Theor Appl Genet (1996) 93: 1161–1166

G. Thaller · I. Hoeschele

A Monte Carlo method for Bayesian analysis of linkage between single markers and quantitative trait loci. I. Methodology

Received: 30 May 1996 / Accepted: 14 June 1996

Abstract A Bayesian approach to the statistical mapping of Quantitative Trait Loci (QTLs) using single markers was implemented via Markov Chain Monte Carlo (MCMC) algorithms for parameter estimation and hypothesis testing. Parameter estimators were marginal posterior means computed using a Gibbs sampler with data augmentation. Variables sampled included the augmented data (marker-QTL genotypes, polygenic effects), an indicator variable for linkage, and the parameters (allele frequency, QTL substitution effect, recombination rate, polygenic and residual variances). Several MCMC algorithms were derived for computing Bayesian tests of linkage, which consisted of the marginal posterior probability of linkage and the marginal likelihood of the QTL variance associated with the marker.

Key words Linkage analysis · Bayesian method · Markov chain Monte Carlo · Quantitative-trait loci

Introduction

The availability of many mapped genetic markers and the large family sizes of livestock species allow one to search for quantitative trait loci (QTLs) anywhere in the genome. Traditional methods for identifying major genes or marker-QTL linkages are the Maximum Likelihood (ML)-based approaches of complex segregation and of combined segregation and linkage analysis, respectively. Use of the ML methods requires the researcher, for computational reasons, to either make several simplifying and unrealistic assumptions, e.g., no polygenic effects or relationships among families and no other random effects such as litter effects, or else to account for these effects in an approximate way (Hasstedt 1993).

The development of Markov chain Monte Carlo (MCMC) algorithms (e.g., Smith and Roberts 1993) can eliminate the need for simplifying assumptions. Parameter estimation via Monte Carlo Expectation Maximization (MCEM) algorithms (Guo and Thompson 1992; Thaller et al. 1996) can include any number of additional fixed and random effects and simultaneously account for relationships across families, i.e., full pedigree information. The evaluation of likelihood ratios for testing hypotheses about major genes or linkages via MCMC algorithms is more difficult. An MCMC algorithm combining the importance and Gibbs sampling of the unknown major or QTL genotypes was presented by Thompson and Guo (1991) for likelihoods with broadly similar parameter values. If under the null and alternative hypotheses the parameter values differ substantially, this algorithm will not be useful and more sophisticated techniques will be required, e.g., Monte Carlo mixtures (Geyer 1991).

Even if the inclusion of nuisance parameters is feasible computationally, ML is not likely to be the best method for testing hypotheses about, and for estimating, the parameters of interest, because it does not account for the uncertainty associated with the nuisance parameters. In the presence of many nuisance parameters, likelihood ratio tests may approach their asymptotic distribution only slowly (Zeng 1994). Therefore, Bayesian methods have been suggested by several authors for the identification of major genes or linkages. Hoeschele and VanRaden (1993a,b) derived a Bayesian analysis of linkage between single genetic markers and quantitative trait loci. Hoeschele (1994) extended this method to multiple markers and described a Gibbs sampler for this analysis. Thomas and Cortessis (1992) developed a Bayesian method, implemented via Gibbs sampling, for a simple model of disease etiology, given a single marker. Janss et al. (1995) and Thaller et al. (1996) presented a Bayesian approach for complex segregation analysis of a continuous trait and a categorical trait, respectively.

Communicated by E. J. Eisen

G. Thaller[1] · I. Hoeschele (✉)
Department of Dairy Science, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0315, USA

Present address:
[1]Lehrstuhl fuer Tierzucht, Technische Universität Muenchen, D-85350 Freising-Weihenstephan, Germany

In the Bayesian analysis, inferences about the parameters of interest, or any function thereof, can be derived from their marginal posterior distributions which account for uncertainty in the other parameters. When Bayesian analysis is implemented via MCMC methods, desired inferences from marginal distributions can be obtained even for very general models (e.g., models including additional fixed and random effects). Bayesian analysis can furthermore incorporate prior knowledge about the probability of linkage (Thomas and Cortessis 1992; Hoeschele and VanRaden 1993a,b) and about the distribution of gene effects at QTLs (Hoeschele and VanRaden 1993a,b; Fernando and Grossman 1989; Goddard 1992). By varying the prior assumptions, a robustness study can be conducted to assess the degree of dependency of the outcome of the analysis on the prior information.

In this paper, we present a Bayesian analysis of linkage between single genetic markers and a QTL. This research implements the method of Hoeschele and VanRaden (1993a,b) via MCMC algorithms. In another contribution (Uimari et al. 1996), we will report on Bayesian linkage analysis using multiple linked markers. Here, we first derive Bayesian point estimators for the parameters, which are computed via a Gibbs sampler with data augmentation. We then derive MCMC versions of Bayesian tests for linkage. In a companion paper (Thaller and Hoeschele 1996) we apply the methods to simulated granddaughter designs (GDD) or half-sib designs used in cattle for the mapping of QTLs.

