Journal of Animal and Feed Sciences, 15, 2006, 685–694

Markov Chain Monte Carlo sampling on multilocus genotypes

M. Szydłowski[1]

The August Cieszkowski Agricultural University of Poznań,
Department of Genetics and Animal Breeding
Wołyńska 33, 60-637 Poznań, Poland

(Received 1 February 2006; revised version 9 October 2006; accepted 6 November 2006)

ABSTRACT

Markov Chain Monte Carlo (MCMC) methods are used to solve complex problems in animal genetics. The MCMC samplers may mix slowly, making computation impractical. In this paper the behaviour of the whole locus sampler (L-sampler) in analysis of multilocus data was examined. To evaluate the mixing we monitored estimates for number of genes shared identical by descent between relatives. It was demonstrated in simulation study that linkage between loci may drastically reduce the efficiency of the L-sampler, leading to incorrect inference. Two samplers were considered to improve mixing of Markov chain: the multimeiosis sampler (MM-sampler) and the multilocus sampler (ML-sampler). It was concluded that MM- and ML-samplers improve mixing but do not guarantee practical irreducibility. A situation causing bad mixing was identified and some tips to tackle the problem were given.

KEY WORDS: multilocus linkage analysis, Markov Chain Monte Carlo, MCMC samplers, identity by descent

INTRODUCTION

Markov Chain Monte Carlo (MCMC) methods have proven useful in solving complex genetic problems. They were used in major gene detection (Janss et al., 1995), QTL mapping (George et al., 2000), candidate gene analysis (Stachowiak et al., 2006) and marker assisted genetic evaluations (Liu et al., 2004). The method provides solution by multiple sampling of unknown parameters and missing data. To solve a genetic problem it is usually needed to generate multiple samples of

[1] Corresponding author: e-mail: mcszyd@jay.au.poznan.pl

genotype configuration in a pedigree. The sampled genotypes should be consistent with the collected pedigree and genotype records and the assumed genetic model.

Only irreducible Markov chain can be used for statistical inference. The genotype sampler is theoretically irreducible, if from all genotype configurations, the chain can reach any other genotype configuration with positive probability, in some number of iterations. There are a few samplers which can be used to construct irreducible multilocus sampling scheme (Kong, 1991; Thompson and Heath, 1999). Although theoretically irreducible, the chain may still mix slowly and disables practical computation. The chain mixes slowly if consecutive samples are highly correlated. This problem is more severe for tightly linked loci. An efficient sampler that produces samples with low autocorrelation is important for future application of the MCMC methods in genetics.

The objective of this study was to examine behaviour of some multilocus genotype samplers in the analysis of linked markers. The impact of linkage between markers on the convergence of Markov chain was evaluated. Two multilocus samplers are proposed and their ability to improve mixing is tested.

MATERIAL AND METHODS

The study was based on simulated data sets. The data sets were analysed by the use of various samplers. Four basic samplers and 2 combinations of the basic samplers (hybrid samplers) were used.

L-sampler

The L-sampler produces joint sample of genotypes at single locus for entire pedigree (Kong, 1991). The L-sampler can be applied to each locus in turn to form theoretically irreducible Markov chain on multilocus genotypes. Other single-locus updates can be treated as the reduction of the L-sampler. The L-sampler uses generalised Elston-Stewart algorithm (*pedigree peeling algorithm*) for calculation of genotype probabilities for each pedigree member in given sequence (Cannings at al., 1978). The calculation for an individual takes into account the likelihood of individual's own record and the records for the pedigree members preceding the individual in the sequence. In this way the genotype probabilities for the last individual take into account all data and the entire pedigree is *peeled* (evidence collection step). The sampling step (random propagation step) is conducted backward starting from the last individual peeled and conditioning on already sampled genotypes. The peeling process is relatively fast method for probability calculations on pedigree but still may be computationally demanding. The speed depends on the structure of the pedigree and the number of missing genotype records. The L-sampler can be modified to sample multilocus genotypes, however, the computational burden grows exponentially with the number of loci.

M-sampler

Thompson and Heath (1999) developed the whole meiosis sampler (M-sampler) that produces joint sample of a set of segregation indicators on single chromosome. A segregation indicator describes the origin of a gene and it is zero if the copied gene is parent's maternal gene and one if it is parent's paternal gene. When applied iteratively to all meioses (chromosomes), the M-sampler forms a tool to sample multilocus descent graphs (Lange and Sobel, 1991). In the presence of three or more alleles per locus transition from one valid descent graph to another may require simultaneous updates at more than single meiosis. This is what makes the M-sampler reducible. The M-sampler relies on the conditional independence structure of the segregation indicators, which permits the summation to be performed sequentially along the chromosome (*chromosomal peeling*): such an algorithm was developed by Lander and Green (1987). In the chromosomal peeling, computation is linear in the number of loci and exponential in the number of meioses.

