analysis, the computational problems are intensifying due to the ever-increasing number of polymorphic markers available (Sobel & Lange, 1996) and the relative ease with which individuals can now be genotyped. In particular, exact methods are completely in-

*Corresponding Author

tractable on the large complex pedigrees which frequently arise in animal populations. Consequently, pedigree information is either discarded altogether and data collected on simple designs extracted from a much larger pedigree, or the structure itself is approximated by cutting loops to facilitate computation (Wang, Fernando, Stricker & Elston, 1996). Alternatively, Markov chain Monte Carlo (MCMC) methods (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth & Teller, 1953) can be employed to estimate probabilities and likelihoods of interest. (For an overview of these applications, see Thompson, 2001). However, MCMC methods have not really been tested extensively on these large problems and tend to be viewed with some suspicion in practice, due to the unreliability of the resulting estimates (Hoeschele, Uimari, Grignola, Zhang & Gage, 1997).

Here, we consider the problem of detecting a quantitative trait locus (QTL) from possibly incomplete marker data on individuals related via a half-sib design. As we will show, MCMC methods are required even for this very simple scenario. Our approach to this problem involves the use of graphical models and we will argue that graphical models provide the ideal framework for the development and testing of different MCMC sampling schemes which is crucial to real progress in this area. The natural modularity inherent in these genetic applications makes them ideal for a graphical model representation. Yet, although graphical models feature explicitly in several specific applications (Kong, 1991; Jensen & Kong, 1999; Lund & Jensen, 1999), the general applicability of this approach to solving complex problems in genetics has not been widely appreciated. The use of graphs in genetics dates back to the path analysis diagrams of Wright (1934). Indeed, a standard representation of a pedigree such as the marriage node graph representation of Figure 1 (Section 2) is itself a graphical model representing the qualitative aspects of Mendelian inheritance by which an individual's genetic properties depend only on the genes of his parents (Spiegelhalter, 1998). However, with a more general graphical model approach, this idea is pushed a little further in the reducing of a complex problem down to its basic components thereby fully exploiting *all* the conditional independence structures of the problem at hand for performing calculations at the most local level possible. The result is a highly flexible modelling environment which can be more readily adapted to changes in the problem than purpose-designed software. Although our interest is primarily in their potential usefulness in tackling general pedigrees, we will try to dispel some of the suspicion surrounding graphical models by explaining what they are and by demonstrating their relevance to computational problems in genetics with a simple application to QTL mapping on a half-sib design.

2 Genetics and Genetic Mapping

A pedigree is defined to be a set of individuals with a fully specified set of inter-relationships amongst them. Pairs of pedigree members are called *spouses* only if they have common offspring in the pedigree and every such spouse pairing defines a *marriage*. The *founders* of the pedigree are those individuals without parents and are either members of some baseline generation back to which ancestry has been traced or individuals who have married into the pedigree in subsequent generations. By definition, founders are assumed to be unrelated. Although we will focus on a very simple pedigree in this paper, the large highly looped animal pedigrees with which we are concerned pose enormous computational problems for probability and likelihood calculations. This will be discussed further in Section 4.

For a diploid individual, the DNA in each normal cell forms *homologous* pairs of long strings or chromosomes, one of each pair deriving from the DNA of his mother and the other from the DNA of his father. A segment of chromosome coding for a functional protein is known as a *locus* and we refer to the DNA at this locus as a *gene*. Different forms of the DNA at the locus are called *alleles* and the unordered pair of alleles, one on each chromosome, at a given locus is called the *genotype*. The observable characteristic is the *phenotype* (e.g. affected/normal, blood group etc.) where we note

that this term refers to any data, even when observed individuals have been typed. The underlying stochastic process whereby an individual passes a copy of one of his two genes at a locus with probability $\frac{1}{2}$ to each offspring is called *Mendelian segregation* and is a well-accepted assumption for many traits. However, segregations of genes at loci on the same chromosome may be correlated if the loci are close together on the chromosome, or *linked*. As the pair of sex chromosomes behaves a little differently from the others—the *autosomes*—we will restrict our attention to autosomal traits here.

During gamete formation in a process called *crossing over*, the maternal and paternal chromosomes in any homologous pair exchange segments of genetic material so that the chromosome inherited by an offspring from a parent is a mixture of DNA segments from the grandparental chromosomes. Crossovers are less likely to occur between loci which are physically close on the chromosome and the two alleles inherited by the offspring at these loci from a single parent will thus tend to have the same grandparental origins, or be *in phase*. The *genetic map distance* between two loci is defined as the expected number of crossovers to occur between them in a gamete. It is measured in centiMorgans where a Morgan is the unit in which one crossover is expected to occur. An odd number of crossovers results in a *recombination* when the alleles at these loci are out of phase. There are various mapping functions (Ott, 1999) relating genetic map distance to the probability of observing a recombination. The one we will use in this paper is due to Haldane (1919) which assumes that there is no genetic interference and hence that crossovers in non-overlapping intervals occur independently. Under this model, the relationship between the genetic distance d between any two loci and the corresponding recombination fraction r is given by

$$r = \frac{1}{2}(1 - e^{-2d})$$

with inverse function

$$d = -\frac{1}{2}log(1 - 2r).$$

From their definition as expectations, one advantage of map distances is that they are additive, whereas recombination fractions are not, so they are sometimes more convenient to work with, especially when multiple loci are involved.

Quantitative traits, such as height, weight etc., exhibit variation without natural discontinuities. This is a consequence of the simultaneous segregation of many genes (*polygenic* variation) superimposed by some truly non-genetic continuous variation (Falconer & Mackay, 1996). A QTL can be thought of as a segment of chromosome affecting a quantitative trait and is essentially a "gene" with an effect on the trait of interest which, although sizeable, is not large enough to cause an observable discontinuity and hence cannot be detected using Mendelian methods. A *marker* locus is usually a known position on the chromosome characterized by a specific DNA sequence or observable variations in the sequence and which has no effect on the trait under study. Ideally, in order to be useful for linkage detection, it should be highly variable or *polymorphic* so that non-relatives tend to have different alleles. In principle, identification of QTLs by linkage with marker loci involves scoring individuals for their genotypes at the marker loci and phenotype for the quantitative trait under study. If there were a QTL coding for the trait between any pair of marker loci, differences in mean records for the continuous trait among the various classes of marker genotype should be evident. Genetic linkage calculations become more intensive with the consideration of several markers jointly and when marker data are incomplete.

2.1 *Detecting QTLs by Linkage with Marker Loci*

Several methods have been proposed for QTL detection (Hoeschele *et al.*, 1997) including models for crosses of inbred lines, crosses of outbred lines, outbred populations, with diallelic QTL, multiallelic QTL, and nonparametric methods. Here we will briefly outline the main classes of methods for parametric models in which a putative diallelic QTL is segregating in an outbred population.

The simplest methods are based on the least squares principle (see Haley, Knott & Elsen, 1994) for example). These methods have been heavily used because of their computational simplicity. Because of the speed with which a single analysis can be performed, permutation tests requiring the analysis of a large number of permuted datasets can be carried out (Churchill & Doerge, 1994; Good, 1994) for the calculation of significance thresholds of the relevant test statistics. However, the method only applies to specific and simple designs such as half-sibships or full-sib pairs. It does not utilise all the information in the distribution of the data, it does not allow for the inclusion of random polygenic effects and it does not enable estimation of any QTL parameters other than QTL position.

The second group of approaches utilizes the principle of maximum likelihood in which phenotypes are modelled as a mixture of normal distributions pertaining to each QTL genotype. Initially, methods were developed to analyse line crosses using several linked markers (Lander & Botstein, 1989; Jansen & Stam, 1994; Jansen, 1996). For outbred populations, methods were developed for full-sib groups (Knott & Haley, 1992), for half-sib groups (MacKinnon & Weller, 1995) and for more general pedigrees (Guo & Thompson, 1992; Jansen, Johnson & Van Arendonk, 1998). In general, for complex pedigrees with genetic models involving several loci, exact likelihood computations are infeasible and one must resort to Monte Carlo based approaches. Guo & Thompson (1992) and Jansen *et al.* (1998) applied a Monte Carlo EM algorithm using the Gibbs sampler to obtain draws from the necessary conditional distributions, given the data and the current values of the parameters, in order to compute the conditional expectations which are part of the EM equations. The likelihood approach in principle provides a general framework for fitting a variety of models. The drawback, compared to least squares, is the higher computational demand, especially when used in conjunction with Monte Carlo methods.

The third group comprises Bayesian methods; inferences are based on the marginal posterior distribution of the relevant parameters or of functions of these. The required integration leading to the marginal distribution of interest is often achieved via Markov chain Monte Carlo. Earlier contributions assumed models with a single marker and a diallelic QTL applied to simple pedigree structures (Thaller & Hoeschele, 1996a,1996b). Uimari, Thaller & Hoeschele (1996) and George, Mengersen & Davis (2000) extended the method to deal with multiple markers. In the study of George *et al.* (2000), the ordering of the QTL along the marker map was achieved using the reversible jump MCMC sampler of Green (1995). More recently, a number of important contributions to the literature use various sampling strategies to improve the behaviour of the Markov chain, and assume different levels of complexity in the pedigree structure. These make use of the flexibility offered by the Bayesian approach which allows for treating the number of putative QTL as an unknown random variable (Heath, 1997; Uimari & Hoeschele, 1997, Sillanpää & Arjas, 1998, 1999; Stephens & Fisch, 1998; Lee & Thomas, 2000; Yi & Xu, 2000; Uimari & Sillanpää, 2001).

One of the most challenging aspects of MCMC-based linkage analysis is the updating of genotypes of individuals, especially, in large, complex pedigrees. Earlier approaches used local updating schemes where each locus for each individual was sampled conditionally on all other parameters and individuals in the pedigree (Thomas & Cortessis, 1992; Guo & Thompson, 1992). It soon became evident that this approach results in very poor mixing of the Markov chain (Jensen & Sheehan, 1998; Sheehan, 2000). A variety of block-updating schemes have been proposed as an attempt to solving this problem. In some, each locus across all individuals is sampled jointly (Heath, 1997; Sillanpää & Arjas, 1999), whereas in others, all loci for each individual in the pedigree are sampled in one pass as in the meiosis-by-meiosis sampler of Thompson & Heath (2000). There are also combination

samplers which alternate between blocking individuals at a single locus and blocking loci within an individual (Thompson, 2000; Hurme *et al.*, 2000; Thomas *et al.*, 2000). While the joint updating schemes generally lead to a very significant improvement in the behaviour of MCMC algorithms, multiple tightly linked marker loci can still result in poor mixing. Therefore, a more general blocking structure in which several loci are updated jointly for several individuals is needed. As will be demonstrated in Section 6, graphical models have enormous potential for devising such samplers. One such blocking sampler is that of Jensen, Kjærulff & Kong (1995) (and Jensen (1997)) which has been successfully applied to a linkage analysis with one marker locus and one disease locus (Jensen & Kong, 1999) and to a complex segregation analysis for a quantitative trait (Lund & Jensen, 1999).

2.2 *A Simple QTL Mapping Problem on a Half-sib Design*

In animal populations, data are frequently collected on simple designs extracted from a much larger pedigree structure. The half-sib design of Figure 1 is one of the simplest and will be the focus of our discussion for the rest of this paper. Consider the QTL detection problem where the trait of interest is milk yield in dairy cows, for example. In a typical half-sib design, 10 to 15 bulls are chosen (the sires) each of which has about 50 daughters. All animals are (ideally) typed at the marker loci and the daughters, in addition, have a phenotypic record for the trait (i.e. milk yield) giving information on the putative QTL. All information on the dams of the daughters is ignored and in the absence of such knowledge, they are all assumed to be different and unrelated to the sire as befitting founders of this simple pedigree. Furthermore, the population from which sires were sampled is assumed to be in Hardy–Weinberg and linkage equilibrium and in particular, sires are assumed to be an unselected sample.