Methods

Parameter estimation

Bayesian linkage analysis was implemented via the Gibbs sampler in combination with data augmentation (Tanner and Wong 1987). In each Gibbs cycle, all parameters and missing data were sampled from conditional distributions given the observed data and current sample values of all other variables. The parameter vector (θ) included the substitution effect (α) and gene frequency (p) at a biallelic QTL, marker-QTL recombination rate (r), an overall mean and additional fixed effects (β), and polygenic, σ_u^2, and residual variance, σ_e^2. Allele frequencies at the marker locus were assumed known, as these can be accurately and independently estimated from the marker genotypes, but can easily be included in the sampling scheme (Uimari et al. 1996). The missing data were the marker-QTL (***MG***) genotypes and polygenic effects (***u***) of all individuals. The genotypes in ***MG*** were defined as complete multi-locus genotypes (known linkage phases). Also included in the sampling scheme was an indicator variable $\mathscr{L}$ representing either nonlinkage ($\mathscr{L}$=0) or linkage ($\mathscr{L}$=1). Under $\mathscr{L}=0$, r=0.5, and under $\mathscr{L}=1, 0 \leq r < 0.5$. Below, $P(.)$ will denote the (joint) probability of a (set of) discrete variable(s), and $f(.)$ will denote the joint probability density of a set of continuous variables or of both continuous and discrete variables.

The joint posterior distribution of a sample of the missing data and parameters, given the observed phenotypic (***y***) and marker (***M***) data, is

$$f(r, \theta_{-r}, \boldsymbol{MG}, \boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{M}) = P(\mathscr{L}=0 \mid \boldsymbol{y}, \boldsymbol{M}) f(\theta_{-r}, \boldsymbol{MG}, \boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{M}, \mathscr{L}=0)$$
$$+ P(\mathscr{L}=1 \mid \boldsymbol{y}, \boldsymbol{M}) f(\theta, \boldsymbol{MG}, \boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{M}, \mathscr{L}=1) \qquad (1)$$

where $\theta_{-\mathrm{r}}$ equals θ with r omitted, and where

$$f(\theta_{-r}, r, \boldsymbol{MG}, \boldsymbol{u}\ \boldsymbol{y}, \boldsymbol{M}, \mathscr{L}=1) \propto$$
$$f(\boldsymbol{y} \mid \boldsymbol{MG}, u, \theta_{-r})\, P(\boldsymbol{M} \mid \boldsymbol{MG})\, P(\boldsymbol{MG} \mid r, p) f(\boldsymbol{u} \mid \sigma_u^2)$$
$$f(\theta_{-r}) f(r \mid \mathscr{L}=1) \qquad (2)$$

is the joint posterior probability of the parameters and missing data under linkage. The joint posterior under nonlinkage equals the right-hand-side in (2), except that $f(r|\mathscr{L}=1)$ is replaced with $f(r|\mathscr{L}=0)=I(r=0.5)$, where $I(.)$ is the indicator function equal to 1 if the condition is true or zero otherwise. In (2), it was assumed that r and the other parameters are independent a priori, and

$$f(\theta_{-r}) = f(p) f(\alpha) f(\beta) f(\sigma_u^2) f(\sigma_e^2)$$

is the prior density of the other parameters assuming prior independence. Further, $f(\boldsymbol{u} \mid \sigma_u^2)$ is the density of the polygenic effects, $P(\boldsymbol{MG} \mid r, p)$ is the joint probability of a set of genotypes for the pedigree, and $P(\boldsymbol{M} \mid \boldsymbol{MG})$ is 1 for any ***MG*** compatible with ***M*** and 0 otherwise.

The marginal posterior probability of nonlinkage as defined in equation (6) of Hoeschele and VanRaden (1993b) can be rewritten as

$$P(\mathscr{L}=0 \mid \boldsymbol{y}, \boldsymbol{M})$$
$$= \frac{P(\mathscr{L}=0) f(\boldsymbol{y}, \boldsymbol{M} \mid \mathscr{L}=0)}{P(\mathscr{L}=0) f(\boldsymbol{y}, \boldsymbol{M} \mid \mathscr{L}=0) + P(\mathscr{L}=1) f(\boldsymbol{y}, \boldsymbol{M} \mid \mathscr{L}=1)}$$
$$= \left[1 + \frac{P(\mathscr{L}=1)}{P(\mathscr{L}=0)} \frac{f(\boldsymbol{y}, \boldsymbol{M} \mid \mathscr{L}=1)}{f(\boldsymbol{y}, \boldsymbol{M} \mid \mathscr{L}=0)}\right]^{-1} \qquad (3)$$