ML-sampler

The pedigree peeling enables simultaneous updates of genotypes on entire pedigree. The computation is exponential in the number of loci. Hence, simultaneous update on multilocus genotypes for whole pedigree and all loci is not possible. Two-locus sampling may be practical on small pedigrees. The two-locus sampler would solve problem of two tightly linked genes located in long distance from surrounding markers. In a large pedigree the multilocus updates can be made for a part of the pedigree - a block of close relatives, fixing the surrounding genotypes. This sampler is referred to as the multilocus sampler (ML-sampler). We used the ML-sampler for two tightly linked loci, a situation typical for experiments on large populations with a few carefully selected markers. Picking a random individual from the pedigree with its parents, offspring and spouses formed a block of individuals for simultaneous updates with the ML-sampler. In each iteration of the MCMC algorithm the multiple random blocks of individuals were updated. Some individuals occurred in multiple blocks.

MM-sampler

The multimeiosis sampler (MM-sampler) is based on the idea of the M-sampler and enables simultaneous updates of segregation indicators for all loci and multiple meioses. Since the sampler uses the chromosomal peeling, the computation grows rapidly with the number of meioses, making simultaneous updates of more than a few meioses impractical. In our implementation of the MM-sampler, the block of meioses was random: starting from an initial pair of paternal and maternal meioses (vectors of paternal and maternal segregation indicators for a single offspring), the block was stretched out to meioses for randomly chosen close relatives. Next meioses entered the block until the total number of possible realizations of segregation indicators across all loci did not exceed the preset limit.

Hybrid samplers

Two hybrid samplers were considered: the first one was the combination of the L- and MM-updates, and the other one was the combination of the L- and ML-updates. The L-sampler was used in both hybrid samplers to ensure theoretical irreducibility of the chains.

To examine the impact of linkage on the behaviour of the L-sampler, multiple data sets were simulated. The pedigree used for the simulation consisted of 107 individuals and four generations (Figure 1). Three series of data sets were generated with the markers spaced every 30, 10 and 5 cM. For comparison, an additional series of data sets was simulated under no linkage. Each series consisted of 1000 data sets. Three codominant markers were simulated with the number of alleles ranged between 2 and 15. Drawing alleles from the uniform distribution produced the genotypes of the founders. Dropping genes down the pedigree assuming various distances between loci produced genotypes of the non-founders. The simulated records were made available only for a part of the pedigree, mainly for young individuals. The genotypes at a locus were known or completely

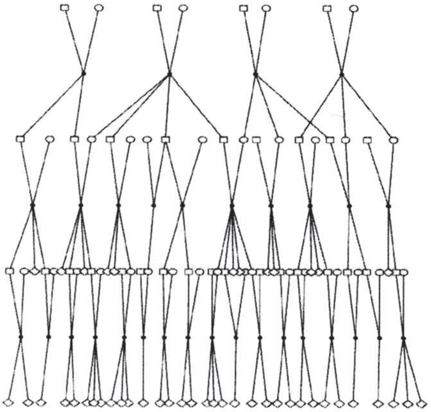

Figure 1. Pedigree structure used for simulation

deducible for about half of the pedigree. The hybrid sampler formed by the L- and MM-updates was examined only on 10 data sets from the 5 cM series that caused the largest problem for the L-sampler. The hybrid sampler that used the combination of the L- and ML-updates was used for two-locus problem on large complex pig pedigree. The pedigree consisted of 2380 individuals in multiple generations and included many loops. Two codominant markers of 15 and 2 alleles were simulated in the genetic distance of 1 cM. The simulated records were made available for 1782 animals in more recent generations. Within the pedigree, 1359 animals had their genotype known at first locus and 1190 animals were fixed at the other locus.

To examine the mixing behaviour of the samplers we compared summaries calculated from two independent runs of the MCMC samplers. The dissimilarity between summaries indicates bad mixing. The diagnostic summary considered in this study was the number of genes shared identical by descent (IBD) in a pair of relatives. The number of genes shared IBD is important summary in genetic analyses, i.e. homozygosity mapping and some variance components QTL mapping methods. All the data sets were analysed twice using different seeds for the random number generator, and thus, distinct starting points. We ran 10,000 iterations burn-in and then collected 1,000 samples from every 10-th iteration. The allelic frequencies were updated every iteration. The genetic distances between markers were fixed at simulated values. The estimates of the IBD sharing were calculated as the mean values of all samples. For all the data sets the same pair of seeds for random number generator was used, thus, the variation in difference between estimates originated entirely from the data sets. For the pig pedigree the genes shared IBD in both loci were counted.

