A BAYESIAN APPROACH TO ORDERING GENE MARKERS

A.W. George[1], K. L. Mengersen[1] and G. P. Davis[2]

[1] School of Mathematics, Queensland University of Technology, Brisbane, QLD, 4067
[2] CSIRO Tropical Agriculture, University of Queensland, Brisbane, QLD, 4170

SUMMARY

Gene markers have traditionally been ordered using Maximum Likelihood (ML) techniques. In this paper a Bayesian alternative for estimating the recombination rates and ordering markers is presented and is implemented via a Markov chain Monte Carlo (MCMC) algorithm. By focusing on a half-sib design for which there is a class of missing data, namely the dam genotypes, we illustrate the improved parameter inference and ease with which common complexities in design are accommodated. Results are presented from the analysis of half-sib data where information is available on the gene markers CSSM24, CSSM5, CYP21, RM33 and RM185. Using our methodology to order the markers and estimate the associated parameters, the marker ordering CSSM5 RM33 CYP21 RM185 CSSM24 has a posterior probability of 0.97 and the mean recombination rates were estimated to be 0.106, 0.168, 0.059, 0.167. This agrees with the International Reference Family map.

Keywords: Bayesian analysis, gene markers, linkage analysis

INTRODUCTION

Gene markers and gene marker maps are being used in animal species for parentage testing (Moore and Vankan, 1994), disease resistance diagnosis (Fuji *et al.*, 1991) and quantitative trait loci (QTL) detection (Georges *et al.*, 1995; Anderson *et al.*, 1994) These maps are a representation of the linear order of the genes along a chromosome with the distances between adjacent genes proportional to the frequency of recombination between them. The closer two genes are together, the less likely a recombination event is to occur. Thus, the frequency of recombination between genes serves as a measure of genetic distance and is the basis of gene marker map construction.

Gene marker maps rely heavily on accurate estimates of marker order, which are ideally obtained using three generation pedigrees with informative grandparents, genotypes from all individuals and large full sib family size. Such a structure is difficult and expensive to create in some species, particularly cattle, and has led to the creation of separate reference and resource families used for gene mapping and QTL detection respectively (Hetzel, 1989). Resource families are often designed so that an entire class of animal is not genotyped. In the commonly used half-sib and granddaughter designs, for example, genotype information on the dams is not collected and is consequently regarded as missing.

In this paper, we illustrate a Bayesian approach to ordering markers for half-sib data. The approach presented adopts MCMC in order to perform the required computations.

METHODS

Consider a large half-sib family with a single male parent mated to a large number of females, each of which produces a single offspring. N markers are genotyped on the male parent (sires) and all progeny but there is no such information on the female parents (dams). It is assumed recombination events are independent and markers in the dam population are in linkage equilibrium.

ML has commonly been used for linkage analysis of markers for making maps (Ott, 1991), however, the published ML method for the half-sib design is based on the assumption that the allele frequencies are known in the ungenotyped population (Georges *et al.*, 1995). This technique is very sensitive to this assumption. Furthermore, only point estimates of recombination rates between markers and standard errors are obtained from this approach.

We circumvent these limitations by formulating the problem through a Bayesian paradigm and applying MCMC methods to the joint posterior density. Under the Bayesian approach there is no assumption regarding allele frequencies. Also the marginal posterior density for each of the parameters is obtained and a wide range of inferences can be elicited based on realisations from these distributions.

MCMC techniques are a Pandora's box of sample based methods specifically designed to sample complicated distributional forms. In the context of a Bayesian treatment of the gene ordering problem, parameter realisations are obtained by repeatedly sampling the joint posterior density where the joint posterior is the product of the likelihood and the priors. The MCMC algorithm used in this paper consists of two tasks, parameter estimation and model selection.

The first task, parameter estimation utilises the Metropolis-Hastings algorithm (Hastings, 1970) to realise parameter values for the recombination rates and the allele frequencies. These parameter values form a Markov chain whose equilibrium distribution is the parameter's marginal posterior density. By obtaining the marginal posterior density, a large range of inferences are possible eg. means, modes, probability statements, credible intervals.

The second task, model selection, utilises the reversible jump sampler (Green, 1996) which allows transitions between models with different marker orderings and marker phase configurations. Based upon the frequency with which a model is visited, posterior model probabilities are easily obtained.

A full description of the methodology is beyond the scope of this paper but can be found in George *et al.* (1996).

The data analysed form part of the gene marker database on the CBX experiment which aims to detect QTL for carcass and meat quality traits (Hetzel *et al.*, 1997). Five markers from Chromosome 23 were genotyped on three families in the CBX (CSSM24, CSSM5, CYP21, RM33, RM185). The largest family provided 124 progeny with complete data. The analysis was used to order the markers and estimate recombination rates and allele frequencies in the dam population.