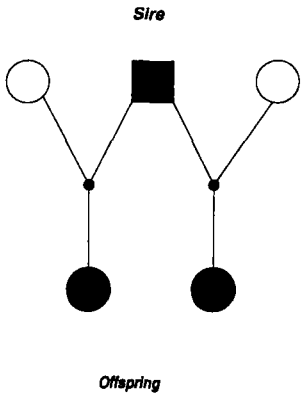


Figure 1. *A half-sib design depicted as a marriage node graph where individuals and their marriages are represented as nodes with squares for males, circles for females and dots for marriages. Here we have a sire with 2 offspring assumed to have distinct and unrelated dams.*

We will begin with the most rudimentary model for this mapping problem and assume that there is a single QTL coding for our trait and we are only concerned with determining whether or not this QTL is to be found between two known markers. A model with a diallelic QTL is assumed with alleles Q and q and allele frequencies p_Q and $1 - p_Q$. In this simple case, we will also hold that the marker loci are diallelic with alleles M, m and N, n, respectively and corresponding frequencies $p_M, 1 - p_M, p_N$ and $1 - p_N$. In this application, it is assumed that the map distance between the two markers is known and so the probability of recombination between the two markers, r_{MN}, is given. The QTL position is not known so the recombination fractions, r_{MQ} and r_{QN} between each

marker and the QTL as shown in Figure 2, have to be estimated. Under the assumption of no genetic interference, recombinations in non-overlapping intervals are independent and the probability of a double recombinant is $r_{MQ}r_{QN}$. Hence, we can parameterise the problem in terms of either r_{MQ} or r_{QN}.

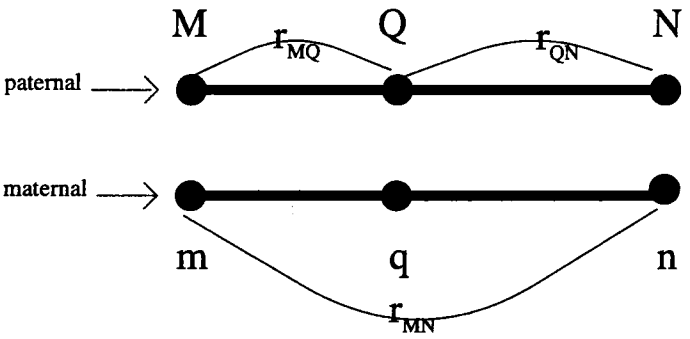


Figure 2. *The simple QTL-mapping problem with two flanking (diallelic) markers and one QTL.*

We begin with an explicit derivation of the joint and associated fully conditional distributions of interest for a Bayesian MCMC analysis of this simple design (Section 3). Although these must be standard, the details are not altogether trivial and are not easily found in the literature. Any changes to either the model or the design, of course, would require a complete reformulation of these distributions. As we will discuss in Section 5, a graphical model representation for the same analysis can accommodate such modifications far more easily.

3 Derivation of a Bayesian MCMC Implementation

We begin with a specification of notation and the relevant prior distributions for fitting a normal linear mixed model (Gelfand *et al.*, 1990). These will be consistent throughout the paper. Then we write out the joint posterior distribution from which the fully conditional distributions for the parameters of interest are derived.

3.1 Specifying the Joint Posterior Distribution

The phenotypic record on offspring j of sire i will be denoted by y_{ij} where $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n_i$, for the quantitative trait under study. The full record vector is $\mathbf{y}$ and is of dimension $n = n_1 + n_2 + \ldots + n_k$.

Let $\mathcal{M} = \{\mathcal{M}_s, \mathcal{M}_o\}$ denote the (known) marker information for the pair of flanking markers at the M and N loci associated with all the typed individuals. In particular, $\mathcal{M}_s = \{\mathcal{M}_i\}$ for $i = 1, \ldots, k$ gives the marker data for all the sires and $\mathcal{M}_o = \{\mathcal{M}_{ij}\}$ for $i = 1, \ldots, k$, $j = 1, \ldots, n_i$ is the vector of marker data for all n offspring. For notational simplicity we will assume at this stage that the marker allele frequencies are known. As we are only considering the case where both marker loci are diallelic here, it suffices to list the frequencies for the M and N alleles, p_M and p_N, respectively. Unknown allele frequencies are easily dealt with, as is shown in Section 5.

Analogously, we let $\mathcal{Q} = \{\mathcal{Q}_s, \mathcal{Q}_o\}$ be a random variable whose realization defines a particular configuration of the vector of (unobserved) QTL genotypes of all sires ($\mathcal{Q}_s = \{\mathrm{Q}_i\}$), and of all offspring ($\mathcal{Q}_o = \{\mathrm{Q}_{ij}\}$) where in this case, Q_i and Q_{ij} refer to a single QTL genotype and can take any of the four distinct values $\{QQ, Qq, qQ, qq\}$. When the distinction is required, the leftmost

allele of the genotype is of paternal origin, and the allele on the right is of maternal origin. The frequency of the allele Q, p_Q, is unknown and derives from a prior distribution taken to be Beta with known parameters, a and b:

$$\pi(p_Q) \sim Beta(p_Q|a, b).$$

The known marker map positions for the M and N loci are represented by λ_M and λ_N, respectively, while the unknown QTL location will be denoted by λ_Q where we will assume the ordering $\lambda_M < \lambda_Q < \lambda_N$. The map distance between the two markers is $d_{MN} = \lambda_N - \lambda_M$. As a prior distribution on $d_{MQ} = \lambda_Q - \lambda_M$, the map distance between the QTL and the first marker, we assume a uniform distribution over the interval $(0, d_{MN})$ which is equivalent to putting a uniform prior on the QTL location λ_Q over the interval $(\lambda_N - \lambda_M)$.

The *"fixed"* effects in the model describe the effect of each (unordered) QTL genotype on the data. We represent these by the vector $\mu = (\mu_1, \mu_2, \mu_3)'$ and will use

$$\pi(\mu) \propto \textit{constant}$$

as an improper prior distribution on μ. The *random* effects, or *sire* effects, are given by $\mathbf{s} = (s_1, \ldots, s_k)'$ where s_i denotes the average additive genetic effect of the i^{th} sire on the phenotypes of his daughters and which cannot be explained by the QTL. It is assumed that these are normally distributed with common variance. In particular,

$$\mathbf{s}|\sigma_s^2 \sim N(\mathbf{0}, \mathbf{I}\sigma_s^2)$$

where $\sigma_s^2 \in \mathbb{R}_+$ is the sire variance component. As a prior distribution on σ_s^2 we assume a scaled inverted chi-square distribution with ν_s degrees of freedom and scale factor S_s, both known:

$$\pi(\sigma_s^2) \quad \propto \quad (\sigma_s^2)^{-(\frac{\nu_s}{2}+1)} exp(\frac{-\nu_s S_s}{2\sigma_s^2}) \tag{1}$$

Note: If we let σ_u^2 denote the polygenic variance i.e. the total additive genetic variance unexplained by the QTL, we have that $\sigma_s^2 = \frac{1}{4}\sigma_u^2$ since half the genes of an offspring are shared with its sire.

Let $\sigma_{res}^2 \in \mathbb{R}_+$ be the residual variance component which models all the variation in the data that cannot be explained by the sire effect or the QTL. In terms of the above, $\sigma_{res}^2 = \frac{3}{4}\sigma_u^2 + \sigma_e^2$ where σ_e^2 is the environmental variance. From the restrictions imposed by the model, we have that

$$0 \leq h^2 = \frac{4\sigma_s^2}{\sigma_s^2 + \sigma_{res}^2} \leq 1$$

and hence

$$\sigma_s^2 \leq \frac{\sigma_{res}^2}{3} \tag{2}$$

where h^2 is the polygenic heritability with $h^2 \in [0, 1]$. Note that σ_{res}^2 is also assumed to have a scaled inverted chi-square prior distribution with known degrees of freedom ν_{res} and scale factor S_{res}.

Denote the vector of unknown parameters by $\theta = (\mu', \mathbf{s}', \sigma_s^2, \sigma_{res}^2)'$. The assumed sampling model for the data is then:

$$\begin{aligned}
y_{ij}|(\mathrm{Q}_{ij} = QQ, \theta) &\sim N(\mu_1 + s_i, \sigma_{res}^2) \\
y_{ij}|(\mathrm{Q}_{ij} = Qq, \theta) &\sim N(\mu_2 + s_i, \sigma_{res}^2) \\
y_{ij}|(\mathrm{Q}_{ij} = qQ, \theta) &\sim N(\mu_2 + s_i, \sigma_{res}^2) \\
y_{ij}|(\mathrm{Q}_{ij} = qq, \theta) &\sim N(\mu_3 + s_i, \sigma_{res}^2)
\end{aligned}$$

or

$$\mathbf{y}|\mathcal{Q}, \boldsymbol{\theta} \sim N\left(\mathbf{X}\boldsymbol{\mu} + \mathbf{Z}\mathbf{s}, \mathbf{I}\sigma^2_{res}\right) \tag{3}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are known incidence matrices associating $\boldsymbol{\mu}$ and $\mathbf{s}$, respectively, with the data. Equivalently, we can write

$$y_{ij} = s_i + q_{ij} + e_{ij} \tag{4}$$

where

$$e_{ij} \sim N(0, \sigma^2_{res})$$

and

$$q_{ij} = \begin{cases} \mu_1 & \text{if } Q_{ij} = QQ \\ \mu_2 & \text{if } Q_{ij} = Qq \text{ or } qQ \\ \mu_3 & \text{if } Q_{ij} = qq \end{cases} .$$

Note that the expression in (3) implies that

$$Cov(s_i, e_{ij}) = 0 \; \forall \; i, j.$$

Furthermore, given the QTL genotypes for the offspring, $\mathcal{Q}_o$, and the sire effects, $\mathbf{s}$, it is assumed that phenotypic records are conditionally independent.

For convenience, all conditioning on the known marker allele frequencies, p_M and p_N, known map distance between the markers, d_{MN} and known hyper parameters of prior distributions will be suppressed from the notation. The joint posterior distribution is:

$$\begin{aligned} f(p_Q, \mathcal{Q}, \lambda_Q, \boldsymbol{\theta}|\boldsymbol{y}, \mathcal{M}) &\propto f(\mathbf{y}, \mathcal{M}, p_Q, \mathcal{Q}, \lambda_Q, \boldsymbol{\theta}) \\ &\propto f(\mathbf{y}|\mathcal{Q}, \boldsymbol{\theta}) \Pr(\mathcal{Q}|\mathcal{M}, \lambda_Q, p_Q) \pi(\boldsymbol{\theta}) \pi(\lambda_Q) \pi(p_Q) \end{aligned} \tag{5}$$

where $\pi(\boldsymbol{\theta})$, $\pi(\lambda_Q)$ and $\pi(p_Q)$ represent the prior distributions for $\boldsymbol{\theta}$, λ_Q and p_Q, respectively. Given the model assumptions, note that the prior distribution for $\boldsymbol{\theta}$ can be factorised as:

$$\pi(\boldsymbol{\theta}) \propto f(\mathbf{s}|\sigma^2_s) \pi(\sigma^2_s) \pi(\sigma^2_{res}) \pi(\boldsymbol{\mu}). \tag{6}$$

The fully conditional posterior distributions for the unknown parameters $\mathcal{Q}$, λ_Q, $\boldsymbol{\mu}$, $\mathbf{s}$ and p_Q, given the phenotypic records for the quantitative trait, are all derived from the joint posterior (5) and will be written in the form

$$f(\cdot|., data) \text{ or } Pr(\cdot|., data).$$

These will be detailed below in the remainder of this section.

In a full likelihood approach, the parameter (row) vector of interest would be $\boldsymbol{\beta} = (\boldsymbol{\mu}', \lambda_Q, \sigma^2_{res}, \sigma^2_s, p_Q)'$ and the analysis would involve joint maximization of the following function with respect to $\boldsymbol{\beta}$:

$$\begin{aligned} L(\boldsymbol{\beta}|y) &\propto \int_{\Re^t} \sum_{\mathcal{Q}} \Pr\left(\mathcal{Q}|\mathcal{M}, \lambda_Q, p_Q\right) f(\mathbf{y}|\mathcal{Q}, \boldsymbol{\theta}) f\left(\mathbf{s}|\sigma^2_s\right) d\mathbf{s} \\ &= \sum_{\mathcal{Q}} \Pr\left(\mathcal{Q}|\mathcal{M}, \lambda_Q, p_Q\right) f\left(\mathbf{y}|\mathcal{Q}, \boldsymbol{\mu}, \sigma^2_s, \sigma^2_{res}\right) . \end{aligned} \tag{7}$$

The required summation over all QTL genotypes $\mathcal{Q}$ which implicitly involves summation over all possible phases in the sires (see Section 3.2 below) is an enormous computational undertaking for any reasonably sized half-sib design (Section 2.2). Of course, for more complicated designs and particularly for general pedigrees, the likelihood (7) will not have a closed analytical form and

sampling for all quantities of interest will be essential. Even on simple designs, if direct estimation of the sire polygenic effects is of interest, maximum likelihood methods do not apply and MCMC methods must be employed. For this, all sire families need to be considered jointly and the inclusion of these within- and between-family polygenic effects adds greatly to the computational complexity of the problem (Section 5). In the maximum likelihood approach of Georges *et al.* (1995), such computational problems have been largely circumvented by the treatment of each sire and his offspring in a separate analysis but neither sire effects nor the sire variance component can be inferred in this way.