RESULTS

For the series of the data sets simulated for unlinked markers the L-sampler produced consistent estimates of IBD sharing parameters. Due to small chain size some small discrepancies between estimates from two runs occurred. For 95% of the tests the calculated sum of squared differences between the estimates was below 0.31, and for 99% of the tests the statistics was below 0.46. Table 1 shows the results for the 30 cM, 10 cM and 5 cM series. Clearly, close linkage

Table 1. The mixing problem of the L-sampler - the percentage of the MCMC tests within three series

Sum of squared differences[1]	Unlinked markers	Genetic distance		
		30 cM	10 cM	5 cM
	%			
<0.31	95	94	58	47
<0.46	99	97	60	53

[1] 95 and 99% thresholds under no linkage

ruins the performance of the L-sampler. For some pairs of relatives extremely different estimates for IBD sharing were calculated. The number of the tests with at least one such pair was 4 in the 30 cM series, and 88 and 218 in the 10 cM and 5 cM series, respectively. Such dissimilarity in estimates is produced when two chains are stuck at distinct initial values.

Ten the most difficult data sets from the 5 cM series were reanalysed by the hybrid sampler that used the L- and MM-update steps alternately. The MM-sampler improved mixing, but did not solve all problems. In six of the 10 tests, the sums of squared differences dropped to negligible level, strongly suggesting good mixing. For the rest, the statistics drooped, however, several pairs of relatives with very inconsistent estimates for numbers of genes shared IBD still occurred.

The behaviour of the hybrid sampler that used the L- and ML-samplers was examined on large complex pedigree. The ML-sampler showed its capability of improving mixing, at least in cases similar to the one considered here. Figure 2 presents the plots of IBD sharing estimates from two runs. The use of the ML-sampler led to similar mixing behaviour as expected under no linkage model. The two-locus updates cost as much time as the L-updates at two loci.

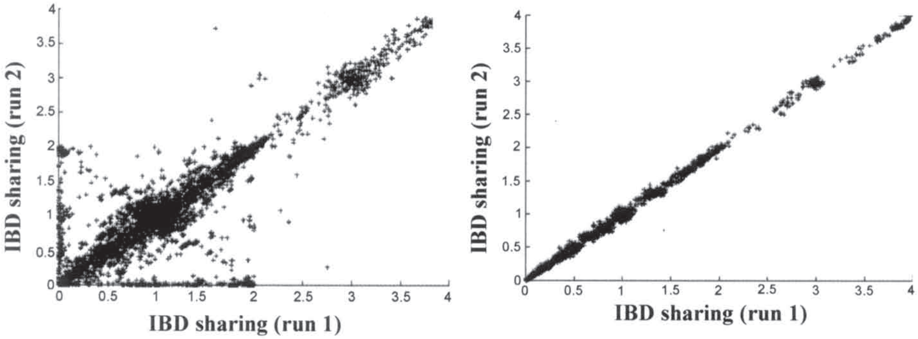

Figure 2. Plots of the estimates for the number of genes shared IBD in two closely linked loci in large pedigree analysed with the L-sampler applied on whole pedigree (left) and the combination of the L- and ML-samplers (right). The position of each point is determined by two estimates from the independent MCMC runs

DISCUSSION

It was shown that the mixing problem could be common difficulty in MCMC analysis of multilocus pedigree data. The problem is not restricted to close linkage nor to complex pedigrees. The availability of dense genetic maps makes the problem even more common. The tendency to offer sampling algorithm in easy-to-use computer package makes it easy to oversee the problem or even worse to ignore it. Thus, similar user-friendly convergence checking programs are required.

When various pedigrees were considered, the mixing problem appeared more often in complex structures. The simplification made in pedigree structure may improve mixing. A complex pedigree can be simplified by breaking the inbreeding loops and making the pedigree structure acyclic. Ignoring some relationships and splitting a pedigree to a set of the nuclear families make further simplifications. Such modifications on pedigree makes also computation faster and more samples possible. This study shows, however, that even zero-loop pedigree can cause mixing problem and the simplification of the pedigree structure does not guarantee improvement in mixing. Moreover, the simplification of pedigree structure has the disastrous effect of reducing the power of linkage analysis (Dyer et al., 2000).