where the marginal likelihood under linkage is

$$f(\boldsymbol{y}, \boldsymbol{M} \mid \mathscr{L}=1)$$
$$= \int_0^{0.5} \int_{p_l}^{p_l} \int_{\alpha_{\min}}^{\infty} \int_0^c \int_0^c \int_{-\infty}^{\infty} \cdots$$
$$\int_{-\infty}^{\infty} \sum_{MG} P(\boldsymbol{MG} \mid r, p)\, P(\boldsymbol{M} \mid \boldsymbol{MG})\, f(y \mid \boldsymbol{MG}, \alpha, \boldsymbol{\beta}, u, \sigma_e^2)$$
$$f(u \mid \sigma_u^2) f(r \mid \mathscr{L}=1)$$
$$f(p, \alpha, \boldsymbol{\beta}, \sigma_e^2, \sigma_u^2)\, du\, d\boldsymbol{\beta} d\sigma_u^2\, d\sigma_e^2\, d\alpha\, dp\, dr \qquad (4)$$

and the marginal likelihood under nonlinkage is identical to (4) except that $f(r|\mathscr{L}=1)$ is replaced by $f(r|\mathscr{L}=0)$. In (4), $\alpha_{\min}$ is a lower limit on α, while p_{l} and p_{u} are lower and upper limits on p, respectively, and c is a constant (see Hypothesis testing).

Sampling from the joint posterior distribution in (1) is not feasible because the marginal probabilities of (non)linkage are unknown [the integrations in (4) are not feasible]. Therefore, a Gibbs sampler was derived based on the joint posterior distribution of the event of (non)linkage, the parameters, and the missing data, or

$$P(\mathscr{L}, \theta, \boldsymbol{MG}, \boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{M}). \qquad (5)$$

Sampling from (5) was performed by sampling in turn from the conditional distribution of the linkage indicator variable $\mathscr{L}$ and of r, and from the conditional distribution of the other parameters and the missing data, or

$$P(\mathscr{L}, r \mid \theta_{-r}, \boldsymbol{MG}, \boldsymbol{u}, \boldsymbol{y}, \boldsymbol{M}),\ P(\theta_{-r}, \boldsymbol{MG}, \boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{M}, r). \qquad (6)$$

The conditional probability of nonlinkage equals

$$P(\mathscr{L}=0 \mid \boldsymbol{\theta}_{-r}, \boldsymbol{MG}, u, y) = \frac{P(\mathscr{L}=0)\, P(\boldsymbol{MG} \mid \mathscr{L}=0, p)\, f(y \mid \boldsymbol{MG}, u, \boldsymbol{\theta})\, f(u \mid \sigma_u^2)\, f(\boldsymbol{\theta}_{-r})}{numerator + P(\mathscr{L}=1) \int_{0.0}^{0.5} P(\boldsymbol{MG} \mid r, p)\, f(\boldsymbol{y} \mid \boldsymbol{MG}, u, \boldsymbol{\theta})\, f(u \mid \sigma_u^2)\, f(\boldsymbol{\theta})\, dr}$$
$$= \frac{P(\mathscr{L}=0)\, P(\boldsymbol{MG} \mid \mathscr{L}=0)}{numerator + P(\mathscr{L}=1) \int_{0.0}^{0.5} P(\boldsymbol{MG} \mid r)\, f(r \mid \mathscr{L}=1)\, dr} = P(\mathscr{L}=0 \mid \boldsymbol{MG}) \qquad (7)$$

where $P(\boldsymbol{MG}|r)$ denotes the part of $P(\boldsymbol{MG}|p,r)$ which depends only on r. The probability density of the missing data and parameters under linkage is

$$f(\boldsymbol{\theta},\boldsymbol{u},\boldsymbol{MG}\,|\,\boldsymbol{y},\boldsymbol{M},\mathscr{L}=1) \propto \left[\prod_{i=1}^{N_b} P(MG_i\,|\,p)\right]\left[\prod_{i=N_b+1}^{N} P(MG_i\,|\,\boldsymbol{MG}_{p_i},r,p)\right] \left[\prod_{i=1}^{N} P(M_i\,|\,MG_i)\,f(y_i\,|\,\boldsymbol{\theta},u_i,MG_i)\right] f(u\,|\,\boldsymbol{\theta})\,f(\boldsymbol{\theta}_{-r})\,f(r\,|\,\mathscr{L}=1) \qquad (8)$$

where N is number of animals in the pedigree, N_b is number of base animals, N is the total number of individuals, p_i=[s_i,d_i], s_i is the sire of i, d_i the dam of i, $f(y_i|.)$ is set to one if animal i does not have an observed phenotype, and $P(M_i|MG_i)$ is set to one for an individual without marker information. Further, $P(MG_i|MG_{pi},r,p)=P(MG_i|MG_{si},MG_{di},r)$ for an individual with both parents known, and $P(MG_i|MG_{pi},r,p)=P(MG_i|MG_{si},r,p)$ for an individual with only the sire known. The conditional probability of the parameters and the missing data under nonlinkage is identical to that in (8) except that r is fixed at 0.5. The integration required in (7) was performed using algorithm AS 63 (Appl. Statist. 22:409) for the integration of the truncated beta density (see sampling of parameters below).

