A survey of current Bayesian gene mapping methods

John Molitor, Paul Marjoram, David Conti and Duncan Thomas*

Department of Preventive Medicine, University of Southern California, 1540 Alcazar Street, CHP-220, Los Angeles, CA 90089-9011, USA

**Correspondence to*: Tel: +1 323 442 1632; Fax: +1 323 442 2349; E-mail: jmolitor@usc.edu

Date received (in revised form): 3rd June 2004

Abstract

Recently, there has been much interest in the use of Bayesian statistical methods for performing genetic analyses. Many of the computational difficulties previously associated with Bayesian analysis, such as multidimensional integration, can now be easily overcome using modern high-speed computers and Markov chain Monte Carlo (MCMC) methods. Much of this new technology has been used to perform gene mapping, especially through the use of multi-locus linkage disequilibrium techniques. This review attempts to summarise some of the currently available methods and the software available to implement these methods.

Keywords: *Bayesian analysis, haplotype, gene mapping*

Introduction

Bayesian methods have become extremely popular in genetic analysis, in part because they allow for the incorporation of background information into the model. The popularity of Bayesian methods may, however, also be due to the ease with which complex likelihoods can be handled through modern computationally intensive Markov chain Monte Carlo (MCMC) techniques.[1] MCMC techniques iteratively update a parameter's value based upon current estimates of values for all other parameters in the model. Likelihoods that can be difficult to estimate jointly can often be handled easily by examining one parameter at a time, conditional on other parameters in the model. While MCMC methods are, by themselves, not Bayesian methods, they are most often utilised in a Bayesian context, as the random nature of parameters in a Bayesian model allow for MCMC methods to be utilised in a natural way. Many excellent introductions to MCMC methods exist.[2]

Much of this powerful Bayesian-based computational machinery has been applied to the field of gene mapping. Marker association studies using single nucleotide polymorphisms (SNPs) are recognised as providing potential for linkage disequilibrium (LD) mapping of genetic polymorphisms contributing to complex traits. Often, association mapping is performed by examining a single-locus model at each candidate marker and then testing the statistical significance of the association at each position. Recently, methods that utilise full haplotype information have been proposed.[3–5] These methods attempt to deal with the complex interplay between markers without explicitly modelling all possible combinations of these interactions. Bayesian methods and MCMC parameter estimation have increasingly been used to formulate and fit these models.

In the remainder of this paper, several current Bayesian gene-mapping methods using multiple markers will be outlined and available software highlighted. Much has been written in this field and, rather than intending this summary to be exhaustive, the authors have instead attempted to illustrate some of the methods that represent major trends in this area. Important related issues — such as haplotype assignment,[6] haplotype tagging of SNPs[7] and the determination of haplotype block structure[8] — will not be emphasised.

LD mapping

LD refers to a non-random association of alleles within haplotypes. It is these associations that are used in gene mapping techniques.[9] Bayesian methods utilise LD through the use of likelihoods that exploit these allelic associations. There are three general approaches to doing this. All of them try to avoid the inadequacy of traditional methods that treat markers as being independently associated with disease. The first approach is to examine the association of continuous sets of markers (ie haplotypes) with disease (see below). In this approach, a complete haplotype is usually treated as the basic unit of interest. Often, the location of a putative disease-causing mutation is used as a point of reference for haplotype risk estimation. Another approach is to examine the association between alleles and disease status but to model dependency between markers using a hierarchical structure (see below). The main difference between these two approaches is that the first approach starts with haplotypes, with all the rich inter-marker dependency haplotypes contain, and then tries to determine the marker — and ultimately the allele — most associated with disease. The second approach models allelic associations directly and then deals with inter-marker dependency at higher levels in the model. A third, arguably more ambitious, approach is to approximate ancestral trees without actually modelling the entire coalescent (see below).