3.2 *The Fully Conditional Posterior Distribution for* $\mathcal{Q}$

It is convenient at this point to augment with the random variable F_i representing the phase of the i^{th} sire. Information on phase defines the haplotype which is a specification of the alleles that each gamete of the sire carries for the markers and the QTL. From (5), the fully conditional posterior distribution of $\mathcal{Q}$ is given by:

$$\Pr(\mathcal{Q}|., data) \propto f(\mathbf{y}|\mathbf{Q}_o, \boldsymbol{\theta}) \Pr(\mathbf{Q}_s, \mathbf{Q}_o|\mathcal{M}, \lambda_Q, p_Q). \tag{8}$$

From the assumption that offspring phenotype records are conditionally independent given offspring genotypes and sire effects, the first term factorises as follows:

$$f(\mathbf{y}|\mathbf{Q}_o, \boldsymbol{\theta}) = \prod_{i=1}^{k} \prod_{j=1}^{n_i} f(y_{ij}|\mathbf{Q}_{ij}, \boldsymbol{\theta}).$$

For the half-sib design under consideration here, the second term can be written as:

$$\Pr(\mathbf{Q}_s, \mathbf{Q}_o|\mathcal{M}, \lambda_Q, p_Q) = \prod_{i=1}^{k} \Pr[\mathbf{Q}_i|p_Q] \sum_{F_i} \Pr[F_i|\mathcal{M}] \prod_{j=1}^{n_i} \Pr[\mathbf{Q}_{ij}|\mathbf{Q}_i, \mathcal{M}, \lambda_Q, p_Q, F_i].$$

However, by straightforward manipulation,

$$\begin{aligned}
\Pr[F_i|\mathcal{M}_i, \mathcal{M}_{i1}, \ldots, \mathcal{M}_{in_i}] &= \frac{\Pr[\mathcal{M}_i, \mathcal{M}_{i1}, \ldots, \mathcal{M}_{in_i}|F_i] \Pr[F_i]}{\sum_{F_i} \Pr[\mathcal{M}_i, \mathcal{M}_{i1}, \ldots, \mathcal{M}_{in_i}|F_i] \Pr[F_i]} \\
&= \frac{\Pr[\mathcal{M}_i] \Pr[F_i] \prod_{j=1}^{n_i} \Pr[\mathcal{M}_{ij}|\mathcal{M}_i, F_i]}{\sum_{F_i} \Pr[\mathcal{M}_i, \mathcal{M}_{i1}, \ldots, \mathcal{M}_{in_i}|F_i] \Pr[F_i]} \\
&\propto \prod_{j=1}^{n_i} \Pr[F_i|\mathcal{M}_i, \mathcal{M}_{ij}]
\end{aligned} \tag{9}$$

although we note that, from a sampling point of view, it is more useful to write

$$Pr(F_i|\mathcal{M}_i, \mathcal{M}_{ij}) \propto Pr(\mathcal{M}_{ij}|\mathcal{M}_i, F_i)$$

in expression (9). Substituting back into (8) yields:

$$\begin{aligned}
\Pr(\mathcal{Q}|., data) \propto \quad & \prod_{i=1}^{k} \Pr[\mathbf{Q}_i|p_Q] \sum_{F_i} \prod_{j=1}^{n_i} \Pr[F_i|\mathcal{M}_i, \mathcal{M}_{ij}] \times \\
& f(y_{ij}|\mathbf{Q}_{ij}, \boldsymbol{\theta}) \Pr[\mathbf{Q}_{ij}|\mathbf{Q}_i, \mathcal{M}_i, \mathcal{M}_{ij}, \lambda_Q, p_Q, F_i].
\end{aligned} \tag{10}$$

Again, a simple manipulation of the last term shows that

$$\Pr[\mathbf{Q}_{ij}|\mathbf{Q}_i, \mathcal{M}_i, \mathcal{M}_{ij}, \lambda_Q, p_Q, F_i] = \frac{\Pr[\mathbf{Q}_{ij}, \mathcal{M}_{ij}|\mathbf{Q}_i, \mathcal{M}_i, \lambda_Q, p_Q, F_i]}{\Pr[\mathcal{M}_{ij}|\mathbf{Q}_i, \mathcal{M}_i, \lambda_Q, p_Q, F_i]}. \tag{11}$$

3.2.1 Sampling the QTL genotypes

In the case of the half-sib design, it is possible to sample all QTL genotypes simultaneously (Janss, Thompson & Van Arendonk, 1995). This can be accomplished by drawing first from $[\mathbf{Q}_i|., data]$ and then from $[\mathbf{Q}_{i1}, \mathbf{Q}_{i2}, \ldots, \mathbf{Q}_{in_i}|\mathbf{Q}_i, ., data]$. Since families (sires and their offspring) are independent, a draw from $\Pr(\mathcal{Q}|., data)$ involves calculating:

$$\Pr(\mathcal{Q}|., data) = \prod_{i=1}^{k} \Pr\left(\mathbf{Q}_{i1}, \mathbf{Q}_{i2}, \ldots, \mathbf{Q}_{in_i}|\mathbf{Q}_i, ., data\right) Pr(\mathbf{Q}_i|., data) .$$

The first term in the above expression requires n_i computations from (11). Using (10), a draw from $[\mathbf{Q}_i|., data]$ for sire i involves computing:

$$\Pr\left(\mathbf{Q}_i|., data\right) = \sum_{\mathbf{Q}_{i1}} \sum_{\mathbf{Q}_{i2}} \cdots \sum_{\mathbf{Q}_{in_i}} \Pr\left(\mathbf{Q}_{i1}, \mathbf{Q}_{i2}, \ldots, \mathbf{Q}_{in_i}, \mathbf{Q}_i|., data\right)$$

$$\propto \Pr\left(\mathbf{Q}_i|p_Q\right) \sum_{F_i} \prod_{j=1}^{n_i} \Pr\left[F_i|\mathcal{M}_i, \mathcal{M}_{ij}\right] \sum_{\mathbf{Q}_{ij}} f\left(y_{ij}|\mathbf{Q}_{ij}, \theta\right) \Pr\left[\mathbf{Q}_{ij}|\mathbf{Q}_i, \mathcal{M}_i, \mathcal{M}_{ij}, \lambda_Q, p_Q, F_i\right] .$$

This is an example of what is meant by *joint* or *block* sampling.

3.3 The Fully Conditional Distribution for λ_Q

From (5), we have:

$$f\left(\lambda_Q|., data\right) \propto \Pr\left(\mathcal{Q}|\mathcal{M}, \lambda_Q, p_Q\right) \pi\left(\lambda_Q\right) . \tag{12}$$

This expression does not have a standard form and therefore a Metropolis–Hastings step can be used to draw samples from it. Let λ_Q^* denote a candidate value generated from the proposal $u\left(\lambda_Q^*|\lambda_Q\right)$, where λ_Q denotes the previous realization from (12). Then the proposal is accepted with probability $\alpha\left(\lambda_Q^*, \lambda_Q\right)$ given by:

$$\alpha\left(\lambda_Q^*, \lambda_Q\right) = \begin{cases} \min\left[\frac{f(\lambda_Q^*|., data)u(\lambda_Q^*|\lambda_Q)}{f(\lambda_Q|., data)u(\lambda_Q|\lambda_Q^*)}, 1\right], & \text{if } f\left(\lambda_Q|., data\right) > 0. \\ 1, \text{ otherwise} \end{cases} \tag{13}$$

The candidate generation density could be a uniform distribution on the interval $\left(Max\{\lambda_M, \lambda_Q - h\}, Min\{\lambda_N, \lambda_Q + h\}\right)$, where h is chosen such that the acceptance rate is in the range 20% to 50%.

3.4 The Fully Conditional for μ and s

From (5), the fully conditional posterior distribution of $\begin{pmatrix} \mu \\ \mathbf{s} \end{pmatrix}$ is:

$$\begin{aligned} f\left(\begin{pmatrix} \mu \\ \mathbf{s} \end{pmatrix}|., data\right) &\propto f\left(\mathbf{y}|\mathcal{Q}, \theta\right) \pi\left(\theta\right) \\ &\propto f\left(\mathbf{y}|\mathcal{Q}, \theta\right) f\left(\mathbf{s}|\sigma_s^2\right) \end{aligned} \tag{14}$$

from the factorisation in (6). As given in (3), Section 3.1, the sampling model for the data is:

$$\mathbf{y}|\mathcal{Q}, \mu, \mathbf{s}, \sigma_{res}^2 \sim N\left(\mathbf{X}\mu + \mathbf{Z}\mathbf{s}, \mathbf{I}\sigma_{res}^2\right)$$

where $\mathbf{X}$ and $\mathbf{Z}$ are known. Since the prior distribution of sire effects (Section 3.1) is

$$\mathbf{s}|\sigma_s^2 \sim N\left(\mathbf{0}, \mathbf{I}\sigma_s^2\right),$$

it follows from properties of the Normal distribution, that the form of (14) is given by:

$$\begin{pmatrix} \mu \\ \mathbf{s} \end{pmatrix} |., data \sim N\left(\begin{pmatrix} \widehat{\mu} \\ \widehat{\mathbf{s}} \end{pmatrix}, \mathbf{C}^{-1}\sigma_{res}^2\right)$$

where

$$\mathbf{C} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{I}c \end{bmatrix}$$

and $\widehat{\mu}$ and $\widehat{\mathbf{s}}$ satisfy:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{I}c \end{bmatrix} \begin{bmatrix} \widehat{\mu} \\ \widehat{\mathbf{s}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

with $c = \sigma_{res}^2 / \sigma_s^2$. Again, invoking properties of the normal distribution we can write,

$$\mu|\mathbf{s}, ., data \sim N\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{y} - \mathbf{Z}\mathbf{s}\right), \left(\mathbf{X}'\mathbf{X}\right)^{-1}\sigma_{res}^2\right] \tag{15}$$

and

$$\mathbf{s}|\mu, ., data \sim N\left[\left(\mathbf{Z}'\mathbf{Z} + \mathbf{I}c\right)^{-1}\mathbf{Z}'\left(\mathbf{y} - \mathbf{X}\mu\right), \left(\mathbf{Z}'\mathbf{Z} + \mathbf{I}c\right)^{-1}\sigma_{res}^2\right]. \tag{16}$$

3.5 The Fully Conditional for σ_{res}^2 and σ_s^2

Again from the joint distribution (5) we have that

$$f\left(\sigma_{res}^2|., data\right) \propto f\left(\mathbf{y}|\mathcal{Q}, \theta,\right)\pi(\sigma_{res}^2)$$

where $\pi(\sigma_{res}^2) \sim \nu_{res} S_{res} \chi_{\nu_{res}}^{-2}$ (Section 3.1). From standard properties of the Normal distribution, it follows that this posterior distribution is also proportional to a scaled inverted chi-square distribution with $\tilde{\nu}_{res}$ degrees of freedom and scale parameter $\tilde{S}_{res}$ where $\tilde{\nu}_{res} = n + \nu_{res}$ and

$$\tilde{S}_{res} = \frac{(\mathbf{y} - \mathbf{X}\mu - \mathbf{Z}\mathbf{s})'(\mathbf{y} - \mathbf{X}\mu - \mathbf{Z}\mathbf{s}) + \nu_{res} S_{res}}{n + \nu_{res}}.$$

