Genet. Res., Camb. (2003), **81**, *pp.* 205–212. *With* 1 *figure.*
DOI: 10.1017/S0016672303006232 Printed in the United Kingdom

An algorithm for sampling descent graphs in large complex pedigrees efficiently

First published in *Genetical Research*, volume 81, issue 3, 2003.
Published by Cambridge University Press.
Genetical Research online: http://journals.cambridge.org/action/displayJournal?jid=GRH
This article is also available online at: http://dx.doi.org/10.1017/S0016672303006232

JOHN M. HENSHALL[1]* AND BRUCE TIER[2]
[1] *CSIRO Livestock Industries, J. M. Rendel Laboratory, Rockhampton, QLD, Australia*
[2] *Animal Genetics and Breeding Unit**, University of New England, Australia*

(*Received 1 April 2002 and in revised form 12 November 2002*)

Summary

No exact method for determining genotypic and identity-by-descent probabilities is available for large complex pedigrees. Approximate methods for such pedigrees cannot be guaranteed to be unbiased. A new method is proposed that uses the Metropolis–Hastings algorithm to sample a Markov chain of descent graphs which fit the pedigree and known genotypes. Unknown genotypes are determined from each descent graph. Genotypic probabilities are estimated as their means. The algorithm is shown to be unbiased for small complex pedigrees and feasible and consistent for moderately large complex pedigrees.

1. Introduction

Many methods are currently used for estimating genotypic and identity by descent (IBD) probabilities in human and animal pedigrees. Genotypic and IBD probabilities are of interest to geneticists studying the transmission of genes through complex pedigrees, where a gene might be for a genetic disorder, a molecular test or marker. Commonly, the genotypes of some individuals in the population are known with certainty, partially known for other individuals (in that some genotypes can be excluded) and unknown for the remainder of the population. Most methods for estimating genotypic and IBD probabilities are suitable for small pedigrees, but have disadvantages when applied to large and complex pedigrees, where 'complex' implies the presence of many marriage and inbreeding loops.

Gene dropping provides unbiased estimates, but it is only feasible for very small pedigrees. Methods based on peeling (e.g. Elston & Stewart, 1971; van Arendonk *et al.*, 1989; Fernando *et al.*, 1993; Stricker *et al.*, 1995; Janss *et al.*, 1995*b*; Kerr & Kinghorn, 1996) provide unbiased estimates of genotypic probability for small pedigrees or pedigrees without loops. However, exact peeling is computationally infeasible for large complex pedigrees. Pedigree simplification and iterative peeling are two feasible methods used for large complex pedigrees, but estimates can no longer be guaranteed to be unbiased (Fernando *et al.*, 1993).

To solve the problem, a number of Markov chain Monte Carlo (MCMC) methods have been used to estimate genotypic and IBD probabilities (e.g. Lange & Sobel, 1991; Sobel & Lange, 1993, 1996; Guo & Thompson, 1994; Janss *et al.*, 1995*a*). These methods sample either genotypes or descent graphs. They can produce unbiased samples for large complex pedigrees, provided that the sampling algorithm can traverse the parameter space efficiently; however, impediments to traversing the parameter space can be severe. Furthermore, when there are more than two alleles at the locus, MCMC methods for sampling genotypes are not necessarily irreducible (Sheehan & Thomas, 1993). MCMC methods which operate by sampling descent graphs need not be subject to irreducibility problems, but as noted by Sobel & Lange (1996), 'mixing' may be very poor. Adjacent samples are highly correlated and it may be infeasible to obtain sufficient samples to guarantee estimates have a low probability of error.