Haplotype methods

A popular Bayesian method for which there is available software is the BLADE algorithm, which is named after the associated paper 'Bayesian Analysis of Haplotypes for Linkage Disequilibrium Mapping', by Liu *et al.*[4] This method explicitly models positions of historical recombination and mutation events based upon an initial set of founders and can deal with missing marker data, multiple founders and unphased chromosomes. This method deals with case-control data and explicitly models ancestral haplotypes on which the original, disease-causing mutations occurred. As with the method of McPeek and Strahs,[3] this method estimates β, the recombination distance from the disease locus to the left-most marker, along with the recombination event to the left and right of the disease locus. Software to implement this method is available at http://fas.harvard.edu/~junliiu/TechRept/03folder/bladev2.tgz. The software is available for Linux on x86 processors and uses a command line interface. The latest version, version 2, allows for inference on phased and unphased haplotypes.[10]

Spatially-based haplotype methods

Spatially-based gene mapping methods are usually based upon the idea that 'similar' haplotypes are likely to carry a common disease-causing variant and hence have the same or similar risk. A similar idea is employed in the field of spatial statistics,[11,12] in which 'regions' (haplotypes in this case) often display some kind of spatial dependence structure, and regions of higher risk are often clustered together.

The key to applying spatial statistical methods to haplotype analysis is to decide upon the distance metric one will use to determine how 'close' one haplotype is to another. In haplotype analysis, the distance metric could be as simple as the proportion of marker loci at which two haplotypes are the same, or the length of the longest contiguous segment over which they are identical by state. Alternatively, if one wanted to estimate the location of a single disease-bearing mutation, one could calculate the length of segment shared by the two haplotypes around the position of the hypothesised mutation.

Thomas *et al.*[13] and Molitor *et al.*[14] used spatial smoothing techniques to perform fine-mapping in a Bayesian context. In order to impose a kind of dependency structure on the haplotype effects so that similar haplotypes are induced to have similar risks, a conditional autoregressive (CAR) prior is used.[15] A matrix of weights is used to indicate the 'closeness' of one haplotype to another, with close haplotype pairs weighted with high values and distant haplotype pairs weighted with low values. Conditionally, the prior for each haplotype risk is expressed as a univariate normal distribution centred on the weighted average of all the haplotype risks.

Clustering methods

Bayesian clustering methods are similar to spatial smoothing techniques in that similar haplotypes are induced to have similar risk. Rather than smoothing the risks based upon spatial similarity, however, haplotypes are placed into spatially homogeneous clusters with constant risk. Molitor *et al.*[16] applied this approach to gene mapping by assuming that each cluster is determined by a 'centre' corresponding to a prototypical haplotype, which can be seen as analogous to the ancestral haplotype from which the other haplotypes in the cluster are derived. The identities of the centres will define the way that haplotypes are allocated to their respective clusters. Given a set of haplotype centres, any observed haplotype will be placed into the cluster corresponding to the closest centre. Here, the risk for a haplotype cluster *c* is defined as $\gamma_c \sim N(\alpha, \sigma_\gamma^2)$. A simple model is then used.

$$\text{Probit}\ \ [\Pr(y_i = 1)] = \gamma_{c_{h_i}} \tag{1}$$

The above model is written for haploid data, but could be extended to handle diploid data by adding a second risk term in equation (1) (plus potential interaction terms) and then treating the haplotypes as latent variables in the MCMC algorithm. As with spatial smoothing techniques, a distance metric is chosen that contains the location of a putative mutation and this location can be estimated as part of the modelling process. Command-line Linux-based software to implement this method for case-control data can be obtained on request from the corresponding author.