Similarly, from (5):

$$f\left(\sigma_s^2|., data\right) \propto f\left(\mathbf{s}|\sigma_s^2\right)\pi(\sigma_s^2)$$

from which we obtain:

$$\sigma_s^2|., data \sim \tilde{\nu}_s \tilde{S}_s \chi_{\tilde{\nu}_s}^{-2}$$

where $\tilde{\nu}_s = \nu_s + k$ (k being the number of sires) and

$$\tilde{S}_s = \frac{\mathbf{s}'\mathbf{s} + \nu_s S_s}{k + \nu_s}.$$

3.6 *The Fully Conditional Distribution for p_Q*

For the j^{th} daughter of sire i, let $Q_{(ij,1)}$ denote the allele inherited at the QTL locus from the sire and $Q_{(ij,0)}$ the QTL allele received from the dam. From (5),

$$Pr\left(p_Q|.,data\right) \propto \Pr\left(Q|\mathcal{M},\lambda_Q,p_Q\right)\pi\left(p_Q\right)$$

$$= \pi\left(p_Q\right)\prod_{i=1}^{k} Pr(Q_i|p_Q)\sum_{F_i} Pr(F_i|\mathcal{M})\prod_{j=1}^{n_i} Pr\left(Q_{ij}|Q_i,\mathcal{M}_i,\mathcal{M}_{ij},F_i,\lambda_Q,p_Q\right)$$

$$= \pi(p_Q)\prod_{i=1}^{k} Pr(Q_i|p_Q)\sum_{F_i} Pr(F_i|\mathcal{M})\prod_{j=1}^{n_i} Pr(Q_{(ij,1)}|Q_i,\mathcal{M}_i,\mathcal{M}_{ij},F_i,\lambda_Q)\times$$

$$Pr(Q_{(ij,0)}|p_Q)$$

$$= \pi(p_Q)\prod_{i=1}^{k} Pr(Q_i|p_Q)(\prod_{j=1}^{n_i} Pr(Q_{(ij,0)}|p_Q))\times$$

$$\sum_{F_i} Pr(F_i|\mathcal{M})(\prod_{j=1}^{n_i} Pr(Q_{(ij,1)}|Q_i,\mathcal{M}_i,\mathcal{M}_{ij},F_i,\lambda_Q))$$

$$\propto \pi\left(p_Q\right)\prod_{i=1}^{k}\Pr\left(Q_i|p_Q\right)\prod_{j=1}^{n_i}\Pr\left(Q_{(ij,0)}|p_Q\right). \tag{17}$$

The last line in (17) follows because the term

$$\Pr\left[Q_{(ij,1)}|Q_i,\mathcal{M}_i,\mathcal{M}_{ij},F_i,\lambda_Q\right]$$

does not depend on p_Q. The term $\Pr\left[Q_{(ij,0)}|p_Q\right]$ corresponds to the pmf of a Bernoulli distribution (the allele received from the mother is either Q with probability p_Q, or q with probability $1-p_Q$) and the term $\Pr\left[Q_i|p_Q\right]$ is derived from the assumption of Hardy–Weinberg equilibrium at the QTL locus. Therefore the term

$$\prod_{i=1}^{k}\prod_{j=1}^{n_i}\Pr\left[Q_{(ij,0)}|p_Q\right]\Pr\left[Q_i|p_Q\right]$$

involves the count (over all $n+2k$ gametes) of the number of Q and q alleles. Let n_Q be the number of Q alleles and n_q the number of q alleles for the given QTL genotype configuration with $n_Q+n_q=n+2k$. Then, since $\pi(p_Q)\sim Beta(a,b)$ (Section 3.1), the fully conditional posterior distribution of p_Q is

$$p_Q|.,data \sim Beta\left(p_Q|a+n_Q,b+n_q\right). \tag{18}$$

4 Graphical Models

A probabilistic approach to dealing with uncertainty in expert systems dates back to the 1980s (Pearl, 1988) when it was realised that calculations on seemingly intractable high dimensional problems can be efficiently performed by imposing a set of simplifying conditional independence assumptions which essentially split the problem up into small manageable pieces. This simplification enables the representation of a complex problem in graphical form which in turn informs the development of efficient algorithms both for performing calculations and making inference on model parameters. These models are sometimes called *Bayesian networks* (Jensen, 1996) but we will adopt the terminology of Cowell *et al.* (1999) and use the more general term, *graphical modelling* to refer to methods which exploit local dependencies to express complex relationships for modelling

and computation.

We define a graph to be a set of vertices or *nodes* and a set of *edges* where an edge is an unordered pair of nodes. The nodes represent the variables in the model and the edges represent links between them. Edges can be either *directed* with arrows indicating the direction of the link, or undirected. Directed edges represent probabilistic *influence* or *causal* mechanisms whereas undirected edges refer to correlations between the variables. The terminology traditionally used in this area, and unfortunately for our applications, derives from genetics. For instance, for nodes labelled A and B, we say that A is a *parent* of B, or B is a *child* of A, if there is a directed edge from A to B. In contrast with the biological interpretation of these terms, a node in a graph can have more than two parents, as shown in Figure 3. A *path* is defined to be a sequence of directed edges, each sharing a common node with both preceding and succeeding edges. If there is a path from node A to node B (i.e. we can arrive at B by following arrows from A), we say that A is an *ancestor* of B and B is a *descendant* of A. A path beginning and ending with the same node is a directed *cycle*. Analogously, a *trail* is a sequence of undirected edges and forms an undirected cycle if it begins and ends with the starting node. If all the edges of a graph are directed, it is a directed graph and if it has no directed cycles, it is a *directed acyclic graph* or DAG (Cowell *et al.*, 1999). For a general DAG, where the nodes are variables, $\{v \in V\}$, with some joint probability distribution function, we have that any node, given its parents, is conditionally independent of all nodes which are not descendants. In other words, the form of the joint distribution function is related to the structure of the graph and necessarily takes the form:

$$\prod_{v \in V} f\,(v|pa(v)) \tag{19}$$

where $pa(v)$ denotes the set of parent nodes of the node v. This is known as the "directed local Markov property" (Lauritzen *et al*, 1990).

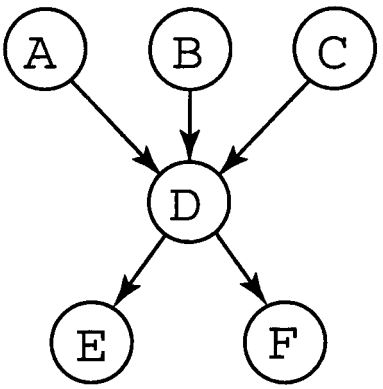


Figure 3. *A simple graphical model with nodes A, B and C all parents of D while E and F are both child nodes of D. Note that if nothing is known about D besides what can be inferred from its parents, then A, B and C are all independent. Conditional dependencies between A, B and C are imposed, however, if information on F, say, influences the certainty of D.*

As detailed by Lange & Elston (1975), a pedigree can be considered as a directed graph with two kinds of node and two kinds of directed edge or *arc*. Individuals and marriages are represented as nodes, and the connecting edges are *marriage arcs*, directed from an individual to his marriages, and *descent arcs*, directed from a marriage to the resulting offspring. As genes are always passed down from parents to offspring, the directions on the edges can be omitted as in the marriage node graph of Figure 1. A path in a pedigree necessarily involves an alternating sequence of marriage and descent arcs, and since an individual cannot be his own ancestor or descendant, a pedigree is a DAG,

by definition. Undirected cycles can be formed in many ways, however, and these are usually called *loops*. An example of such a loop arises when two biologically related individuals marry causing two separate paths of descent from a common ancestor to the node representing their marriage. Other loops include marriage rings, exchange loops, multiple marriage loops and all kinds of overlapping combinations of the above (Cannings *et al.*, 1978). A pedigree without loops is a graph without any cycles and is often called a *tree*.

Algorithms for exact calculation on directed acyclic graphs with a conditional distribution specified for each node, generally involve the following steps:

1. Remove the directions from the existing edges and add further undirected edges between all pairs of parent nodes with a common child node. This is referred to as *moralising* the graph i.e. by "marrying" the parents.
2. The moral graph is now *triangulated* by adding more edges until there are no cycles involving more than three nodes. Finding a good triangulation requires an algorithm to find an optimal ordering for node *elimination*.
3. Once the graph has been triangulated, maximal sets of pairwise connected nodes, or *cliques* can be identified and these cliques are then connected in what is known as a *junction tree*. The goal of any triangulation algorithm is to generate cliques which are as small as possible. However, finding an optimal elimination sequence is known to be NP-hard.

(See Jensen (1997), Jensen & Kong (1999) or Lund & Jensen (1999) for a brief overview and Lauritzen & Spiegelhalter (1988) or Cowell *et al.* (1999) for details on exact methods of computation.) The cliques in the junction tree correspond essentially with the *cutsets* of the peeling algorithms in the statistical genetics environment (Cannings *et al.*, 1978) and finding optimal peeling sequences is of the same order of difficulty as finding an optimal node elimination sequence for graph triangulation. The general problem with exact methods is that their storage requirements are exponential in the size of the largest clique, or cutset, and these tend to get very large when the graph has (undirected) cycles or loops. We have already discussed how loops can occur in pedigrees. There are many ways of forming loops in a general graph (Section 5).

Graphical models lend themselves readily to a Bayesian interpretation where all unknown quantites are regarded as random variables and so data, latent variables and model parameters can all be represented as nodes in the graph with associated distributions. They are most useful for problems such as arise naturally in genetics applications, where extensive conditional independence assumptions allow for communication of structure without recourse to large sets of equations as we had in Section 3 (Spiegelhalter, 1998). This modularity enables direct exploitation of local computational methods and hence easy extension of the model, in principle, to arbitrary levels of complexity.

HUGIN (Andersen *et al.*, 1989) is a commercially available software package for computing probabilities on general Bayesian Networks. For MCMC applications (particularly in a Bayesian framework) the BUGS package (Gilks, Thomas & Spiegelhalter, 1994), which is currently available free of charge, performs Gibbs sampling on graphical models. For these big problems in genetics, MCMC sampling schemes are required and these tend to have slow mixing problems unless some form of block or joint updating is used. (See Jensen *et al.* (1995), for example.) The *random propagate* algorithm (Dawid, 1992) implemented in HUGIN enables efficient sampling of a random configuration from the correct distribution on a graph. For this reason, we have chosen to use HUGIN for joint updating of blocks of variables conditional on the values of all the variables not included in the given block within an MCMC framework. The MCMC code has been written outside HUGIN. The aim is not to produce a rival program to the existing statistical genetics software. Indeed, it is highly probable that some existing program will be more efficient for any specific application. However, by placing these genetics problems in the general framework of graphical modelling, we aim to produce a more flexible modelling environment which allows modification to more complicated problems

without major rewriting of the necessary software.

5 A Graphical Model for the QTL Problem

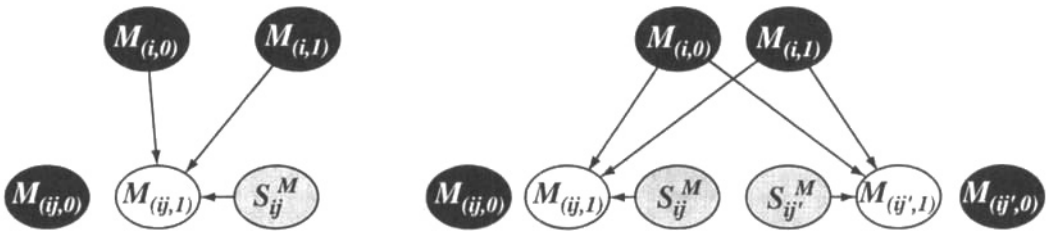
We now formulate the Bayesian analysis of Section 3 in a graphical modelling framework. The general idea of graphical models is demonstrated in the gradual building of the model shown in Figure 8 at the end of this section. As the full graph for the whole design comprises over 200,000 nodes, we will focus on the pedigree of Figure 1 with only one sire and two daughters. For the purposes of illustration by way of keeping the graph relatively uncomplicated, we also omit explicit representation of some of the parameters at this point, although we do include them in our simulation analysis as shown in Figure 9 in Section 6.