Attempts to improve the performance of descent graph sampling algorithms have focused on the

* Corresponding author. CSIRO Livestock Industries, Private Mail Bag, 1, Armidale, New South Wales, 2350, Australia. Tel: +61 2 67761302. Fax: +61 2 67761333. e-mail: John.Henshall@csiro.au
** AGBU is a joint institute of NSW Agriculture and The University of New England.

correlation of adjacent samples, and on drawing legal descent graphs through genotype elimination. At one extreme are MCMC algorithms such as those of Sobel & Lange (1996) where the autocorrelation of legal candidate samples is high. To reduce this autocorrelation composite transmission rules have been proposed (e.g. Lange & Sobel, 1991; Sobel & Lange, 1993, 1996), in which structured groups of elements of the descent graph are changed together. However, changing more of the descent graph reduces the probability that the result will be legal, so more samples may be required. At the other extreme are methods for sampling uncorrelated descent graphs, such as those of Henshall *et al.* (1999, 2001). By applying the genotype elimination through inheritance constraint (GEIC) algorithm, all samples are legal, but the density of the samples drawn is due not only to the likelihood of the sample but also to properties of the algorithm used to obtain the sample (the GEIC sampling density). While it is easy to use importance sampling or a Metropolis–Hastings step to adjust for the GEIC sampling density, for pedigrees of reasonable size this adjustment results in a small number of effective samples, resulting in estimates of low accuracy.

In this paper a new MCMC method for sampling descent graphs is proposed in which the independent descent graph method of Henshall *et al.* (2001) is placed into a MCMC context. A Metropolis–Hastings (MH) step is used to accept or reject candidate graphs which have far less autocorrelation than the candidate graphs used in other MCMC methods. The paper shows that the algorithm can produce unbiased estimates on small complex pedigrees and that it may be feasible for moderately large complex pedigrees.

2. Method

Henshall *et al.* (1999, 2001) describe a method (GEIC) for estimating genotypic and IBD probabilities from independently sampled descent graphs. Each descent graph is sampled *de novo* and consequently adjacent samples are completely uncorrelated. A description of this method, reprinted from Henshall *et al.* (2001), is contained in the Appendix.

This new algorithm puts the GEIC algorithm into an MCMC framework. An initial descent graph is sampled using GEIC. Subsequent samples are obtained by using the GEIC and MH algorithms. A subset of the primary descent graph is retained from the previous sample, and the remainder of the primary descent graph is sampled using GEIC. A MH step is used to accept or reject the candidate. This sampling procedure in the algorithm differs from that described in Henshall *et al.* (2001) in two important ways.

Firstly, a partial primary descent graph is sampled from the inheritance constraints of the current sample primary descent graph. This is a legal subset of the current descent graph. GEIC is then used to complete a new primary descent graph. Base alleles and a secondary descent graph are then sampled as in Henshall *et al.* (2001).

Secondly, the MH algorithm is used to accept or reject candidate samples, based on the likelihood of the sample and the probability of moving from the current descent graph to the candidate descent graph, rather than using importance sampling to weight samples. This probability is similar to the importance sampling density in the method of Henshall *et al.* (2001), and is a function of the number of elimination and base gamete sampling steps required to produce each sample. The MH algorithm is of benefit here as adjacent samples are correlated.

The new algorithm is an MCMC descent graph sampling algorithm, but has the ability to make long jumps between adjacent samples. Accordingly it is referred to here as the long jumping descent graph sampler (LJDGS). A full description of the algorithm follows:

1. Obtain a legal descent graph and associated likelihood ($\pi(x)$) and GEIC sampling density ($g(x)$) using the method of Henshall *et al.* (2001).
2. Repeat
   (a) Sample a subset of the primary descent graph to retain.
   (b) Apply GEIC to the pedigree, constraining the subset to remain unchanged, to obtain a new primary descent graph with likelihood $\pi(y)$ and GEIC sampling density $g(y)$.
   (c) Apply the MH algorithm, with $q(x, y) = g(y)$, $q(y, x) = \mathrm{g}(x)$, and acceptance criterion
   $$\min((\pi(y)q(y, x))/(\pi(x)q(x, y)), 1)$$
   (d) If the candidate sample is accepted, then set $\pi(x) = \pi(y)$ and $g(x) = g(y)$
   (e) Accumulate the most recently accepted descent graph and associated genotypes
3. Summarize parameters of interest as means of the samples.

