Proc. Assoc. Advmt. Anim. Breed. Genet. ***16****: 107-110*

COMBINING TWO MARKOV CHAIN MONTE CARLO APPROACHES FOR LINKAGE AND ASSOCIATION STUDIES WITH A COMPLEX PEDIGREE AND MULTI MARKER LOCI

S.H. Lee[1], J.H.J. Van der Werf [1] and B. Tier[2]

[1]School of Rural Science and Agriculture, University of New England, Armidale, NSW2351
[2]Animal Genetics and Breeding Unit[*], University of New England, Armidale, NSW 2351

SUMMARY

In QTL mapping using linkage and/or linkage disequilibrium, an important process is to find the pattern of inheritance states and haplotype configurations, a process known as haplotype reconstruction. Haplotype reconstruction is routinely based upon observed pedigree information and marker genotypes for individuals in the pedigree. It is not feasible for the exact methods to use all such information for large complex pedigree especially when there are many missing genotypes. Markov Chain Monte Carlo (MCMC) approaches have been widely used to handle a complex pedigree with sparse genotypic data. However they often have reducibility problems or are slow to converge. Combining two different MCMC approaches results in improvement of computational speed and mixing properties. It allows obtaining reliable estimates such as identity by descent coefficients between individuals within a reasonable time.

INTRODUCTION

Finding the pattern of inheritance states and haplotype configurations is an essential step for linkage and linkage disequilibrium (LD) mapping. Inheritance states are used to predict pedigree-based identity by descent (IBD) coefficients arising through segregation in a recorded pedigree. Haplotype configurations allow LD-based IBD coefficients to be estimated among ancestors beyond the recorded pedigree.

Haplotype reconstruction is necessarily based on the observed pedigree and marker data. This can cause difficulties as the pattern of inheritance states is hard to derive especially with typical complex pedigrees and many missing genotypes. Exact likelihood methods such as pedigree peeling (Elston and Stewart 1971; Cannings et al. 1978) or chromosome peeling (Lander and Green 1987) increase exponentially in computational complexity with the number of markers or the number of pedigree members. Furthermore, missing genotypic data severely affects the computational task in exact likelihood methods.

Markov chain Monte Carlo (MCMC) algorithms are an alternative and flexible method to reconstruct haplotypes and estimate IBD probabilities. These algorithms make it possible to deal with complex pedigrees and many missing genotypic data (Sheehan et al. 1989; Lange and Matthysse 1989; Thompson 1994). However, reducible sites often occur and mixing problems also appear in using multiple marker loci (Thompson and Heath 1999; Canning and Sheehan 2002). By updating segregation indicators jointly for all marker loci in a single meiosis, the meiosis Gibbs sampler (Thompson and Heath 1999) greatly improves mixing of the Markov chain with multiple markers.

[*] AGBU is a joint venture of the NSW Depatment of Primary Industries and the University of New England

However there are still significant reducibility problems (Thompson and Heath 1999; Heath 2003). A different MCMC method, called the random walk approach (Sobel and Lange 1996), uses multiple random moves to update variables, which allow the chain to pass through illegal states between legal states. This can remedy reducibility problems. However, illegal or less likely configurations are often proposed, which are mostly rejected by a Metropolis mechanism (Metropolis et al. 1953). Therefore the computational efficiency of the random walk approach is much less than that of the meiosis Gibbs sampler where updated variables are always accepted.

We propose a sampler which combines the meiosis Gibbs sampler and the random walk approach. In the combined sampler, the meiosis sampler is firstly used for all sites. Subsequently, if potential reducible sites are detected, the random walk approach is applied to those sites. Therefore the combined sampler should be much faster than the random walk approach and there should be no or few reducible sites.

MATERIALS AND METHODS

Posterior distribution of segregation states. Probability of one realization of segregation states (*S*), given marker data, can be derived from (1).

$$pr(S|G) = \frac{pr(G|S)pr(S)}{\sum pr(G|S)pr(S)} \quad (1)$$

where *G* represents the observed marker data, *pr(S)* is prior probability of the segregation indicators, *Pr(G|S)* is the probability of the observed marker data given *S*, and the denominator is summed over the probabilities of all possible configurations of *S*. Since the computation of the denominator is infeasible in general pedigrees, a MCMC approach is required to obtain the posterior distribution of the segregation indicators.

Updating schemes for segregation indicators in MCMC cycles. The meiosis sampler is firstly applied to all loci for every individual. During the meiosis sampler, it is possible to detect potential reducible sites. After a cycle of the meiosis sampler, a random walk is carried out for the potential reducible sites that were never updated in the meiosis sampler. After enough moves of the random walk (e.g. number of moves ~ number of meioses x number of markers), all reducible sites have equal chance to be updated and they can have new variables. Note that the smaller the number of reducible sites, the faster the combined method.

Initial legal configuration for the Markov chain. A MCMC approach requires a starting configuration, consistent with observed marker data. The genotype elimination through inheritance constraint (GEIC) algorithm (Henshall et al. 2001) is suitable for finding a legal configuration of segregation indicators at each locus.

IBD probabilities based on LD and linkage information: IBD probabilities between all members are estimated based on LD and linkage in each sampling round. Sampled haplotypes for base animals are used to estimate LD-based IBD probabilities between unrelated base animals, using the method of Meuwissen and Goddard (2001). Sampled segregation indicators at multiple loci for descendants are used to estimate IBD probabilities between relatives given LD-based IBD probabilities of base animals (Lee and van der Werf 2005).