5.1 A Single Locus

We begin with the sire and one daughter for the first marker locus with alleles M and m, say. For the sire, we create two nodes representing the allelic states of his maternal and paternal genes at the locus, the values at which are randomly assigned from a Bernoulli distribution parameterised with the appropriate allele frequency, p_M. Under the assumptions of the model (i.e. random union of gametes), these distributions are independent. The parameter p_M is assumed known at this stage and is hence not explicitly represented in the model. However, as will be shown in Section 6, it can easily be estimated by including it as an additional node with an appropriate sampling distribution. Indexing maternally inherited genes by 0 and those paternally inherited by 1, as before (Section 3.6), we define the following nodes and their probability distributions for any particular sire:

$$\begin{aligned} M_{(i,0)} &\sim Ber\,(p_M) \\ M_{(i,1)} &\sim Ber\,(p_M) \end{aligned}$$

where $M_{(i,0)}$ denotes the maternal gene of sire i at the first marker locus and $M_{(i,1)}$ denotes his paternal gene. These are the two black nodes at the top of Figure 4(a).



(a) Sire and one offspring

(b) Sire and two offspring

Figure 4. *A sire and offspring with genes at one locus. Nodes corresponding to genes sampled from the population are shown in black. The genes inherited from the sire are shown in white. The segregation indicator nodes are shown in light grey.*

Segregation of genes from sire i must now be considered for each daughter and it is particularly convenient here to do this using *meiosis* or *segregation* indicators (Thompson, 1994; Sobel & Lange, 1996; Thompson, 2001). These are binary variables taking the values 0 and 1 to represent maternal and paternal inheritance respectively and are defined for each daughter. At this first marker locus (the "M-locus"), the meiosis indicator for the j^{th} daughter of sire i is denoted by S_{ij}^M and represented by the node at the bottom right of Figure 4(a). Its value is randomly assigned from a Bernoulli

distribution with probability $\frac{1}{2}$ since, in the absence of information on other loci, inheritance is assumed to be Mendelian. The gene inherited by the daughter from her sire (her paternal allele) will be a copy of the paternal gene in the sire if the meiosis indicator has a value of 1 and will be a copy of the sire's maternal gene, otherwise. The daughter's paternal gene is denoted by $M_{(ij,1)}$ and is represented by the white node in the bottom centre of Figure 4(a). Since we have no information on the dams, we assume that the maternal gene, $M_{(ij,0)}$, is randomly drawn from the general population with allele frequency p_M and we represent this by the black node on the bottom left of Figure 4(a). Specifically,

$$M_{(ij,0)} \sim Ber\,(p_M)$$

$$M_{(ij,1)} = \begin{cases} M_{(i,0)} & \text{if } S_{ij}^M = 0 \\ M_{(i,1)} & \text{if } S_{ij}^M = 1. \end{cases}$$

To complete the graph in Figure 4(a), we note from above that the nodes $M_{(i,0)}$, $M_{(i,1)}$ and S_{ij}^M are all *parents* of the node $M_{(ij,1)}$ since the value assigned to the daughter's paternal gene depends on all three. Accordingly, arrows are directed from each of these parent nodes to the *child* node.

In Figure 4(b), a second daughter has been added. This involves replicating the nodes and connections described above for the first daughter. In accordance with Mendel's First Law, the variables corresponding to one daughter are statistically independent of the variables corresponding to the other daughter given the state of the variables corresponding to the sire. Hence there are no arrows directly connecting any nodes representing one daughter to any nodes representing the other. Because of this replication from one to several offspring, we will focus on one offspring and continue to build on Figure 4(a).

Note that the model assumptions are also explicitly represented in Figure 4. Independence of the maternal and paternal genes in an individual, due to the assumption of random union of gametes, is indicated by the lack of connecting arrows between them. Similarly, we infer independence of the offspring maternal genes from each other and from the genes of the sire reflecting the assumption that as they are all founders of the pedigree, dams are unrelated both to each other and to the sire.

We complete our modelling of the single-locus scenario by adding nodes representing the genotypes for both individuals with values completely determined by the assigned genes and in this case, directly observable as marker phenotypes. Let M_i denote the genotype at the "M-locus" for sire i and M_{ij} the corresponding genotype for his j^{th} daughter. Nodes representing these values are shown in Figure 5 with the M_i node shown as a *child* of $M_{(i,1)}$ and $M_{(i,0)}$ and similarly M_{ij} as a child of $M_{(ij,1)}$ and $M_{(ij,0)}$ to reflect the necessary dependencies.

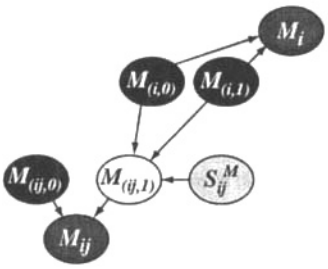


Figure 5. *Figure 4(a) with genotype nodes (in dark grey) added for both individuals. Again, nodes corresponding to genes sampled from the population are black, genes sampled from the sire are white and segregation indicator nodes are light grey.*

5.2 Two or More Linked Loci

For the same two individuals, we now consider the addition of a second locus—the QTL locus—which is linked to the first. Just as for the "M-locus", the sire's maternal and paternal genes at the QTL are represented by two nodes with values independently and randomly assigned from a Bernoulli distribution, this time with parameter p_Q:

$$\begin{aligned} Q_{(i,0)} &\sim Ber\left(p_Q\right) \\ Q_{(i,1)} &\sim Ber\left(p_Q\right). \end{aligned}$$

The parameter, p_Q, is itself a random quantity with an assumed Beta prior distribution (Section 3.1) but we will not represent it as a node at this point to avoid over-cluttering the graph. In other words, the graph of Figure 6 assumes that p_Q is known. Since there is no observation on the sire directly related to the QTL, we do not represent the sire's QTL genotype explicitly in the graph. As in Section 5.1, the daughter's maternal QTL gene is randomly sampled from the population while the paternal gene is inherited from her sire according to the value of the relevant segregation indicator:

$$\begin{aligned} Q_{(ij,0)} &\sim Ber\left(p_Q\right) \\ Q_{(ij,1)} &= \begin{cases} Q_{(i,0)} & \text{if } S_{ij}^{Q} = 0 \\ Q_{(i,1)} & \text{if } S_{ij}^{Q} = 1 \end{cases} \end{aligned}$$

with $Q_{(ij,0)}$ and $Q_{(ij,1)}$ denoting the maternal and paternal genes of the j^{th} daughter of sire i at the QTL, respectively. Arrows from the two QTL genes in the sire and from the QTL segregation indicator are directed to the paternal gene in the daughter, as before. A further node representing the daughter's QTL genotype Q_{ij} is included in the graph, despite the fact that it is unobservable (by definition), since we are interested in a quantitative phenotype with distribution depending on the genotypic state at the QTL. This is a *child* node of her two QTL genes as is indicated by the arrows in Figure 6.

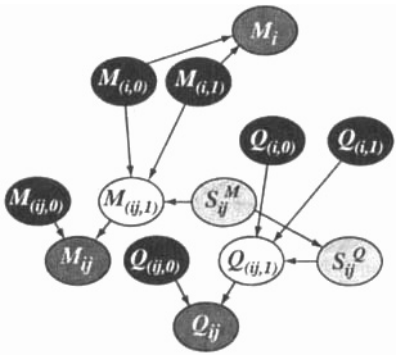


Figure 6. *Two individuals with two linked loci. Nodes corresponding to genes sampled from the population are shown in black. The genes sampled from the sire are shown in white. The segregation indicator nodes are light grey. Genotypes are shown in dark grey.*

However, we now need to take linkage (and implicitly phase in the sire) into account. At this point, it is simpler to think in term of recombination fractions rather than genetic distances. Under the no-interference model, whether or not the sire's maternal or paternal QTL gene is passed to the daughter depends only on what was inherited at the "M locus" and on the recombination fraction r_{MQ} between the two loci. In other words, the daughter's segregation indicator at the QTL, S_{ij}^{Q}, depends on her segregation indicator at the linked locus and on r_{MQ} which we will also assume known, for

now. Estimating this quantity is a trivial extension (Section 6) but leads to a more complicated graph (Figure 9). Specifically, we have

$$S_{ij}^{Q} \sim \begin{cases} Ber(r_{MQ}) & \text{if } S_{ij}^{M} = 0 \\ Ber(1 - r_{MQ}) & \text{if } S_{ij}^{M} = 1. \end{cases}$$

This dependency is represented in Figure 6 by an arrow connecting the two segregation indicators.

Additional marker loci can be included in a completely analagous fashion. To include the "N-locus", we define

$$\begin{aligned} N_{(i,0)} &\sim Ber\,(p_N) \\ N_{(i,1)} &\sim Ber\,(p_N) \end{aligned}$$

for the sire's genes and

$$\begin{aligned} S_{ij}^{N} &\sim \begin{cases} Ber(r_{QN}) & \text{if } S_{ij}^{Q} = 0 \\ Ber(1 - r_{QN}) & \text{if } S_{ij}^{Q} = 1 \end{cases} \\ N_{(ij,0)} &\sim Ber\,(p_N) \\ N_{(ij,1)} &= \begin{cases} N_{(i,0)} & \text{if } S_{ij}^{N} = 0 \\ N_{(i,1)} & \text{if } S_{ij}^{N} = 1 \end{cases} \end{aligned}$$

for the segregation indicator and daughter's genes where r_{QN} is the recombination fraction between the QTL and the N-locus. This is in fact a redundant parameter since r_{QN} can be expressed in terms of r_{MN} and r_{MQ} (Section 2.2). Further nodes, N_i and N_{ij}, are defined for the (observable) genotypes of sire and offspring at this locus which depend on the relevant genes. Arrows reflecting all these dependencies are added, exactly as before, to get the model in Figure 7.

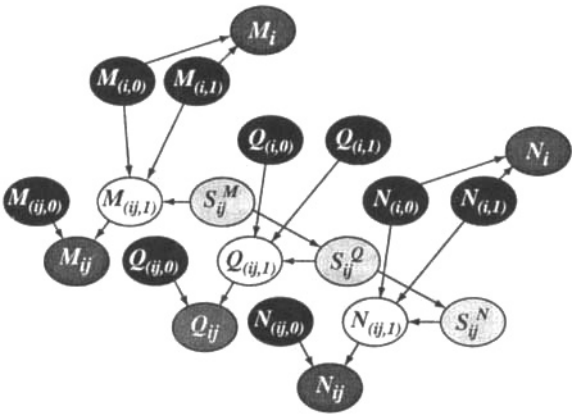


Figure 7. *Two individuals with three loci. Nodes corresponding to genes sampled from the population are shown in black. The genes sampled from the sire are shown in white. The segregation indicator nodes are shown in light grey.*

In addition to marker data, we also have phenotype information for the continuous trait of interest on the daughters. According to our model (Section 3), this depends both on the genotype at the QTL and on the unlinked polygenic effect inherited by the daughter from her sire, and these dependencies are reflected by the arrows shown in Figure 8. The node, s_i, represents the effect of sire i and has the Normal distribution $N(0, \frac{1}{4}\sigma_u^2)$ where σ_u^2 is all the additive genetic variance unexplained by the QTL

(Section 3) and $\frac{1}{4}\sigma_u^2 = \sigma_s^2$. Recall that y_{ij} is the phenotype record on offspring j of sire i and μ_1, μ_2 and μ_3 are the fixed effects corresponding to each of the three unordered genotypes at the QTL: QQ, Qq and qq. Then, the observed phenotype is also assumed to derive from a normal distribution:

$$y_{ij} \sim \begin{cases} N\left(s_i + \mu_1, \frac{3}{4}\sigma_u^2 + \sigma_e^2\right) & \text{if} \quad Q_{(ij,0)} = Q_{(ij,1)} = Q \\ N\left(s_i + \mu_2, \frac{3}{4}\sigma_u^2 + \sigma_e^2\right) & \text{if} \quad Q_{(ij,0)} \neq Q_{(ij,1)} \\ N\left(s_i + \mu_3, \frac{3}{4}\sigma_u^2 + \sigma_e^2\right) & \text{if} \quad Q_{(ij,0)} = Q_{(ij,1)} = q \end{cases}$$

where σ_e^2 represents the environmental variance and $\frac{3}{4}\sigma_u^2$ is the proportion of the additive genetic variance unexplained by the QTL or the sire. Note that $\frac{3}{4}\sigma_u^s + \sigma_e^2 = \sigma_{res}^2$ as we had before in Section 3. Assuming $\sigma_s^2, \sigma_{res}^2, \mu_1, \mu_2, \mu_3$ are known for now, the corresponding model for one sire with two daughters is shown in Figure 8. In Section 6, we will see that the quantities $p_M, p_N, p_Q, \lambda_Q, \sigma_S^2, \sigma_{res}^2, \mu_1, \mu_2, \mu_3$ can all easily be considered as random and the corresponding graph is shown in Figure 9.