(i) *Subset sampling algorithm*

It is critical to the success of the algorithm that the method of sampling subsets of the primary descent graph to retain (step 2a) is balanced, and does not affect the candidate generating density $g(y)$. While one can simply prove that a particular subset sampling algorithm does not satisfy this criterion by exception, to prove that a particular subset sampling algorithm does satisfy this criterion would require consideration

of all possible genotype–pedigree configurations. For other than very small or simple pedigrees this may not be possible. However, exhaustive testing has failed to find fault with the strategy described below.

A primary descent graph consists of the set of paths connecting informative gametes to base gametes. Before choosing the subset to be retained, a binary variable is initialized to 'save' for all gametes in the primary descent graph. All paths connecting informative gametes to base gametes are traversed. With each step (gamete) on the traverse, a random number between zero and one is drawn. If the random number exceeds a predetermined non-zero probability b, then the binary variable associated with the gamete switches from the 'save' state to a 'discard' state. All binary variables associated with gametes on the path between, and including, it and the base gamete are set to 'discard' (the initial change to 'discard' breaks the link between the informative and base gametes). It is possible for paths connecting a number of informative gametes to intersect at the same gamete. If when traversing a path a gamete with a variable set to 'discard' is found then the remainder of that path has also been set to 'discard' and is not reset again. For any individual on a path where one of its gametes' variables has been set to 'discard', the state of the variable associated with its other gamete (and all gametes between it and the connected base gamete) is also set to 'discard'. To facilitate mixing, the gametes of full-sib animals are also set to 'discard'. Gametes that remain set to 'save' retain their current inheritance state when a new primary descent graph is sampled.

The variable b need not remain constant, and the algorithm may be 'tuned' to different pedigrees by varying the method of selecting b.

3. Test analyses

Two test pedigrees were used to test the LJDGS algorithm.

A pedigree with 11 individuals (pedigree A, Table 1) was used to validate that the method was able to produce unbiased estimates. This pedigree is small, allowing the calculation of exact genotypic probabilities for comparison purposes. A single locus, with 4 alleles with founder allele frequencies (0·5, 0·25, 0·2, 0·05) was assumed, and test data sets obtained by sampling base gametes and Mendelian transmission. On each test data set genotypes were made available for four randomly chosen individuals, and assumed unknown on the remaining seven individuals.

From pedigree A, 1000 random data sets were analysed twice, once with adjacent samples independent – essentially the algorithm used by Henshall *et al.* (2001) but with a MH step instead of importance sampling (IDGS), and once with the LJDGS, with correlated descent graph samples. With LJDGS, the parameter b did not remain constant, but for each sample was drawn from a $\beta(1,1)$ distribution. Genotypic and IBD probabilities were estimated as the mean of 10 000 samples. As the MH algorithm is used, the effective number of samples is fewer than 10 000. A simple estimate of the effective number of samples was used. The sample which was retained for the most MH cycles ($r_{\max}$) was assumed to have contributed one to the effective number of samples, the effective number of samples then being $n_s/r_{\max}$, where n_s is the total number of samples. This measure is less appropriate for LJDGS than IDGS, because in LJDGS adjacent samples are correlated. Exact genotypic probabilities were obtained using MENDEL (Lange *et al.*, 1988), which uses a peeling-based algorithm.

Table 1. *Pedigree A*

id	Father	Mother
1	0	0
2	0	0
3	0	0
4	2	1
5	2	3
6	2	4
7	5	4
8	5	3
9	7	6
10	7	8
11	10	9

For each analysis, genotype was sampled for base individuals and Mendelian transmission was sampled for non-base individuals. Unordered genotype was then made available for only four randomly chosen individuals.