To derive the conditional sampling distributions of the parameters, the posterior probability density (8) must be written more explicitly in terms of the unknown parameters; in particular, the genotype probabilities in (8) must be rewritten in terms of parameters p and r. Prior ignorance about the parameters was represented by an improper flat prior for β, proper uniform priors for p, r, and the variances, and a uniform prior on[0,∞] for α. Prior information on α might be represented by an exponential prior (Hoeschele and VanRaden 1993a) or by a normal prior truncated to the left at zero (Goddard 1992). Then, (8) is equivalent to

$$f(\boldsymbol{\theta},\boldsymbol{u},\boldsymbol{MG}\,|\,\boldsymbol{y},\boldsymbol{M},\mathscr{L}=1) \propto p^{\gamma_p}(1-p)^{\eta_p}\,r^{\gamma_r}(1-r)^{\eta_r}\prod_{k=1}^{m}(q_k)^{\gamma_k} \left[\prod_{i=1}^{n} P(M_i\,|\,MG_i)\,f(y_i\,|\,\alpha,\boldsymbol{\beta},u_i,MG_i,\sigma_e^2)\right] f(\boldsymbol{u}\,|\,\sigma_u^2)\,I(p_l<p<p_u)\,I(0\le r<0.5)\,f(\alpha) \qquad (9)$$

where the γ and η terms are appropriate allele and recombinant counts, respectively, based on the current $\boldsymbol{MG}$ genotypes of individuals, and q_k is frequency of marker allele k.

The event of linkage ($\mathscr{L}$=1) or nonlinkage ($\mathscr{L}$=0) was sampled according to $P(\mathscr{L}=0|\boldsymbol{MG})$ in (7) by sampling a uniform $U(0,1)$ variate x and setting $\mathscr{L}$=0 and r=0.5 if $x<P(\mathscr{L}=0|\boldsymbol{MG})$. If linkage was sampled ($x>P(\mathscr{L}=0|\boldsymbol{MG})$), then $\mathscr{L}$=1 and r was sampled subsequently from the beta distribution Be(γ_r+1,η_r+1) truncated to the right at 0.5.

Fully conditional sampling densities of each parameter can be derived from (9) by collecting and rearranging all terms dependent on the particular parameter in order to obtain a standard distribution to sample from. The sampling distribution for p is Beta(γ_p+1,η_p+1) as in Janss et al. (1995). If considered as unknown in the analysis, allele frequencies at the marker would be sampled from a Dirichlet distribution (Devroye 1986) with parameters determined by the allele counts γ_k.

Sampling distributions for fixed (β) and random effects ($\boldsymbol{u}$) are normal, conditional on a set of genotypes, and can be derived from (9) using linear mixed model theory (e.g., Wang et al. 1993). Mean and variance of each univariate, conditional, normal sampling distribution can be obtained from MME in α, β, and $\boldsymbol{u}$, for a given set of genotypes, from analogy with Gauss-Seidel iteration by replacing current and previous solutions with current and previous sampling values, setting the variance equal to the reciprocal of the diagonal of the equation and the mean equal to the solution. The u effects of parents with several final offspring (sires in a GDD) were sampled from a distribution marginalized with respect to the final offspring but conditional on all other variables as in Janss et al. (1995). Mean and variance of this sampling distribution are obtained by forming MME for the sire and his final progeny with right-hand-sides adjusted for all other effects and absorbing final progeny into the sire equation.

QTL parameter α can be sampled analogously to the elements in β and $\boldsymbol{u}$ from its fully conditional, normal standard sampling distribution if a prior representing a uniform or normal distribution truncated below at zero is used, with sample values below zero being rejected. If an exponential prior distribution for α were chosen instead, the resulting sampling distribution would not be standard and would have to be sampled from via techniques for nonstandard distributions (Devroye 1986), e.g., rejection sampling, adaptive rejection sampling (Gilks and Wild 1992), Metropolis-Hastings within the Gibbs sampler (Chib and Greenberg 1995; Uimari et al. 1996), or adaptive rejection Metropolis sampling within the Gibbs sampler (Gilks et al. 1995).

The fully conditional distributions of the variance components were inverse chi-square, as in the linear mixed model. Variances were sampled as in Wang et al. (1993) except that the degrees of freedom were reduced by 2 resulting from a proper uniform prior for the variances instead of the prior in Wang et al. (1993) now known to cause an improper posterior distribution (e.g., Hoeschele 1989; Hobert and Casella 1994) in the linear mixed model. An improper uniform prior has been shown to produce a proper posterior (Carlin 1992; Gelman and Rubin 1992; Hobert and Casella 1994).

Genotypes of all individuals, except parents with many final offspring, were sampled according to their fully conditional probabilities (Guo and Thompson 1992). The $\boldsymbol{MG}$ genotypes of parents with many final offspring were sampled from posterior probabilities marginalized with respect to the genotypes of the final offspring, as in Janss et al. (1995). For an individual with known inheritance of the marker alleles, sampling its $\boldsymbol{MG}$ was equivalent to sampling $\boldsymbol{G}$ given $\boldsymbol{M}$. For an individual with unknown inheritance at some marker loci (same marker genotype as a single known parent or as both parents), all possible multi-locus ($\boldsymbol{MG}$) genotypes were identified and sampled conditional on parental and nonfinal offspring $\boldsymbol{MG}$ genotypes, final offspring phenotypes, and the individual's marker genotype. For a base animal, all possible multi-locus ($\boldsymbol{MG}$) genotypes were obtained by combining its possible marker linkage phases with the four possible QTL genotypes and sampled conditional on the $\boldsymbol{MG}$ of offspring. Consequently, in each Gibbs cycle a complete set of multi-locus genotype realizations was obtained, allowing sampling of r, p, and the q_k (if unknown) from standard distributions. This approach differs from that of Thomas and Cortessis (1992) who sampled only $\boldsymbol{G}$ given $\boldsymbol{M}$ which led to a nonstandard sampling distribution for r.