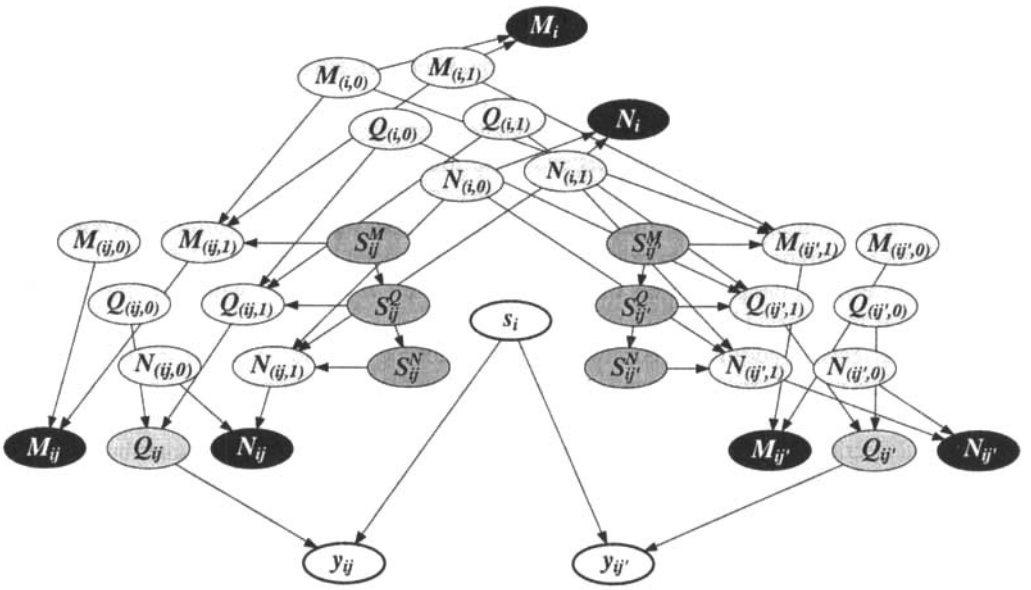


Figure 8. *Graphical model for the half-sib design of Figure 1 where only one sire with two daughters is considered.*

Note that although the pedigree of Figure 1 is clearly unlooped, the corresponding graph of Figure 8 for this three-locus genetic model (Section 3.1) has many loops. The loop defined by the node ordering

$$\{M_{(ij,1)}, M_{(i,1)}, M_i, M_{(i,0)}, M_{(ij,1)}\}$$

is a very simple example whereas a more complex loop is created by the trail

$$\{S_{ij}^M, M_{(ij,1)}, M_{(i.0)}, M_{(ij',1)}, S_{ij'}^M, S_{ij'}^Q, Q_{(ij',1)}, Q_{(i,0)}, Q_{(ij,1)}, S_{ij}^Q, S_{IJ}^M\}.$$

In an industrial-sized half-sib design where there are up to 100 offspring associated with any particular sire, the loop possibilities are uncountable although the pedigree is still a tree. The graphical model representation of Figure 8 shows explicitly why computational problems arise when calculations with multiple loci are required. It should not be surprising to learn that an exact likelihood analysis

for the problem represented in Figure 8 is only feasible when small numbers of offspring in each sire family are involved. For a reasonably large (i.e. typical) design, exact computational methods break down and MCMC methods must be used.

6 Implementation on a Simulated Example

A Bayesian MCMC analysis for the model described in Section 3 was carried out on a simulated half-sib design using a graphical models program written around the HUGIN package (`http://www.hugin.com`).

6.1 Simulated Design and Prior Distributions

The design was a realistically large half-sib design comprising 15 sires, each with 100 offspring. The two diallelic marker loci were known to be 10 cM apart and the QTL was placed at the centre of this interval resulting in a recombination fraction of 0.05 between either marker and the QTL which, in turn, corresponds to a map position of $\lambda_Q \sim 0.053$. Allele frequencies for the simulations were set to be 0.4 for both p_M and p_N while p_Q was 0.5. Genes were assigned to the founders in the design from the appropriate Bernoulli distributions and offspring genotypes were simulated in accordance with Mendelian segregation and the linkage model described in Section 5.2. As a further addition to the model of Section 3, the allele frequencies of the markers were also assumed unknown in the analysis and were assigned the Beta prior distribution introduced in Section 3.1 for the case of the QTL. For all three loci, the parameters of the Beta prior were $a = b = 1$ which defines a uniform distribution over the interval [0, 1].

The "fixed" effects, or additive effects for the QTL genotype were set to be: $\mu_1 = 10$, $\mu_2 = 0$ and $\mu_3 = -10$ and the prior distributions on these were assumed to be improper uniform. A quantitative genetic effect unlinked to either the markers or the QTL was simulated in accordance with the description in Section 3.1. For this, it was assumed that the conditional prior distribution of sire effects, s_i, given the sire variance σ_s^2, was Gaussian, with mean vector zero and with variance $\mathbf{I}\sigma_s^2$, where $\mathbf{I}$ is the identity matrix of order 15×15. The sire variance and residual variance were assigned independent scaled inverted chi-square distributions (Section 3.5), with degrees of freedom $v_s = 5$ and $v_{res} = 5$ respectively, and corresponding respective scale parameters $S_s = 7.5$ and $S_{res} = 92.5$. For the simulation experiment, we took $\sigma_s^2 = 7.5$ and $\sigma_{res}^2 = 92.5$.

The phenotypic data for the quantitative trait were simulated for all 1500 offspring of the 15 sires in accordance with the sampling model described in Section 3.1 for the given parameter values and priors:

$$y_{ij} = s_i + q_{ij} + e_{ij}$$

where

$$q_{ij} = \begin{cases} \mu_1 & \text{for} \quad Q_{ij} = QQ \\ \mu_2 & \text{for} \quad Q_{ij} = Qq \text{ or } qQ \\ \mu_3 & \text{for} \quad Q_{ij} = qq \end{cases}$$

and $e_{ij} \sim N(0, \sigma_{res}^2)$.

The full graphical model for this analysis is shown in Figure 9 for a single sire with two offspring where we have elaborated on the graph we built in Section 5 by adding nodes for the QTL genotype effects μ_1, μ_2, μ_3, variance components $\sigma_s^2, \sigma_{res}^2$, QTL map location λ_Q, and allele frequencies, all of which are held to derive from the prior distributions given above.

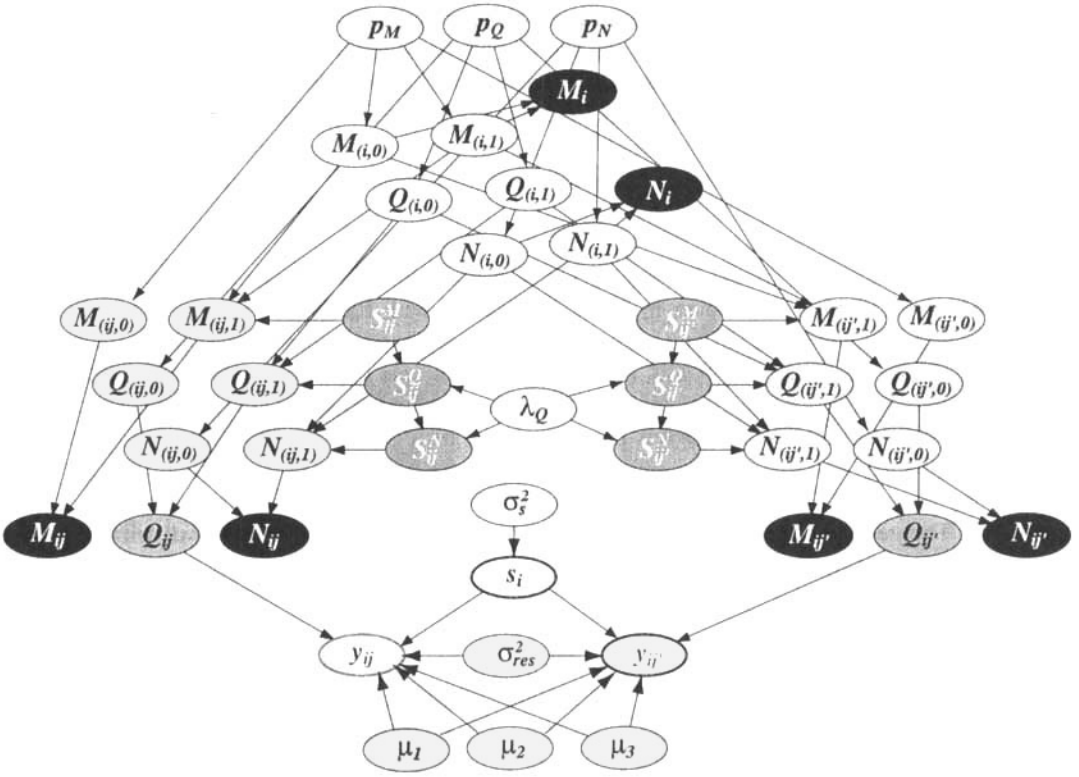


Figure 9. *Graphical model for the full Bayesian analysis on a half-sib design with one sire and two daughters.*

6.2 Implementation

The model was implemented via a two-phase Gibbs sampling approach. In the first phase, the discrete nodes in the graph of Figure 8 comprising all segregation indicators, genes and genotypes were sampled jointly. The second phase included sampling of the allele frequencies, the QTL position, the two variance components (one at a time) and the "location parameters" which comprise the three QTL genotypic effects and fifteen sire effects. Variance components and allele frequencies were all sampled using a single-site Gibbs sampler from their respective fully conditional posterior distributions in Sections 3.5 and 3.6. The QTL position, λ_Q, was sampled using a Metropolis–Hastings algorithm as described in Section 3.3 and this was converted to the recombination fraction r_{MQ} for the block sampling step involving the segregation indicators. The 18 genotype and sire effects were drawn in one pass from their joint fully conditional posterior distribution (Section 3.4). The package HUGIN was used in the first stage of the MCMC algorithm. The program sets up a graph representing all the individual genes, segregation indicators, quantitative phenotypes and the various connections or links between the nodes. The package allows simultaneous sampling of all the discrete nodes of the graph in Figure 8 conditionally on all the parameters and the observed phenotypes. This kind of blocking, or joint updating of large groups of variables, should greatly facilite mixing of the MCMC samplers. Of course, other blocking schemes can be tried and sampler performance should be closely examined.

After a fair amount of experimentation, the reported inferences were based on results from a single long chain of 98,000 updates. The first 200 samples were discarded (burn-in) following which every 10^{th} sample was kept yielding a final sample size or actual chain length of 9780. Convergence

was studied informally, using trace plots and comparing results obtained from independent chains. Estimates of effective chain lengths were obtained for all the parameters based on one of the methods proposed by Geyer (1992). The minimum effective chain length which corresponded to the estimate of the QTL position λ_Q was 376. For all the other parameters, effective chain length was larger than 2000.

6.3 Results

Monte Carlo estimates of the 2.5% and 97.5% percentiles of the posterior distributions for a few selected parameters are shown in Table 1, together with the "true" values from which the data were actually simulated. In any particular sample, we expect to observe random fluctuations about these "true" values. In this case, for instance, the sample gene frequency of the QTL allele was 0.47, whereas the "true" value was actually 0.50. Consequently, the reported posterior interval does not cover the "true" value for this parameter. The results for the QTL genotypic effects are represented by $\Delta_a = |\mu_1 - \mu_3|$, which describes the difference between the two homozygotes, and by $\Delta_d = \mu_2 - \frac{\mu_1+\mu_3}{2}$, the *dominance deviation*, which measure the effect of the heterozygous genotype. The general conclusion from the results in Table 1 is that the posterior distributions give good coverage overall for the simulated values.

