GRAPHICAL MODELS FOR COMPLEX GENETICS PROBLEMS

N. A. Sheehan[1], B. Guldbrandtsen[2] and D. A. Sorensen[2]

[1]Dept. of Public Health and Epidemiology, University of Leicester, 22-28 Princess Road West, Leicester LE1 6TP, UK
[2]Dept. of Animal Breeding and Genetics, Danish Institute of Agricultural Sciences, PO Box 50, DK-8830 Tjele, Denmark

INTRODUCTION

The analysis of genetic data on general pedigrees can pose enormous computational problems for exact methods of probability and likelihood calculation such as the *peeling* algorithms (Elston and Stewart, 1971; Cannings *et al.*, 1978) when the pedigree or the genetic model under consideration is overly complex. The obvious representation of a pedigree as a graph leads naturally to an exploitation of *graphical models* (Lauritzen, 1996), which have their origins in the development of a probabilistic approach to dealing with uncertainty in expert systems (Pearl, 1988). The idea behind such models is to reduce a complex problem into small manageable subcomponents, thus facilitating understanding of the computational issues involved and informing the development of more efficient algorithms for handling such problems. The advantage to be gained by formally setting complex genetics applications into a more general computational framework is in the development of a flexible modelling environment whereby different problems can be tackled by essentially the same software.

GRAPHICAL MODELS FOR GENETICS APPLICATIONS

A *directed acyclic graph* (DAG) is a set of nodes representing the variables in the model and a set of connecting directed *edges* representing probabilistic influence or causal links between them with the further property that there is no sequence of linking directed edges beginning and ending with the same node (i.e. no directed cycles). Under the usual assumptions of the genetic model, such as Mendelian inheritance, and given that an individual cannot be his own ancestor or descendant, a pedigree problem can be represented as a DAG for which the *local Markov property* (Lauritzen *et al*. 1990) is satisfied. Such structures are often called *Bayesian networks* (Jensen 1996). Algorithms for performing calculations on general Bayesian networks which fully exploit *all* the conditional independencies inherent in the problem (e.g. Lauritzen and Spiegelhalter 1988) can then be applied. These algorithms are essentially the same as the peeling method but are a little more clever computationally. They all break down when the relevant graph has too many interconnecting *undirected* cycles, or *loops*, and the quantities of interest must then be estimated either by Markov chain Monte Carlo (MCMC) methods (Hastings 1970) or by simplifying some aspects of the problem. However, MCMC methods have not been tested sufficiently on large complex problems and tend to be viewed with some suspicion due to the potential unreliability of the resulting estimates.

A SIMPLE MAPPING PROBLEM

In order to illustrate the idea, we will construct a graphical model for an elementary QTL mapping problem on a half-sib design. The details of this construction and its extension to a fully Bayesian analysis of the problem are given in Sheehan *et al.* (in press) and a practical implementation is presented in Guldbrandtsen *et al.* (subm).

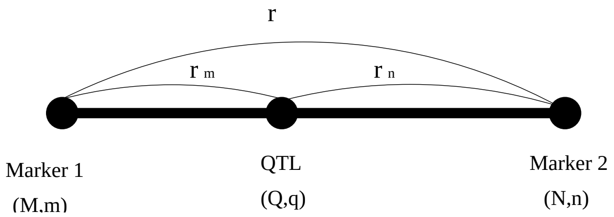

Figure 1. A QTL with two flanking markers (taken from Sheehan and Sorensen, in press)

The mapping problem is illustrated in figure 1 where we assume a single diallelic QTL somewhere between two diallelic marker loci where the marker map positions and allele frequencies are known. The recombination fractions r_m and r_n are not known but only one has to be estimated under the assumption of no genetic interference. We have phenotypic records for the QTL on the offspring and marker data for all individuals. The model for the data is

$$y_{ij} = s_i+\mu_{ij}+e_{ij}$$

where y_{ij} is the phenotype of offspring j of sire i, s_i is the effect of sire i, μ_{ij} is the effect of the QTL genotype of offspring ij and e_{ij} is the random residual of individual ij. The sire effects follow a Normal distribution, i.e. $s_i \sim N(0,\sigma_u^2)$. Similarly, we assume, $e_{ij} \sim N(0,\sigma_{res}^2)$ for the random residuals. For illustration, we will focus on a model assuming all parameters (including r_n and r_m) are known. It is easy to extend the final graph of figure 3 to the fully Bayesian model by adding nodes for parameters and hyper parameters and edge for their interrelationships.