Parameter estimators were marginal posterior means, and their $\boldsymbol{MC}$ estimates were averages of all Gibbs samples for the respective parameters.

Hypothesis testing

Two criteria were chosen as tests for linkage between the marker and a QTL. The first criterion was the marginal posterior probability of linkage, $P(\mathscr{L}=1|\boldsymbol{y},\boldsymbol{M})$, and the second was the marginal posterior density of the variance associated with the marker $\sigma^2_{\mathrm{m}}=(1-2r)^2 2p\,(1p)\alpha^2)$, f($\sigma^2_{\mathrm{m}}|\boldsymbol{y},\boldsymbol{M}$). Three different $\boldsymbol{MC}$ methods were used to evaluate $P(\mathscr{L}=1|\boldsymbol{y},\boldsymbol{M})$ which either provide an estimate of the probability directly, or estimates of the marginal likelihoods [(4)], or of the ratio of these likelihoods, f($\boldsymbol{y},\boldsymbol{M}|\mathscr{L}$=1)/f($\boldsymbol{y},\boldsymbol{M}|\mathscr{L}$=0).

For testing linkage versus nonlinkage, different null and alternative hypotheses can be formulated. In this study, the null hypothesis was defined as "the marker is unlinked ($\mathscr{L}=0 \Leftrightarrow r=0.5$) to a QTL with substitution effect a greater than or equal to α_{min} and gene frequency p in the range $0<p_{\mathrm{l}}<p<p_{\mathrm{u}}<1$". The value α_{min} may be set equal to the minimum detectable effect for a given design, and limits p_{l} and p_{u} may be chosen such that at least one family in the design is heterozygous at the QTL. When the hypothesis is formulated in terms of one QTL parameter (here r), such restrictions on the other QTL parameters are necessary because non-linkage can either be expressed as r=0.5, α=0, or p=0.

The marginal posterior probability of non-linkage was estimated directly as the Monte Carlo average of the conditional probabilities of linkage used in the Gibbs sampler, or

$$\hat{P}(\mathscr{L}=0\,|\,\boldsymbol{y},\boldsymbol{M}) = \frac{1}{C}\sum_{c=1}^{C} \frac{P(\mathscr{L}=0)P(\boldsymbol{MG_c}\,|\,r=0.5)}{numerator + P(\mathscr{L}=1)\int_{0.0}^{0.5} P(\boldsymbol{MG_c}\,|\,r)\,f(r\,|\,\mathscr{L}=1)\,dr} \tag{10}$$

where C is the number of Gibbs cycles and sub- or super-script c indicates cycle number, or as the frequency of the cycles in which the nonlinkage event was sampled, i.e.,

$$\hat{P}(\mathscr{L}=0\,|\,\boldsymbol{y},\boldsymbol{M}) = \frac{1}{C}\sum_{c=1}^{C} I(\mathscr{L}_c=0) \tag{11}$$

where $I(.)$ is the indicator function equal to one if in cycle c the sample value for $\mathscr{L}$ was 0 and equal to zero otherwise. The parametric estimator (10) should be preferred but in large samples (10) and (11) yield identical results.

The Monte Carlo estimation of likelihood ratios (classical or Bayesian) is considerably more difficult than point estimation via MCMC methods and still in the development phase. Here we chose to investigate two estimators of marginal likelihoods or of their ratios which have recently been proposed in the literature. The first estimator, suggested by Newton and Raftery (1994), employs an MC estimator of the marginal likelihood based on importance sampling, or

$$f(\boldsymbol{y}) = \frac{\sum_{c=1}^{C} f(y\,|\,\boldsymbol{\theta}_c)\,f(\boldsymbol{\theta}_c)/g(\boldsymbol{\theta}_c)}{\sum_{c=1}^{C} f(\boldsymbol{\theta}_c)/g(\boldsymbol{\theta}_c)} \tag{12}$$

where $g(.)$ is the importance sampling function. Setting $g(.)$ equal to the prior $f(\boldsymbol{q})$ produces an inefficient estimator with large variance, as only few samples from the prior yield non-negligible conditional likelihood values. Setting $g(.)$ equal to the posterior $f(\theta\,|\,\boldsymbol{y})$ leads to an unstable estimator equal to the harmonic mean of the conditional likelihoods, where the rare occurrence of parameter vectors with very small likelihoods dominates the estimate. Because the problems with the aforementioned estimators are of an opposite nature, Newton and Raftery (1994) suggested using a $g(.)$ equal to the density of a mixture of the prior and posterior distributions. As this approach would require sampling both from the prior and the posterior, Newton and Raftery (1994) derived an approximation with samples drawn only from the posterior, which takes the form