Table 1

Simulated values of the parameters (2nd line) and 2.5% and 97.5% percentiles (1st and 3rd lines) from the appropriate posterior distribution.

Parameter:	σ_s^2	σ_{res}^2	p_Q	Δ_a	Δ_d	λ_Q
2.5% Percentile	4.8	81.3	0.409	13.7	-4.62	0.014
Simulated Value	7.5	92.5	0.50	20.0	0.00	0.053
97.5% Percentile	25.3	113.2	0.495	23.1	1.42	0.099

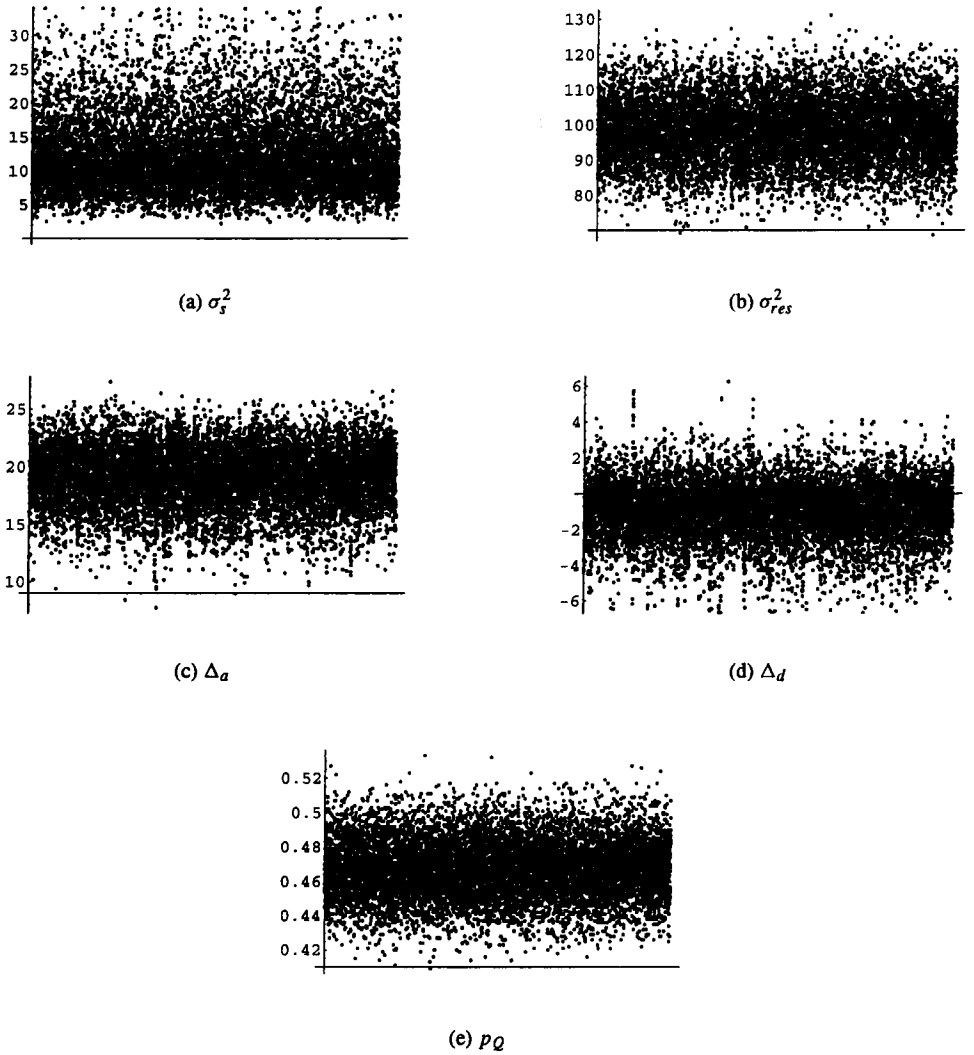
Figure 10 displays trace plots for some of these parameters which indicate that mixing of the Monte Carlo chains appears to be satisfactory. This is also supported by the rather small autocorrelations observed between samples and in the sizes of the effective chain lengths (not shown).

Histograms of the posterior distributions of the parameters in Figure 10 are shown in Figure 11. We note that the marginal posterior distribution of the sire variance is markedly skewed, suggesting that there is relatively little information in the data (15 sire families) to infer this parameter. Some degree of skewness is also noticeable in the posterior distributions of Δ_a and Δ_d pertaining to the QTL effects. As expected, the residual variance shows little sign of asymmetry. Thus, the conclusion from this small simulation study is that the graphical model implementation of this Bayesian MCMC analysis yields the expected results. As with any other implementation, a detailed investigation into the behaviour of the MCMC samplers should accompany any analysis.

7 Discussion

We have chosen to focus our attention on a very simple QTL mapping model on a very simple pedigree, the half-sib design, with view to demonstrating a graphical model application which can be made more general. The actual analysis that we performed by way of demonstrating our approach is not novel in any way (see George *et al.* (2000), for example) and the half-sib design is not particularly interesting in itself. Clearly there are several faults with the design and model presented above. Firstly, strong selection amongst the bulls leads to an obvious violation of the normality assumptions on sire effects. Maternal inheritance is completely ignored in the half-sib design, even when data on dams are available. In reality, all these animals are inter-related in many different ways and the true pedigree structure which is highly complex is also ignored. The genetic map may be incorrect and marker

data are frequently incomplete due to typing errors. Finally, a more realistic model for detecting a QTL would involve multiple highly polymorphic markers with different possible orderings perhaps, (George *et al.*, 2000) and multiple QTLs. What is interesting, however, is that the half-sib design, once made realistically large, poses computational problems for an exact calculation on this simple model and MCMC methods are already required before any of these desirable modifications are considered. The large cliques which cause the computational problems are transparent in this representation.



(a) σ_s^2 (b) σ_{res}^2

(c) Δ_a (d) Δ_d

(e) p_Q

Figure 10. *Trace plots for the individual parameters.*

The point of this paper is not that we claim to be able to deal with computational problems in genetics which cannot be addressed equally well with existing software and methods. Rather we propose to lend a new perspective to the handling of computationally intensive problems in statistical genetics by reformulating the computational problem as a graphical model and exploiting the expert systems approach to exact probabilistic inference on large graphical model networks.

One big advantage of this approach is that any programme which computes probabilities on general graphical models can be used to carry out these analyses. Although the algorithms used on general Bayesian networks are essentially the same as the peeling algorithms, they are a little more efficient computationally in that they tend to be less problem-specific and exploit local dependencies to a greater extent. With the commercially available HUGIN package, for instance, we exploit the efficient "random propagate" routine for joint sampling of blocks of nodes conditionally on the given values at all the remaining nodes in the graph to construct a flexible block updating MCMC sampling scheme. This is a big advantage for the fast development and testing of new methodological approaches in such a computationally challenging area.

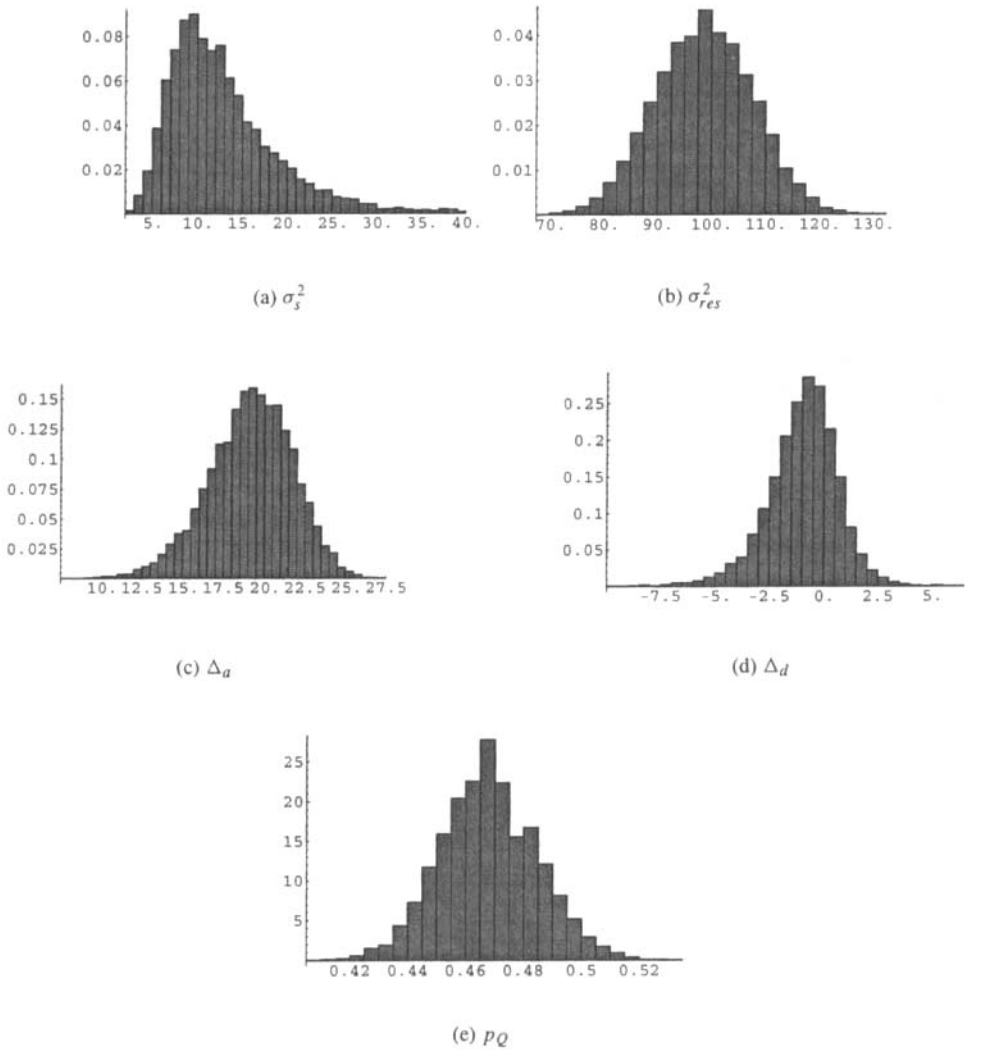


Figure 11. *Distribution of samples of the parameters.*

It is important to note, however, that performing calculations on general pedigrees is going to be just as problematic with this representation as with any other, in the sense that large cliques or cutsets will

cause the peeling algorithms to break down and the MCMC samplers will have the same slow mixing problems etc. The advantage is that all kinds of complex problems in genetics can be accommodated by the same program with very little modification. For example, model extensions and extra pedigree links can easily be incorporated into the graphical model of Figure 9, which was itself extended from the model of Figure 8, by careful addition of extra nodes, links and associated conditional probabilities. The easy handling of phase in the sires is particularly elegant in the graphical model formulation of Section 5 whereas in the Bayesian framework outline in Section 3, the phase problem grows exponentially with the number of loci in the model. The assumption of linkage equilibrium in the sire population means that all phases are equally likely and the first segregation indicator has a Bernoulli ($\frac{1}{2}$) distribution. Sampling segregation indicators jointly with the genotype nodes within HUGIN automatically sums over all phases. Similarly, dealing with missing marker data is far easier in the graphical model formulation of the problem. Allowing the QTL to move across different intervals is not so easy in this particular implementation as the current version of HUGIN does not allow for efficient moving between different networks as would be required in a reversible jump sampling framework.

There will always be a need for problem-specific software in genetics. A pedigree is, after all, a very special type of Bayesian network and some modifications to a general programme are required in order to perform these calculations efficiently. In particular, it is crucial to be able to exploit the often considerable reductions in complexity imposed by the data. Whole sections of the pedigree might be uninformative for a given dataset and should be discarded before a peeling sequence is sought. This process of *clipping* is equivalent to the requirement that data be entered prior to triangulation of a Bayesian network. HUGIN, for example, insists on finding a triangulation before the data are entered whereas a statistical geneticist would always remove the redundant parts of the pedigree before calculating probabilities. While this particular feature could easily be removed, there are many other computational shortcuts that a pedigree-specific program would exploit. Consequently, for any given problem, a general program is bound to compare unfavourably with some purpose-written code. However, one of the big problems in this area is that rapid biological advances are often such that existing efficient task-specific software cannot readily be extended and becomes obsolete thereby impeding quick response to new analytic challenges.