A GRAPHICAL MODEL FOR THE QTL PROBLEM

We begin by considering one sire and one offspring for the first marker locus where alleles M and m have known frequencies p_M and $1-p_M$, say. We index maternally and paternally inherited genes by 0 and 1, respectively, and assume that the two genes in the sire are randomly drawn from the population. Thus, writing the sire's maternal gene at this locus as $M_{(i,0)}$, we have $M_{(i,0)} \sim \mathrm{Ber}(p_M)$, for example. The two genes in the sire determine his genotype $M_{i,}$ so we have directed edges leading to this node from $M_{(i,0)}$ and $M_{(i,1)}$ in figure 2a. A *segregation indicator* S_{ij}^M is introduced to describe transmission of genes from sire i to his j^th offspring at the "M-locus". This is a binary variable taking the value 0 (maternal gene inherited) or 1 (paternal gene inherited) with probability 0.5 i.e. Mendelian segregation. The gene inherited by individual ij from its sire is labelled $M_{(ij,1)}$ in figure 2a and we can immediately infer from the directed edges displayed that its value depends on both genes in the sire and on the segregation indicator determining which one is passed down. Similarly, we can see that the maternally inherited gene, $M_{(ij,0)}$ is drawn randomly from the population, as assumed by the half-sib design, and that both these genes affect the offspring genotype, M_{ij} .

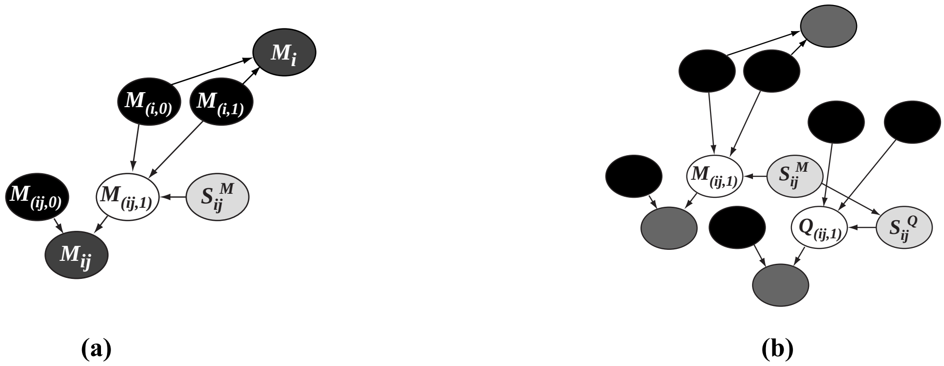

Figure 2. The graphs for one sire and one offspring, with one marker locus (a) and two linked loci (b), respectively. (Taken from Sheehan *et al.* in press).

We will now add the QTL locus, which is linked to the first. As before, the sire's paternal and maternal genes, $Q_{(i,1)}$ and $Q_{(i,0)}$, are independently assigned from a Bernoulli distribution with parameter p_Q (assumed known only for the purposes of illustration). Similarly, the value of the QTL allele $Q_{(ij,1)}$ inherited by individual ij from the sire depends on these and on the segregation indicator S_{ij}^Q at the QTL locus. This time, however, the value of S_{ij}^Q also depends on the value of S_{ij}^M because of linkage and the dependency is a function of the recombination fraction, r_M (figure 1). Hence, $S_{ij}^Q = 1$ (paternal gene inherited) *if* $S_{ij}^M = 1$ *and* there is no recombination. Specifically,

$$S_{ij}^Q \sim \mathrm{Ber}(r_M) \text{ if } S_{ij}^M = 0 \text{ and } S_{ij}^Q \sim \mathrm{Ber}(1\text{-}r_M) \text{ if } S_{ij}^M = 1.$$

Figure 2b depicts these relationships for the case where r_M is known. (When r_M is unknown, it is represented by an extra node with directed edges to both segregation indicators.) Note that a node representing the offspring's QTL genotype, Q_{ij}, has been added even though it is an unobservable quantity, as we have a quantitative phenotype y_{ij} with distribution depending on the genotypic state at the QTL. Adding the second marker locus, which is linked to the QTL, is completely analogous and with the sire effect, s_i, and phenotype nodes y_{ij} we get the final graph in figure 3 for one sire and two offspring.