Although not formulated in a Bayesian framework (the focus of this paper), it is worth mentioning another method based upon clustering of haplotypes that has been proposed by Durrant *et al.*[17] This method is based upon *cladistic* clustering of haplotypes constructed from simple hierarchical averaging techniques. At each partition, clusters of haplotypes from the previous partition are merged together. The cladogram successively partitions haplotypes $\mathbf{T}[h], \mathbf{T}[h-1], \mathrm{K}, \mathbf{T}[1]$. The first partition, $\mathbf{T}[h]$, consists of *h* clusters; subsequent partitions merge together increasingly diverse clusters of haplotypes. The final partition, $\mathbf{T}[1]$, combines all haplotypes into a single cluster. For large genomic regions, similarity is defined within a sliding window of SNPs. The method has been coded in the CLADHC algorithm and can be obtained from the corresponding author.

Hierarchical modelling methods

Hierarchical methods usually treat allelic effects as a parameter of interest and then use a hierarchical structure to capture the dependence between alleles at different markers based on LD. Recognising the variation that exists across measures of LD, Conti and Witte[18] introduced a hierarchical model to incorporate haplotype block structure and the expected spatial dependency among the markers. They assume a linear relation between the measures of LD, $\beta = (\beta_1, \mathrm{K}, \beta_L)$, and marker-specific covariates and spatially correlated random effects,

$$\beta = \mathbf{Z}\gamma + \varepsilon, \quad \text{where} \quad \varepsilon \sim \mathrm{N}_{\mathbf{L}}(0_{\mathbf{L}}, \tau^2_{\mathbf{G}}\mathrm{T}_G). \qquad (2)$$

Here, $\mathbf{Z}$ is a pre-determined second-stage design matrix composed of indicator variables distinguishing which markers are in a particular haplotype block and $\boldsymbol{\gamma}$ is a column vector of coefficients corresponding to the effects on disease of each block defined in $\mathbf{Z}$. ε is a vector of random effects reflecting within-block variability. Spatial dependencies between the markers can be incorporated into the model through the specification of $\mathbf{T_G}$, a $L \times L$ covariance matrix for the random effects. The above model allows for markers within the same block to borrow information from one another to improve estimation. Using a two-stage estimation procedure and a semi-Bayes approach with σ^2 pre-specified, they demonstrate potential improvements in the pattern of LD. Furthermore, this model can be easily extended to a fully Bayesian framework that also includes the estimation of σ^2_G.

Kilpikari and Sillanpää introduced a hierarchical method for multi-locus association analysis of quantitative and binary traits that postulates different parameters for allelic effects at each marker but selects a trait-associated subset of markers among candidates to be analysed at each cycle of the MCMC sampler.[19] Final results from different models are presented as locus-specific probabilities using Bayesian modelling techniques. This model averaging approach has computational advantages, in that a relatively small, computationally manageable subset of all possible models is analysed at each step in the estimation process. This allows the method to be applied efficiently to wide chromosomal segments. The software is freely available for research purposes under the name BAMA at URL http://www.rni.helsinki.fi/~mjs.

Approximate coalescent methods

The pattern of marker data seen in a sample of individuals is shaped by the interplay between the processes of mutation and recombination that occur over the evolutionary history of the sample. This ancestral history or genealogy of the sample is widely and successfully described by a stochastic process known as the coalescent.[20] The use of the coalescent as the foundation of a model-based analysis approach has been shown to provide a great deal of power in such contexts. While the basic form of the coalescent is a simple Markov chain, however, many complicating factors — such as recombination, population structure and selection — would, ideally, need to be added in order to accurately approximate the processes that are likely to have shaped a sample drawn in a fine-mapping context.