$$\hat{f}(\boldsymbol{y}) = \frac{\dfrac{\delta C}{1-\delta} + \sum_{c=1}^{C} \dfrac{f(\boldsymbol{y}\,|\,\boldsymbol{\theta}_c)}{\delta \hat{f}(\boldsymbol{y}) + (1+\delta)\,f(\boldsymbol{y}\,|\,\boldsymbol{\theta}_c)}}{\dfrac{\delta C}{(1-\delta)\,\hat{f}(\boldsymbol{y})} + \sum_{c=1}^{C} \dfrac{1}{\delta \hat{f}(\boldsymbol{y}) + (1-\delta)\,f(\boldsymbol{y}\,|\,\boldsymbol{\theta}_c)}} \tag{13}$$

where δ is the (hypothetical) fraction of samples obtained from the prior, and C is sample size from the posterior. Newton and Raftery (1994) noted that δ values in the range 0.01 to 0.10 worked well in their examples. Equation (13) must be solved iteratively and can be rewritten in the computationally more convenient form

$$\sum_{c=1}^{C} \frac{1-\exp[\ln f(\boldsymbol{y}\,|\,\boldsymbol{\theta}_c) - \ln \hat{f}(\boldsymbol{y})]}{\delta + (1-\delta)\exp[\ln f(\boldsymbol{y}\,|\,\boldsymbol{\theta}_c) - \ln \hat{f}(\boldsymbol{y})]} = 0 \tag{14}$$

which is solved for $\ln f(\boldsymbol{y})$.

In the linkage problem considered here, the observations [$\boldsymbol{y}$ in (13) and (14)] were the observations in $\boldsymbol{y}$ and $\boldsymbol{M}$, and the parameter vector θ also included the missing data ($\boldsymbol{u}$, $\boldsymbol{MG}$). To obtain the ratio $f(\boldsymbol{y},\boldsymbol{M}\,|\,\mathscr{L}=1)/f(\boldsymbol{y},\boldsymbol{M}\,|\,\mathscr{L}=0)$, the numerator was evaluated by Gibbs sampling with r free on (.0,0.5), and the denominator was evaluated by running an additional sampler with r fixed at 0.5. The conditional (on parameters) probability density of the observations required in (13) and (14) was

$$\begin{aligned} f(\boldsymbol{y},\boldsymbol{M}\,|\,\boldsymbol{MG_c},\boldsymbol{u_c},\boldsymbol{\theta}_c &= P(\boldsymbol{M}\,|\,\boldsymbol{MG_c})f(\boldsymbol{y}\,|\,\boldsymbol{MG_c},\boldsymbol{u_c},\boldsymbol{\theta}_c) \\ &= \prod_{i=1}^{N} P(M_i\,|\,MG_{i(c)})f(y_i\,|\,u_{i(c)},MG_{i(c)},\boldsymbol{\theta}_c). \end{aligned} \tag{15}$$

Note that while evaluating (15) would present a numerical problem for a large N, only its logarithm is required in (14).

Meng and Wong (1993) presented an alternative approach to the MC evaluation of likelihood ratios. Their class of MC estimators is of the general form

$$\frac{f_1(\boldsymbol{y})}{f_2(\boldsymbol{y})} = \frac{E_2[f_1(\boldsymbol{\theta},\boldsymbol{y})\,\alpha(\boldsymbol{\theta})]}{E_1[f_2(\boldsymbol{\theta},\boldsymbol{y})\,\alpha(\boldsymbol{\theta})]}$$

where

$$f_i(\boldsymbol{\theta}|\,\boldsymbol{y}) = \frac{f_i(\boldsymbol{\theta},\boldsymbol{y})}{f_i(\boldsymbol{y})} \qquad i=1,\ 2$$

subscript i identifies the hypothesis, E_i denotes expectation with respect to $f_i(\theta\,|\,\boldsymbol{y})$, and $a(\theta)$ is an arbitrary function such that

$$0<\left|\int_{\Omega_1\cap\Omega_2} \alpha(\boldsymbol{\theta})\,f_1(\boldsymbol{\theta}|\,\boldsymbol{y})\,f_2(\boldsymbol{\theta}|\,y)\,d\boldsymbol{\theta}\right|\propto \tag{17}$$

In addition to providing theory for identifying an asymptotically optimum $\alpha(.)$ which must be found iteratively, Meng and Wong (1993) suggested some appealing practical choices for $\alpha(.)$. We used

$$\alpha(\boldsymbol{\theta}) = \frac{1}{\sqrt{f_1(\boldsymbol{y},\boldsymbol{\theta})\,f_2(\boldsymbol{y},\boldsymbol{\theta})}} \tag{18}$$

which leads to an MC estimator of the likelihood ratio equal to

$$\frac{f_1(\boldsymbol{y})}{f_2(\boldsymbol{y})} \approx \frac{\dfrac{1}{C_2}\sum_{c=1}^{C_2} \sqrt{f_1(\boldsymbol{y},\boldsymbol{\theta}_c)/f_2(\boldsymbol{y},\boldsymbol{\theta}_c)}}{\dfrac{1}{C_2}\sum_{c=1}^{C_1} \sqrt{f_2(\boldsymbol{y},\boldsymbol{\theta}_c)/f_1(\boldsymbol{y},\boldsymbol{\theta}_c)}} \tag{19}$$

where the C_2 and C_1 samples in the numerator and denominator, respectively, are from the posterior under hypotheses 2 and 1, respectively.