DISCUSSION

There are no loops in the *pedigree* graph for a half-sib design. However, there are several loops in the graph shown in figure 3 which immediately demonstrates why linkage calculations can be computationally intensive, however simple the *pedigree*. Moreover, different graphical representations for the same problem present widely differing computational challenges and may have different structural and inferential properties. It should be stressed that even with this simple model and design, we are already in a situation where exact methods break down and MCMC methods are required. However, by viewing the problem in a general graphical modelling framework, we can use a commercially available package like HUGIN (Andersen et al, 1989) to carry out joint updating of blocks of variables conditional on the values of all the others within an MCMC framework. This should alleviate problems of slow mixing frequently encountered with single-site sampling schemes, such as the Gibbs sampler. The facility to experiment easily and quickly with different blocking schemes is itself an attractive feature enabling rapid development of methods to handle complex applications in genetics.

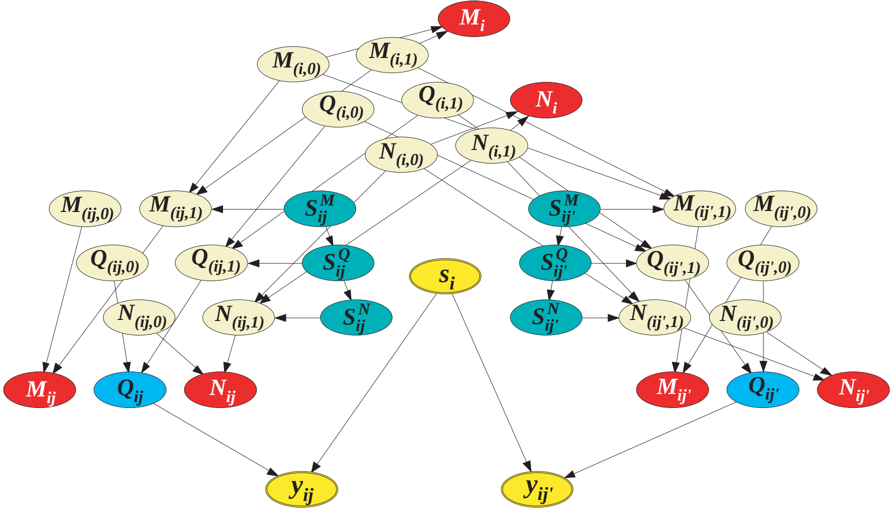

Figure 3. The full graphical model for one sire and two offspring (Sheehan *et al.* in press)

ACKNOWLEDGEMENTS

We acknowledge support from the Wellcome Trust (Grant 056266/Z/98/Z): from the Danish Ministry of Fisheries, Agriculture and Food (FREM98-DJF-1) for B.G. and from the TVW Telethon Institute for Child Health Research, Western Australia for N.A.S.

REFERENCES

Andersen, S.K., Olsen, K.G., Jensen, F.V. and Jensen, F. (1989) In “Proceedings of the 11th International Joint Conference on Artificial Intelligence”, p.1080-1085, Morgan Kaufmann.

Cannings, C., Thompson, E.A. and Skolnick, N. (1978) *Adv. Appl. Prob.* **10**: 26-61.

Elston, R.C. and Stewart, J. (1971) *Hum. Hered.* **21**: 523-542.

Guldbrandtsen, B., Sheehan, N.A. and Sorensen, D.A. (*subm.*) *Proc. 7th WCGALP*

Jensen, F.V. (1996) “An Introduction to Bayesian Networks”. UCL Press, UK.

Hastings, W.K. (1970) *Biometrika* **57**: 97-109.

Lauritzen, S.L. (1996) “Graphical Models”. Clarendon Press, Oxford, UK.

Lauritzen, S.L. and Spiegelhalter, D.J. (1988*) J.Roy.Stat.Soc. B.* **50**: 157-224.

Lauritzen, S.L., Dawid, A.P., Larsen, B. and Leimer, H.G. (1990) *Networks* **20**: 491-505.

Pearl, J. (1988) “Probabilistic Inference in Intelligent Systems”. Morgan Kaufmann, CA.

Sheehan, N.A., Guldbrandtsen, B., Lund, M.S. and Sorensen, D.A. (in press) *Int. Stat. Rev.*

Sheehan, N.A. and Sorensen, D.A. (in press) In ”Highly Structured Stochastic Systems, Editors P. Green, N. Hjort and S. Richardson, Oxford University Press, UK.