Table 1. Correlation and likelihood as the accuracy for each MCMC method with a pedigree spanning 5 generations (*Ne* = 20)

time (sec)	1	4	16	64	256
	Complete genotypic data				
	Correlation (standard error)				
RA[a]	0.837 (0.009)	0.936 (0.007)	0.983 (0.002)	0.995 (0.001)	0.997 (0.001)
MS[b]	0.842 (0.009)	0.861 (0.009)	0.864 (0.008)	0.864 (0.008)	0.865 (0.008)
RAMS[c]	0.871 (0.008)	0.939 (0.006)	0.974 (0.005)	0.991 (0.002)	0.997 (0.001)
	Likelihood				
RA	-1335.154	-794.044	-567.468	-536.153	-529.245
MS	-1098.908	-1093.843	-1090.619	-1090.159	-1092.461
RAMS	-925.292	-736.479	-603.849	-551.810	-524.640
	Incomplete genotypic data				
	Correlation (standard error)				
RA	0.821 (0.011)	0.900 (0.011)	0.941 (0.008)	0.953 (0.008)	0.970 (0.008)
MS	0.930 (0.008)	0.949 (0.007)	0.957 (0.008)	0.961 (0.007)	0.963 (0.007)
RAMS	0.938 (0.01)	0.961 (0.007)	0.975 (0.004)	0.984 (0.003)	0.988 (0.003)
	Likelihood				
RA	-812.550	-489.010	-318.200	-271.709	-215.193
MS	-330.153	-275.463	-248.923	-244.298	-234.314
RAMS	-316.865	-246.756	-208.334	-182.298	-169.921

[a]RA: random walk approach, [b]MS: meiosis Gibbs sampler, [c]RAMS: combined sampler

Simulation study

An effective population size of 100 was simulated for 100 generations for 10 bi-allelic or multi-allelic marker loci at 1 cM intervals, based on Mendelian segregation using the gene-dropping method (MacCluer et al. 1986). This simulation model ensured that the population would have an equilibrium distribution of alleles in all loci. Note that pedigree information is not available for these 100 generations. At generation 101, a population of size *Ne* was simulated for t generations with pedigree recording. In each generation, the number of male and female parents was Ne/2 and they were randomly mated with a total of 2 offspring for each of *Ne*/2 mating pairs. Therefore, the recorded pedigree had complex relationships between animals with a value of $t > 2$.

Complete or incomplete genotypic data were used to investigate the efficiency of three approaches, i.e. the random walk, the meiosis sampler and the combined method. In complete genotypic data, genotypes were available for all pedigreed individuals. In incomplete genotypic data, genotypes were available for progeny in the last generation (ancestral and parental genotypes were all missing but their relationships were used).

The correlation between true IBD probabilities based on true haplotypes and estimated IBD probabilities using the random walk approach (RA), the meiosis sampler (MS) or the combined method (RAMS) was used as the accuracy of each method. The mean and standard error of

correlations over 10 replicates are tabulated against the time spent for sampling segregation indicators.

RESULTS AND DISCUSSION

Table 1 shows correlation between true and estimated IBD probabilities and the highest likelihood among inheritance states sampled, using RA, MS and RAMS with a pedigree spanning 5 generations (N_e =20). In complete genotypic data, the correlation (i.e. the accuracy of IBD estimates) for RA and RAMS are reasonably high and similar to each other after 256 sec. However that for MS is much lower. The likelihood of inheritance state using RA and RAMS gradually increases. However that for MS hardly improves. This is probably due to reducibility problems in MS. With incomplete genotypic data, the accuracy and likelihood for RA is much lower than MS or RAMS until 64 seconds. After 256 seconds, the accuracy and likelihood for RA is higher than MS, but lower than RAMS. This is because with incomplete genotypic data, the reducibility problems are less severe than with complete genotypic data because founder allelic types are less constrained (Thompson and Heath 1999).

In conclusion, simulation results show that the combined sampler, compared to the random walk approach and the meiosis sampler, can remedy reducibility problems, can converge quickly with reasonably high accuracy and can find more likely and desirable inheritance states, with complete or incomplete genotypic data.

REFERENCES

Cannings, C., and Sheehan, N. A. (2002) *Genetics* **162**: 993.

Cannings, C., Thompson, E. A. and Skolnick, M. H. (1978) *Adv. Appl. Probab*. **10**: 26.

Elston, R. C., and Stewart, J. (1971) *Hum. Hered*. **21**: 523.

Heath, S. C. (2003) "Highly Structured Stochastic System" 1st ed. Oxford University Press, Oxford.

Henshall, J. M., Tier, B and Kerr, R. J. (2001) *Genet. Res.* **78**: 281.

Lander, E. S. and Green, P. (1987) *Proc. Natl. Acad. Sci. USA* **84**: 2363.

Lange, K. and Matthysse, S. (1989) *Am. J. Hum. Genet*. **45**: 959.

Lee, S. H., and Van der Werf, J. H. J. (2005) *Genetics* **169**: 455.

MacCluer, J. W., Van der Berg, J. L., Read, B. and Ryder, O.A. (1986) *Zoo Biology* **5**: 147.

Meuwissen, T. H, E. and Goddard, M. (2001) *Genet. Sel. Evol.* **33**: 605.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953) *Chem. Phys.* **21**: 1087.

Sheehan, N. A., Possolo, A. and Thompson, E. A. (1989) *Am. J. Hum. Genet.* **45** (Suppl.) A248.

Sobel, E. and Lange, K. (1996) *Am. J. Hum. Genet.* **58**: 1323.

Thompson, E. A. (1994) *Stat. Sci.* **9**: 355.

Thompson, E. A. and Heath, S. C. (1999) "*Statistics in Molecular Biology and Genetics"* Institute of Mathematical Statistics, American Mathematical Society, Providence, RI.