The chain was run for 60000 iterations with a burn-in length of 2000. The burn-in is the number of iterations discarded from the beginning of the chain to ensure the algorithm is sampling from the correct distribution. The chain's run length was determined through the convergence diagnostics of Raftery and Lewis (1992) and Heidelberger and Welch (1983). Because of significant autocorrelations up to lag 5, every fifth iteration was retained.

RESULTS

After burn-in, very little movement through the model space is observed. This is due to the small recombination rates between the markers. The strength of the data to identify a single model is influenced by both the marker positions and the size of the family and for this analysis the family size is large. Based on the published gene map (Barendse *et al.*, 1996), the true gene ordering is CSSM5 RM33 CYP21 RM185 CSSM24. The recombination rates between the markers are estimated to be 0.09, 0.20, 0.06 and 0.16. Although information from the published gene map could be used to place priors on the recombination rates and the gene marker order, flat priors are used throughout this paper.

There is close agreement between the published results and results presented here. In Table 1, the published recombination rates, the mean parameter estimates and 95% credible intervals for the recombination rates are given where the marker order is CSSM5 RM33 CYP21 RM185 CSSM24 and the marker phase is 11222/22111. The posterior probability associated with this ordering is 0.97.

Table 1. Published recombination rates, mean parameter estimate and 95% credible intervals (CI) for the recombination rates where the marker ordering is CSSM5 RM33 CYP21 RM185 CSSM24 and the marker phase is 11222/22111

	Recombination rate between markers			
	CSSM5 & RM33	RM33 & CYP21	CYP21 & RM185	RM185 & CSSM24
Published	0.09	0.20	0.06	0.16
Estimated[1]	0.106	0.168	0.059	0.167
CI	(0.0533,0.172)	(0.102,0.244)	(0.020,0.116)	(0.104,0.242)

1 mean parameter estimate

The mean allele frequencies and their 95% credible intervals are given in Table 2.

Table 2. Mean parameter estimates and 95% credible intervals (CI) for the allele frequencies where the marker ordering is CSSM5 RM33 CYP21 RM185 CSSM24 and the marker phase is 11222/22111

	CSSM5		RM33		CYP21		RM185		CSSM24	
Allele	1	2	1	2	1	2	1	2	1	2
Estimated[1]	0.104	0.142	0.347	0.193	0.181	0.177	0.265	0.171	0.170	0.084
CI	(0.055, 0.165)	(0.085, 0.210)	(0.264, 0.434)	(0.128, 0.268)	(0.117, 0.255)	(0.114, 0.251)	(0.188, 0.348)	(0.108, 0.244)	(0.110, 0.240)	(0.041 0.142)

1 mean parameter estimate

DISCUSSION

In this paper a technique has been presented to order N gene markers in a half-sib design. A design which there is an entire class of missing genotype information.

With the proposed formulation, marker order is determined by model selection using a stochastic search algorithm, accompanied by a posterior probability for each possible ordering. In addition, external information on marker orderings, be it partial or complete, can be included in the analysis through prior information.

A full posterior distribution is obtained for each recombination rate and a wide range of inferences can be elicited based on realisations from these distributions. Examples of some inferences are: 95% credible intervals, means, modes, probability statements. The ML technique provides only point estimates of recombination rates between markers and standard errors. These are often obtained with considerable cost or approximation due to the complexity of the likelihood and its different expressions as the marker order changes.

Under a Bayesian approach there is no assumption regarding allele frequencies in the dam population. These are included as variables to be estimated and their precision is also explicitly described.

The technique can readily be adapted to other designs and problems. Currently it is being developed for the detection and localisation of a QTL in a half-sib design. The approach revolves around the mixture model where the QTL is positioned through its linkage with N informative gene markers.

REFERENCES

Anderson, L., Haley, C. S., Ellegren, H., *et al.* (1994) *Science* **263**:1771.
Barendse, W., Vaiman, D., Kemp, S.J., *et al.* (1996) *Mammalian Genome* (in press).
Fugi, J., Otsu, K., Zorzato, F., *et al.* (1991) *Science* **253**: 448.
George, A., Mengersen, K., and Davis, G. (1996) Submitted to *Biometrika.*
Georges, M., Nielsen, D., MacKinnon , M., *et al.* (1995) *Genetics* **139**: 907.
Grenn, P. J. (1996) *Biometrika* **82**: 711.
Hastings, W. (1970) *Biometrika* **57**: 97.
Heidelberger, P. and Welch, P. (1983) *Operations Research* **31**: 1109.
Hetzel, D. J. S. (1989) *Proc. NZ. Soc. Anim. Prod.* **49**: 53.
Hetzel, D. J. S., Davis, G. P., Corbet, N. J., *et al.* (1997) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **12**:442
Moore, S. S. and Vankan, D. (1994) *Australasian Biotechnology* **4**: 107.
Ott, J., (1991) Baltimore: Johns Hopkins University Press.
Raftery, A.L. and Lewis, S. (1992) *Bayesian Statistics 4*,(ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith), pp. 763-74. Oxford University Press.