In conclusion, the graphical model provides a powerful and flexible way to view problems in genetics. Preliminary investigations have indicated that there are huge computational advantages to be gained from taking this approach and programs are easily modified to cater for a wide range of problems. In particular, the large complex animal pedigrees which are primarily of interest here, take us into a category of Bayesian networks on which computations are known to be difficult and thus present an exciting challenge to the graphical modelling community. Furthermore, the setting of these large problems in genetics into a more general modelling framework allows for cross-fertilisation of ideas between the hitherto separate communities of genetics and artificial intelligence.

8 Acknowledgements

The authors acknowledge funding for research visits from the ESF Programme on Highly Structured Stochastic Systems which instigated our collaboration and the Wellcome Trust Biomedical Research Collaboration Grant 056266/Z/98/Z. In addition, Daniel Sorensen acknowledges partial financial support from the Danish Agricultural and Veterinary Research Council grant 53-00-0332, Bernt Guldbrandtsen and Mogens Lund acknowledge support from the Danish Ministry of Agriculture, Fisheries and Food grant FREM98-DJF-1 and Nuala Sheehan acknowledges support from the TVW Telethon Institute for Child Health Research, Perth, Western Australia.
We are all grateful to Steffen Lauritzen for his supportive comments and advice throughout.

References

Andersen, S.K., Olesen, K.G., Jensen, F.V. & Jensen, F. (1989). HUGIN—a shell for building Bayesian belief universes for expert systems. In *Proceedings of the 11th International Joint Conference on Artifical Intelligence*, pp. 1080–1085. San Mateo: Morgan Kaufmann.

Cannings, C., Thompson, E.A. & Skolnick, M.H. (1978). Probability functions on complex pedigrees. *Advances in Applied Probability*, **10**, 26–61.

Churchill, G.A. & Doerge, R.W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963–971.

Cowell, R.G., Dawid, A.P., Lauritzen, S.L. & Spiegelhalter, D.J. (1999). *Probabilistic Networks and Expert Systems*, Statistics for Engineering and Information Science. New York: Springer-Verlag, Inc.

Dawid, A.P. (1992). Applications of a general propagation algorithm for probabilistic expert systems. *Statistics and Computing*, **2**, 25–36.

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

Falconer, D.S. & Mackay, T.F.C. (1996). *Introduction to Quantitative Genetics*, fourth edn., Longman Group Ltd.

Gelfand, A.E., Hills, S.E., Racine-Poon, A. & Smith, A.F.M. (1990). Illustration of Bayesian Inference in Normal Data Models using Gibbs Sampling. *Journal of the American Statistical Society*, **85**, 972–985.

George, A.W., Mengersen, K.L. & Davis, G.P. (2000). Localization of a quantitative trait locus via a Bayesian approach. *Biometrics*, **56**, 40–51.

Georges, M., Nielsen, D., Mackinnon, M., Mishra, A., Okimoto, R., Pasquino, A.T., Sargeant, L.D., Sorensen, A., Steele, M.R., Zhao, X., Womack, J.E. & Hoeschele, I. (1995). Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics*, **139**, 907–920.

Geyer, C.J. (1992). Practical Markov Chain Monte Carlo (with discussion). *Statistical Science*, **7**, 473–511.

Gilks, W.R., Thomas, A. & Spiegelhalter, D.J. (1994). A Language and Program for Complex Bayesian Modelling. *Statistician*, **43**, 169–177.

Good, P. (1994). *Permutation Tests: A practical Guide to Resampling for Testing Hypotheses*. New York: Springer-Verlag.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Guo, S.W. & Thompson, E.A. (1992). A Monte Carlo method for combined segregation and linkage analysis. *American Journal of Human Genetics*, **51**, 1111–1126.

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**, 229–309.

Haley, C.S., Knott, S.A. & Elsen, J.M. (1994). Mapping Quantitative Trait Loci in Crosses between Outbred Lines Using Least Squares. *Genetics*, **136**, 1195–1207.

Hastings, W.K. (1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–100.

Heath, S.C. (1997). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *American Journal of Human Genetics*, **61**, 748–760.

Hoeschele, I., Uimari, P., Grignola, F., Zhang, Q. & Gage, K. (1997). Advances in statistical methods to map quantitative trait loci in outbred populations. *Genetics*, **147**, 1445–1457.

Hurme, P., Sillanpää, M.J., Arjas, E., Repo, T. & Savolainen, O. (2000). Genetic basis of climatic adaptation in Scots Pine by Bayesian quantitative trait locus analysis. *Genetics*, **156**, 1309–1322.

Jansen, R.C. (1996). A general Monte Carlo method for mapping multiple quantitative trait loci. *Genetics*, **142**, 305–311.

Jansen, R.C. & Stam, P. (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, **136**, 1447–1455.

Jansen, R.C., Johnson, D.L. & Van Arendonk, J.A.M. (1998). A mixture approach to the mapping of quantitative trait loci in complex populations with an application to multiple cattle families. *Genetics*, **148**, 391–398.

Janss, L.L.G., Thompson, R. & Van Arendonk, J.A.M. (1995). Application of Gibbs sampling for inference in a mixed major gene-polygenic inheritance model in animal populations. *Theoretical and Applied Genetics*, **91**, 1137–1147.

Jensen, C.S. (1997). Blocking Gibbs Sampling for Inference in Large and Complex Bayesian Networks with Applications in Genetics, PhD thesis, Aalborg University, Denmark.

Jensen, C.S. & Kong, A. (1999). Blocking Gibbs sampling for linkage analysis in large pedigrees with many loops. *American Journal of Human Genetics*, **65**, 885–901.

Jensen, C.S. & Sheehan, N. (1998). Problems with determination of noncommunicating classes for Monte Carlo Markov chain applications in pedigree analysis. *Biometrics*, **54**, 416–425.

Jensen, C.S., Kjærulff, U. & Kong, A. (1995). Blocking Gibbs sampling in very large probabilistic expert systems. *International Journal of Human-Computer Studies*, **42**, 647–666.

Jensen, F.V. (1996). *An Introduction to Bayesian Networks*. UCL Press, University College Limited.

Knott, S.A. & Haley, C.S. (1992). Maximum likelihood mapping of quantitative trait loci using full-sib families. *Genetics*, **132**, 1211–1222.

Kong, A. (1991). Efficient methods for computing linkage likelihoods of recessive diseases in inbred pedigrees. *Genetic Epidemiology*, **8**, 81–103.

Lander, E.S. & Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**, 185–199.

Lange, K. & Elston, R.C. (1975). Extensions to Pedigree Analysis. I. Likelihood Calculations for Simple and Complex Pedigrees. *Human Heredity*, **25**, 95–105.

Lauritzen, S.L. & Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their applications to expert systems. *Journal of the Royal Statistical Society, Series B*, **50**, 157–224.

Lauritzen, S.L., Dawid, A.P., Larsen, B.N. & Leimer, H.G. (1990). Independence properties of directed Markov fields. *Networks*, **20**, 491–505.

Lee, J.K. & Thomas, D.C. (2000). Performance of Markov chain Monte Carlo approaches for mapping genes in oligogenic models with an unknown number of loci. *American Journal of Human Genetics*, **67**, 1232–1250.

Lund, M.S. & Jensen, C.S. (1999). Blocking Gibbs sampling in the mixed inheritance model using graph theory. *Genetics, Selection, Evolution*, **31**, 3–24.

Mackinnon, M.J. & Weller, J.I. (1995). Methodology and accuracy of estimation of quantitative trait loci parameters in a halfsib design using maximum likelihood. *Genetics*, **141**, 755–770.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N. & Teller, A.H. (1953). Equations of State Calculations by Fast Computing Machines. *Journal of Chemistry and Physics*, **21**, 1087–1091.

Ott, J. (1999). *Analysis of Human Genetic Linkage*, third edn. Baltimore: The Johns Hopkins University Press.

Pearl, J. (1988). *Probabilistic Inference in Intelligent Systems*. San Mateo, California: Morgan Kaufmann.

Sheehan, N.A. (2000). On the application of Markov chain Monte Carlo methods to genetic analyses on complex pedigrees. *International Statistical Review*, **68**, 83–110.

Sillanpää, M.J. & Arjas, E. (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics*, **148**, 1373–1388.

Sillanpää, M.J. & Arjas, E. (1999). Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. *Genetics*, **151**, 1605–1619.

Sobel, E. & Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics*, **58**, 1323–1337.

Spiegelhalter, D.J. (1998). Bayesian graphical modelling: a case study in monitoring health outcomes. *Applied Statistics*, **47**, 115–133.

Stephens, D.A. & Fisch, R.D. (1998). Bayesian analysis of a quantitative trait locus data using reversible jump Markov chain Monte Carlo. *Biometrics*, **54**, 1334–1347.

Thaller, G. & Hoeschele, I. (1996a). A Monte Carlo method for Bayesian analysis of linkage between single markers and quantitative trait loci. I. Methodology. *Theoretical and Applied Genetics*, **93**, 1161-1166.

Thaller, G. & Hoeschele, I. (1996b). A Monte Carlo method for Bayesian analysis of linkage between single markers and quantitative trait loci. II. A simulation study. *Theoretical and Applied Genetics*, **93**, 1167-1174.

Thomas, A., Gutin, A., Abkevich, V. & Bansal, A. (2000). Multilocus linkage analysis by blocked Gibbs sampling. *Statistics and Computing*, **10**, 259–269.

Thomas, D.C. & Cortessis, V. (1992). A Gibbs Sampling Approach to Linkage Analysis. *Human Heredity*, **42**, 63–76.

Thompson, E.A. (1994). Monte Carlo likelihood in genetic mapping. *Statistical Science*, **9**, 355–366.

Thompson, E.A. (2000). *Statistical Inference from Genetic Data on Pedigrees*, Institute of Mathematical Statistics and the American Statistical Association, NSF-CBMS regional Conference Series in Probability and Statistics, **6**.

Thompson, E.A. (2001). Monte Carlo methods on genetic structures. In *Complex Stochastic Systems*, Eds. O.E. Barndorff-Nielsen, D.R. Cox and C. Kluppelberg, pp. 176–218 (Ch. 4). Chapman & Hall.

Thompson, E.A. & Heath, S.C. (2000). Estimation of conditional multilocus gene identity among relatives. In *Statistics in molecular biology and genetics*, Ed. F. Seiller-Moiseiwitsch, pp. 95–113. IMS Lecture Notes, Institute of Mathematical Statistics. American Mathematical Society.

Uimari, P. & Hoeschele, I. (1997). Mapping linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. *Genetics*, **146**, 735-743.

Uimari, P. & Sillanpää, M.J. (2001). Bayesian oligogenic analysis of quantitative and qualitative traits in general pedigrees. *Genetic Epidemiology*, **21**, 224-242.

Uimari, P., Thaller, G. & Hoeschele, I. (1996). The use of multiple markers in a Bayesian method for mapping quantitative trait loci. *Genetics*, **143**, 1831-1842.

Wang, T., Fernando, R.L., Stricker, C. & Elston, R.C. (1996). An approximation to the likelihood for a pedigree with loops. *Theoretical and Applied Genetics*, **93**, 1299–1309.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, **5**, 161–215.

Yi, N. & Xu, S. (2000). Bayesian mapping of quantitative trait loci under the identity-by-descent-based variance component model. *Genetics*, **156**, 411–422.

Résumé

Les modèles graphiques fournissent une approche efficace et souple pour l'analyse de problèmes complexes en génétique. Alors qu'un logiciel spécifique peut être un outil extrêmement efficace pour une analyse particulière, il est souvent difficile de l'adapter à de nouveaux traitements qui vont au-delà de ses fonctionnalités. En considérant les applications génétiques dans un cadre plus général, on peut utiliser principalement le même logiciel pour traiter de nombreuses difficultés. Ceci constitue un atout dans un domaine où la rapidité des évolutions méthodologiques est fondamentale. Une fois qu'une méthode a été complètement développée et testée, le recours à un logiciel spécifique peut alors s'imposer. L'objectif de l'article est d'illustrer l'usage potentiel de l'approche par les modèles graphiques dans les analyses génétiques, en prenant comme exemple un problème très simple et facile à comprendre.

[*Received September 2001, accepted February 2002*]