For the linkage problem considered here, the alternative hypotheses are 1: $\mathscr{L}$=0 and 2: $\mathscr{L}$=1. The joint probability density of the observations and parameters, required in (19), is given below. Note that leaving r out of θ is necessary to satisfy condition (17).

$$\begin{aligned} &f_i(\boldsymbol{y},\boldsymbol{M},\boldsymbol{\theta}_{-r},\boldsymbol{u},\boldsymbol{MG}) = f(\boldsymbol{y}\,|\,\boldsymbol{u},\boldsymbol{MG},\boldsymbol{\theta}_{-r}) \\ &\cdot f(\boldsymbol{u}\,|\,\sigma_u^2)\,P(\boldsymbol{M}\,|\,\boldsymbol{MG})\,P_i(\boldsymbol{MG}\,|\,\boldsymbol{p},r)\,f(\boldsymbol{\theta}_{-r}) \qquad i=1,2 \\ &\frac{f_1(\boldsymbol{y},\boldsymbol{M},\boldsymbol{\theta}_{-r},\boldsymbol{u},\boldsymbol{MG})}{f_2(\boldsymbol{y},\boldsymbol{M},\boldsymbol{\theta}_{-r},\boldsymbol{u},\boldsymbol{MG})} = \frac{P_1(\boldsymbol{MG})}{P_2(\boldsymbol{MG})} \end{aligned} \tag{20}$$

where :

$$P_1(\boldsymbol{MG}) = P(\boldsymbol{MG}\,|\,\mathscr{L}=0) = P(\boldsymbol{MG}\,|\,r=0.5) \text{ and}$$

$$P_2(\boldsymbol{MG}) = P(\boldsymbol{MG}\,|\,\mathscr{L}=1) = \int_0^{0.5} P(\boldsymbol{MG}\,|\,r)\,f(r\,|\,\mathscr{L}=1)dr.$$

Substituting (20) in (19) yields the estimator of the ratio $f(\boldsymbol{y},\boldsymbol{M}\,|\,\mathscr{L}=1)/f(\boldsymbol{y},\boldsymbol{M}\,|\,\mathscr{L}=0)$.

The marginal posterior density of any function of the parameters can be estimated from Gibbs output using non-parametric techniques such as average shifted histograms (Scott 1992). Here, the marginal posterior distribution of the variance associated with the marker was of interest. For any sample of the marker-QTL parameters r, p, and α, the corresponding variance was calculated as $(1-2r)^2 2p(1-p)\alpha^2$, and the sample of variances was used to estimate the marginal posterior density. Janss et al. (1995) estimated the marginal posterior density of the variance at a major locus in this way and suggested using the maximum density estimate and the density estimate at zero to compute a likelihood ratio test from these marginal likelihoods.

Conclusions

Parameter estimates obtained with the Bayesian approach are to be interpreted differently from those obtained in a classical ML analysis. While one would expect fairly similar estimates when there is substantial information in the data about all parameters, one would expect larger differences when there is little information. These expectations result from the fact that the Bayesian analysis attaches a nonzero probability to the nonlinkage hypothesis if there is weak evidence for linkage in the data, and the parameter estimates can be interpreted as weighted averages of the estimates obtained under linkage and nonlinkage.

Several tests for linkage, evaluated from Gibbs output, were presented. The most promising appears to be the test based on marginal posterior probabilities of linkage calculated from equations (10) or (11) (see Thaller and Hoeschele 1996). This method generalizes straightforwardly to multiple linked markers (Uimari et al. 1996) and is similar to MCMC sampling with model indicators (Albert and Chib 1994).

In a practical setting, the method presented here should be employed to reanalyze interesting regions of the genome. An initial analysis with a computationally simple method (e.g., linear regression), which does not allow for the estimation of QTL parameters and the use of full pedigree information, should be performed first. A major advantage of a simple method is that exact threshold values for hypothesis testing can be determined via data permutation (Churchill and Doerge 1992). For simple designs, relationships can be established between the *F* test, the likelihood ratio test, and Bayes factors (Kass and Raftery), but this will most likely not be possible for the more complicated designs and models for outcross populations.

Applications of Bayesian linkage analysis using MCMC algorithms can also be found in plant genetics (Satagopan et al. 1996) and in human genetics (Thomas and Cortessis 1992). The method presented in the present paper extends the work of Thomas and Cortessis (1992) to continuous phenotypes and the work of Satagopan et al. (1996) to outcross populations and more complex models of phenotypic variation.