The pairs of relatives with the most inconsistent estimates of the IBD sharing were identified. Most often these pairs were of grandparent-grandchild relationship. Although other types of relationship were found as well, the problem always stemmed from the particular combination of data for parent and offspring. As could be expected the mixing problem occurred when two genotype configurations communicate through the other one of very small probability. For linked loci the least probable states are those with many recombinants. Figure 3 exemplifies this situation.

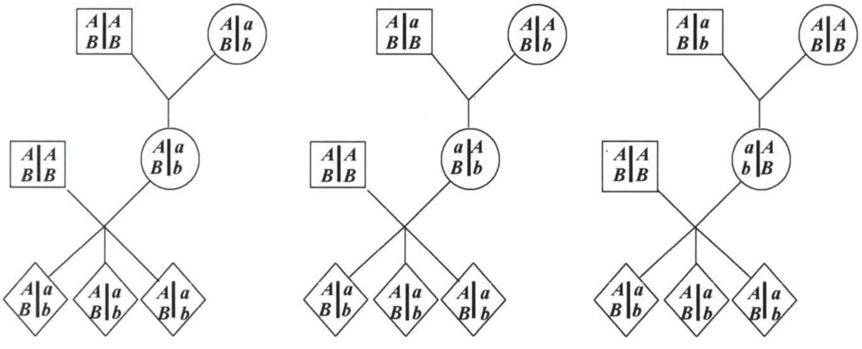

Figure 3. Example of three communicating genotype configurations (states) for small pedigree. If the L-sampler is used, the transition from state 1 (left) to state 3 (right) is achieved *via* state 2 (middle) with three recombination events involved. For linked loci, the situation can cause practical reducibility

The example presents three genotype states on a small pedigree. The genotypes of the father and three children are fixed and the mother is known to be doubled heterozygote. Considering the L-sampler, the transition from the state 1 to 3 is achieved through the state 2 and requires three recombination events. If two loci are linked, the recombination rate (r) between loci is lower than 0.5. In case of linkage, the state 2 has smaller probability (r^3) than the two other states, both with the probability $(1-r)^3$. For closely linked loci or large number of recombinants required in the state 2, this situation makes the chain frozen at state 1 or 3. In consequence, the IBD sharing between grandparent and each of three grandchildren would be estimated close to 1 or 0.

Several implications come from this example. First, one can predict the situations where bad mixing should be anticipated - the larger fullsib/halfsib groups are the more recombinants have to appear to spin the two genotype configurations. In consequence, for a large group of siblings even weak linkage suffices to cause severe mixing problems. Not surprisingly, many animal pedigrees with large number of offspring per sire have been found to cause mixing problems. Next, mixing can be much worse at flanked locus than at terminal loci. This is because the recombination events have to appear in both sides of the flanked gene. Another important issue concerns the M-sampler. In our example, the segregation indicators for the dam and offspring are fully correlated and must be sampled jointly. This cannot be obtained by the use of the M-sampler, which operates on single meiosis. Here the M-sampler is simply reducible, regardless of linkage between loci. Although Thompson and Heath (1999) pointed out that the M-sampler is reducible, it is worth remembering that its use cannot improve mixing in quite common situation considered here.

There is no golden rule for efficient sampling on multilocus genotypes. Some general principles, however, might help to avoid many problems. Firstly, it is strongly recommended to perform simulations from different starting points rather than to produce a single long chain. This rule has been advised by many authors (Gilks et al., 1996). Although the repeated warms-up is time consuming, multi-chain analyses should produce more reliable estimates. Secondly, it is reasonable to use a combination of different sampling tools (hybrid sampler) rather than a single sampler. As it was shown in this paper, the hybrid use of the L- and ML-updates has the ability to improve mixing, although it does not guarantee improvement in all cases. Faster samplers, similar to the MM-sampler considered here, have been proposed (Thomas et al., 2000). Instead of considering all combinations of segregation indicators, the samplers switch between alternative configurations in a set of loci in two or three generation families. Although less general, the samplers can provide fast efficient multilocus updates on small pedigrees. In a large pedigree, like the pig pedigree considered in this paper, they may fail to improve mixing considerably. This is in accordance to the results by Thomas et al. (2000). Another hybrid sampler that combines the M-sampler and the random walk approach was considered by

Lee et al. (2005). Their method allows use of a wider range of data for mapping of QTL, however, there still can be reducibility problems in some cases, e.g., in large half-sib families.