The use of the coalescent in a disease-mapping context is still in its infancy. Initially, several methods used the star-phylogeny, an ancestry in which all sampled haplotypes are assumed to evolve completely independently, to approximate the genealogy of the sample.[21] While the use of a star-phylogeny allows one to avoid a good deal of computational complexity, it fails to capture the correlations induced by the shared ancestry of the sample. Consequently, the variance of the estimates of posterior parameters is likely to be underestimated. Realising this, others have attempted to include more accurate approximations to the coalescent process.[3,22,23] McPeek and Strahs use a star-shaped genealogy, but correct for pair-wise correlations between loci.[3] Their DHSMAP software is available at http://galton.uchicago.edu/~mcpeek/software/dhsmap/. Graham and Thompson introduce the notion of 'recombinant classes' to model the possible existence of several ancestral mutational events from which the sampled cases may have derived.[22] Software is available at http://www.stat.sfu.ca/~jgraham/Papers/Programs/DisequilibriumMapping/. Rannala and Reeve exploit data from an annotated human genome sequence (HGS) as well as data from multiple markers.[23] They use the HGS to generate a prior distribution for the location of functional mutations. Their DMLE + software is available at http://dmle.org/. Perhaps the best of these latter approaches is that of Morris *et al.*, which involves the use of the shattered coalescent.[5] This process captures much of the correlation induced within sampled cases, while approximating that of the controls using more star-like models. Their software is available via e-mail from Andrew Morris at the Wellcome Trust Center for Human Genetics, and is perhaps the most powerful coalescent-based method currently available, although issues of computational complexity prevent its use on datasets involving many markers.

The last word on coalescent-based algorithms for fine-mapping has yet to be written. The key question is this: which parts of the coalescent process need to be included in order to accurately capture the influence of the ancestry on the pattern of LD, and which parts can be ignored in order to gain power by improving computational efficiency? The methods of Terwilliger[21] and Morris *et al.*[5] fall at opposite ends of this spectrum, while those of the other methods referenced in this section lie somewhere between the two.

Conclusion

Bayesian methods are becoming ever more popular in the field of gene mapping, including recent developments in model selection (reviewed in Sillanpää and Corander[24]). Here the authors have summarised some of the more popular currently available methods in this area. (For a more technical review of Bayesian haplotype association methods, see Thomas *et al.*[25]) One area that has not been discussed here is the issue of phase estimation. The standard approach to fine mapping with phase-unknown haplotypes is first to estimate haplotype phase with a program such as PHASE[26] and then to use these estimated haplotypes in a fine-mapping program. One advantage of Bayesian fine-mapping methods, however, is that haplotype phase estimation can be incorporated into a fine-mapping procedure in a unified manner. That is, one can properly account for phase uncertainty in a way that is not possible in a two-stage process. While recent extensions of the expectation-maximisation (E-M) algorithm[6] have provided a frequentist framework for unified inference on haplotype associations allowing for phase uncertainty, many of the previously mentioned Bayesian gene-mapping methods can deal with phase unknown haplotypes in a coherent way.

As mentioned previously, the authors do not claim that this summary is exhaustive. Indeed, given the rate at which this field is progressing, it is quite possible that substantial new methods will be introduced in the time it takes for this paper to reach publication. Although much has been accomplished in this field, clearly more work needs to be done, and, as such, it is likely that this field will continue to undergo rapid expansion.