Acknowledgements The National Science Foundation provided generous support for this project (grant no. BIR-9596247). G. Thaller acknowledges financial support from the Deutsche Forschungsgemeinschaft in the form of a postdoctoral fellowship. I. Hoeschele acknowledges financial support from the European Human Capital and Mobility Fund while on research leave at Wageningen University, The Netherlands. This research was conducted using the resources of the Cornell Theory Center, which receives major funding from the National Science Foundation and New York State. Additional funding comes from the Advanced Research Projects Agency, the National Institutes of Health, IBM Corporation, and other members of the center's Corporate Research Institute.

References

Albert JH, Chib S (1994) Bayesian model checking for binary and categorical response data. Technical Report, Department of Mathematics and Statistics, Bowling Green University, Ohio, USA

Carlin JB (1992) Meta-analysis for 2×2 tables: a Bayesian approach. Stat Med 11:141–159

Chib S, Greenberg E (1995) Understanding the Metropolis-Hastings algorithm. Am Statistician 49:327–335.

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

Devroye L (1986) Non-uniform random variate generation. Springer-Verlag, New York, pp 843–844

Fernando RL, Grossman M (1989) Marker-assisted selection using best linear unbiased prediction. Genet Sel Evol 21:467–477

Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences (with discussion). In: Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds) Bayesian statistics 4. Clarendon Press, Oxford, pp 457–511

Geyer CJ (1991) Reweighting Monte Carlo mixtures. Technical Report No. 568, School of Statistics, University of Minnesota

Gilks WR, Best NG, Tan KKC (1995) Adaptive rejection Metropolis sampling within Gibbs sampling. Appl Statist 44:455–472

Gilks WR, Wild P (1992) Adaptive rejection sampling for Gibbs sampling. Appl Statist 41:337–348

Goddard M (1992) A mixed model for analyses of data on multiple genetic markers. Theor Appl Genet 83:878–886

Guo SW, Thompson EA (1992) A Monte Carlo method for combined segregation and linkage analysis. Am J Hum Genet 51:1111–1126

Hasstedt SJ (1993) Variance components/major locus likelihood approximation for quantitative, polychotomous, and multivariate data. Genet Epidemiol 10:145–158

Hobert JP, Casella G (1994) Gibbs sampling with improper prior distributions. Technical Report BU-1221-M, Biometrics Unit, Cornell University, Ithaca, New York

Hoeschele I (1989) A note on local maxima in maximum likelihood, restricted maximum likelihood, and Bayesian estimation of variance components. J Statist Comp Simul 33:149–160

Hoeschele I (1994) Bayesian QTL mapping via the Gibbs sampler. Proc 5th World Congr Genet Appl Livst Prod, Guelph, Canada 21:241–244

Hoeschele I, VanRaden PM (1993a) Bayesian analysis of linkage between genetic markers and quantitative trait loci. I. Prior knowledge. Theor Appl Genet 85:953–960

Hoeschele I, VanRaden PM (1993b) Bayesian analysis of linkage between genetic markers and quantitative trait loci. II. Combining prior knowledge with experimental evidence. Theor Appl Genet 85:946–952

Janss LLG, Thompson R, van Arendonk JAM (1995) Application of Gibbs sampling in a mixed major gene – polygenic inheritance model in animal populations. Theor Appl Genet 91: 1137–1147

Kass RE, Raftery AE (1995) Bayes factors. J Am Statist Assoc 90:773–795

Meng X-L, Wong WH (1993) Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. Technical Report No. 365, Department of Statistics, The University of Chicago

Newton MA, Raftery AE (1994) Approximate Bayesian inference with the weighted likelihood bootstrap. J Roy Statist Soc B 56:3–48

Satagopan JM, Yandell BS, Newton MA, Osborn TC (1996) Markov chain Monte Carlo approach to detect polygene loci for complex traits. Genetics (in press)

Scott WD (1992) Multivariate density estimation. John Wiley and Sons, New York

Smith AFM, Roberts GO (1993) Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods. J Roy Statist Soc B 55:3–24

Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation. J Am Statist Assoc 82:528–540

Thaller G, Hoeschele I (1996) A Monte Carlo method for Bayesian analysis of linkage between single markers and quantitative trait loci. II. A simulation study. Theor Appl Genet (in press)

Thaller G, Dempfle L, Hoeschele I (1996) Maximum likelihood analysis of rare binary traits under different modes of inheritance. Genetics 143: 1819–1829

Thomas DC, Cortessis V (1992) A Gibbs sampling approach to linkage analysis. Hum Hered 42:63–76

Thompson EA, Guo SW (1991) Evaluation of likelihood ratios for complex genetic models. IMA J Math Appl Med Biol 8:149–169

Uimari P, Thaller G, Hoeschele I (1996) The use of multiple linked markes in a Bayesian method to map quantitative trait loci. Genetics 143:1831–1842

Wang CS, Rutledge JJ, Gianola D (1993) Marginal inferences about variance components in a mixed linear model using Gibbs sampling. Genet Sel Evol 25:41–62

Zeng Z-B (1994) Precision mapping of quantitative trait loci. Genetics 136:1457–1468