There is an obvious need to develop more efficient sampling tools to tackle multilocus problems. To do this, the reduction of genotype space is important task. The allele set-recoding method is important to reduce genotype space (O'Connel and Weeks, 1995). Further reduction is possible by removing final offspring with no records for each locus separately, and fixing genotypes of founders and segregation indicators of children where possible.

A flexible multilocus algorithm for sampling as much of genotypes as possible, given the computer and time limits, can be useful. Starting from a randomly chosen family and a set of consecutive loci, one can stretch out the sampling block on both surrounding individuals and loci, keeping the required computation time below some preset limit. After the extension step, the surrounding space is fixed at the current values and genotype elimination is redone for all individuals in the block. The extension step and genotype elimination step continues until the limits are reached. Next, the sampler updates all parameters for entire block and moves to next block. Note, when the block is small, it is possible to use more powerful technique for genotype elimination than the one iterating on nuclear families.

This flexible blocked Gibbs sampling in combination with the L-updates ensures the irreducibility (not practical) of the Markov chain. The important question to answer is about the rules to construct a block. In our study fixed block worked better than variable blocks. Fixed blocks make also a potential algorithm simpler and faster. Certainly, there is no need to sample the same number of loci for all individuals in the block, the problem is rather to find optimal sampling subspace on multilocus genotype configuration. Hopefully, it is possible to find subspaces of relatively small size enabling fast and efficient sampling.

MCMC methods are frequently proposed to solve multilocus problems in animal genetics and breeding, e.g., fine mapping of quantitative trait loci and marker assisted evaluation. Although computationally attractive the methods need significant improvements. In this paper two samplers were considered to improve the method when applied to linked loci. It was shown that hybrid use of different samplers may improve effectiveness of sampling on multilocus genotype space. However, a reliable genotype sampler for linked loci and large pedigrees is still missing. Further studies are encouraged to develop flexible multilocus samplers.

REFERENCES

Cannings C., Thompson E.A., Skolnick M.H., 1978. Probability functions on complex pedigrees. Advan. Appl. Probab. 10, 26-61

Dyer T., Williams J.T., Gőring H.H.H., Blanger J., 2000. The effect of pedigree complexity on quantitative trait linkage analysis. Genetic Analysis Workshop 12, Vol. 1: Asthma Data. Participant Contributions. San Antonio, Texas (USA)

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. Theor. Appl. Genet. 91, 1137-1147

George A.W., Visscher P.M., Haley C.S., 2000. Mapping quantitative trait loci in complex pedigrees: a two step variance component approach. Genetics 156, 2081-2092

Gilks W.R., Richardson S., Spiegelhalter D.J., 1996. Markov Chain Monte Carlo in practice. Chapman & Hall, London (UK)

Kong A., 1991. Analysis of pedigree data using methods combining peeling and Gibbs sampling. Computer Science and Statistics: In: Proceedings of the 23rd Symposium on the Interface, pp. 379-385

Lander E.S., Green P., 1987. Construction of multilocus linkage map in humans. Proc. Nat. Acad. Sci. USA 84, 2363-2367

Lange K., Sobel E., 1991. A random walk method for computing genetic location scores. Amer. J. Hum. Genet. 49, 1320-1334

Lee S.H., Van der Werf J.H.J., Tier B., 2005. Combining the meiosis Gibbs sampler and the random walk approach for linkage and association studies with a general complex pedigree and multimarker loci. Genetics 171, 2063-2072

Liu Z., Reinhardt F., Szyda J., Thomsen H., Reents R., 2004. A marker assisted genetic evaluation system for dairy cattle using a random QTL model. In: Proceedings of the 2004 INTERBULL Meeting, Sousse (Tunisia). Bull. 32, 170-174

O'Connel J.R., Weeks D.E., 1995. The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. Nat. Genet. 11, 402-408

Stachowiak M., Szydlowski M., Obarzanek-Fojt M., Switonski M., 2006. An effect of a missense mutation in the porcine *melanocortin-4 receptor* (*MC4R*) gene on production traits in Polish pig breeds is doubtful. Anim. Genet. 37, 55-57

Thomas A., Gutin A., Abkevich V., Bansal A., 2000. Multilocus linkage analysis by blocked Gibbs sampling. Stat. Computing 10, 259-269

Thompson E.A., Heath S.C., 1999. Estimation of conditional multilocus gene identity among relatives. Statistics in Mol. Biol. IMS Lecture Notes - Monograph Ser. 33, 95-113