References

1. Beaumont, M.A. and Rannala, B. (2004), ‘The Bayesian revolution in genetics’, *Nat. Rev. Genet.* Vol. 5, pp. 251–261.
2. Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996), Markov chain Monte Carlo in practice, Chapman and Hall, London, UK.
3. McPeek, M.S. and Strahs, A. (1999), ‘Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping’, *Am. J. Hum. Genet.* Vol. 65, pp. 858–875.
4. Liu, J.S., Sabatti, C., Teng, J. *et al.* (2001), ‘Bayesian analysis of haplotypes for linkage disequilibrium mapping’, *Genome Res.* Vol. 11, pp. 1716–1724.
5. Morris, A., Whittaker, J. and Balding, D.J. (2002), ‘Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies’, *Am. J. Hum. Genet.* Vol. 70, pp. 686–707.
6. Stram, D.O., Pearce, C.L., Bretsky, P. *et al.* (2003), ‘Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals’, *Hum. Hered.* Vol. 55, pp. 179–190.
7. Stram, D.O., Haiman, C.A., Hirschhorn, J.N. *et al.* (2003), ‘Choosing haplotype-tagging SNPs based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the multiethnic cohort study’, *Hum. Hered.* Vol. 55, pp. 27–56.
8. Zhang, K., Qin, Z.S., Liu, J.S. *et al.* (2004), ‘Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies’, *Genome Res.* Vol. 14, pp. 909–916.
9. Nordborg, M. and Tavaré, S. (2002), ‘Linkage disequilibrium: What history has to tell us’, *Trends Genet.* Vol. 18, pp. 83–90.
10. Lu, X., Niu, T. and Liu, J.S. (2003), ‘Haplotype information and linkage disequilibrium mapping for single nucleotide polymorphisms’, *Genome Res.* Vol. 13, pp. 2112–2117.
11. Cressie, N.A. (1993), Statistics for Spatial Data Revised Edition, John Wiley & Sons Inc, New York, NY.
12. Mollié, A. (1996), ‘Bayesian mapping of disease’, Markov Chain Monte Carlo in Practice, Chapman & Hall, London, UK, pp. 359–380.
13. Thomas, D., Morrison, J. and Clayton, D. (2001), ‘Bayes estimates of haplotype effects’, *Genet. Epidemiol.* Vol. 21, pp. S712–S717.
14. Molitor, J., Marjoram, P. and Thomas, D. (2003), ‘Application of Bayesian spatial statistical methods to the analysis of haplotype effects and gene mapping’, *Genet. Epidemiol.*, pp. 95–105.
15. Besag, J., York, J. and Mollié, A. (1991), ‘Bayesian image restoration, with two applications in spatial statistics (with discussion)’, *Ann. I. Stat. Math.* Vol. 43, pp. 1–59.
16. Molitor, J., Marjoram, P. and Thomas, D.C. (2003), ‘Fine-scale mapping of diseases with multiple mutations via spatial clustering techniques’, *Am. J. Hum. Genet.* Vol. 73, pp. 1368–1384.
17. Durrant, C., Zondervan, K.T., Cardon, L.R. *et al.* (2004), ‘Linkage disequilibrium mapping via cladistic analysis of SNP haplotypes’, *Am. J. Hum. Genet.*, Vol. 75, pp. 35–43.
18. Conti, D.V. and Witte, J.S. (2003), ‘Hierarchical modeling of linkage disequilibrum: Genetic structure and spatial relations’, *Am. J. Hum. Genet.* Vol. 72, pp. 351–363.
19. Kilpikari, R. and Sillanpää, M.J. (2003), ‘Bayesian analysis of multilocus association in quantitative and qualitative traits’, *Genet. Epidemiol.* Vol. 25, pp. 122–135.
20. Kingman, J.F.C. (1982), ‘On the genealogy of large populations’, *J. Appl. Prob.* Vol. 19A, pp. 27–43.
21. Terwilliger, J. (1995), ‘A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci’, *Am. J. Hum. Genet.* Vol. 60, pp. 777–787.
22. Graham, J. and Thompson, E. (1998), ‘Disequilibrium likelihoods for fine-scale mapping of a rare allele’, *Am. J. Hum. Genet.* Vol. 63, pp. 1517–1530.
23. Rannala, B. and Reeve, J.P. (2001), ‘High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence’, *Am. J. Hum. Genet.* Vol. 69, pp. 159–178.
24. Sillanpää, M.J. and Corander, J. (2002), ‘Model choice in gene mapping: What and why’, *Trends Genet.* Vol. 18, pp. 301–307.
25. Thomas, D.C., Stram, D.O., Conti, D. *et al.* (2003), ‘Bayesian spatial modeling of haplotype associations’, *Hum. Hered.* Vol. 56, pp. 32–40.
26. Stephens, M., Smith, N.J. and Donnelly, P. (2001), ‘A new statistical method for haplotype reconstruction from population data’, *Am. J. Hum. Genet.* Vol. 68, pp. 978–989.