To compare the sampled genotypic probabilities with the exact probabilities, the test statistic used in Henshall *et al.* (2001) was used, $\chi^2 = \sum_{E_{kl} \neq 0} (O_{kl} - E_{kl})^2 / E_{kl}$, where k relates to the individual, l is the unordered (no distinction between paternal and maternal) genotype, E_{kl} is the number of samples expected to occur for genotype l in individual k (calculated from the probabilities obtained using MENDEL and the effective number of samples) and O_{kl} is the effective number of samples which were observed for genotype l in individual k. This statistic has an approximate χ^2 distribution, with $n-11$ degrees of freedom, where n is the number of non-zero E_{kl} in the sum. The distribution of the test statistic is only approximate as both within and across individuals, genotype probabilities are not independent, and the effective number of samples is only an approximation.

A larger pedigree with 1600 individuals was used to evaluate the performance of the LJDGS algorithm on more challenging data. This pedigree, modelled on the simulated pedigree of Heath (1998), consisted of 20 discrete generations, each with 80 individuals. These

were the 10 progeny from each of 8 matings between males and females from the previous generation. A single locus with 16 alleles, with uniform frequencies in the base individuals, was assumed. A single data set was obtained by sampling base gametes and Mendelian transmission. Analyses were then performed with varying proportions of the simulated genotypes made available. For all analyses genotypes for all individuals born in the final generation were available. Three data sets were constructed, with genotypes made available on 0, 25% or 50% of the remaining individuals, with individuals to be genotyped chosen at random (pedigrees B0, B25 and B50).

Each of pedigrees B0, B25 and B50 was analysed using two methods: once with no correlation between adjacent samples (IDGS) and once with correlated descent graph samples (LJDGS). Each analysis involved drawing 10 000 samples, and was repeated five times, using different seeds for the random number generator each time. In each case the first 1000 samples were discarded. With LJDGS, the parameter b did not remain constant, but for each sample was drawn from a $\beta(80,1)$ distribution. Genotypic and IBD probabilities were estimated as means of the last 9000 samples, and the effective number of samples calculated as described above.

Mixing was assessed in four ways. The first was by examining the degree of symmetry in the inheritance of base gametes, as the probability of a base gamete being sampled as 'paternal' should equal the probability of it being sampled as 'maternal'. The test statistic used was

$$\mathrm{SYM} = 1 - \frac{\sum_{i=1}^{n_b} \sum_{j=1}^{n_a} \sum_{k=1}^{j} \left| p_{ijk} - p_{ikj} \right|}{n_b}$$

where n_b is the number of base individuals (=80), n_a is the number of alleles (=16) and p_{ijk} is the probability that individual i inherited allele j from its sire and allele k from its dam. This statistic can take values from zero to one, with lower values less symmetric in the inheritance of base alleles.

For the second measure of mixing unordered genotypic probability estimates were considered across the five replicates. The number of genotypic probability estimates in which at least one replicate had a zero probability while at least one replicate had a probability greater than 0·00, 0·01, 0·02, 0·05 or 0·10 was calculated. The statistics Z_0, Z_{01}, Z_{02}, Z_{05} and Z_{10} are these counts, expressed as percentages of the total number of cells with non-zero probabilities, excluding cells in which there was no variation.

For the third measure of mixing unordered genotypic probability estimates were again considered across replicates. The statistics S_{01}, S_{02}, S_{05} and S_{10} are the number of cells in which the standard deviation (over replicates) of the genotypic probability estimate exceeded 0·01, 0·02, 0·05 or 0·10 respectively, expressed as percentages of the total number of cells with non-zero standard deviations.

The fourth measure of mixing was to compare the within- and across-replicate variation in the sampled descent graphs. The statistic W is the percentage of gametes (excluding base gametes) for which the across-replicate variation was significant at the 1% level. To compute this statistic, the within-sample variance was calculated assuming a binomial distribution, using the estimated number of effective samples. The degrees of freedom were also calculated with respect to the estimated number of effective samples, and approximate significance level obtained from significance levels for an F distribution.

While these tests may identify inadequate mixing, that inadequate mixing has not been identified is not a guarantee of adequate mixing.

4. Results

The genotypes for pedigree A are well estimated (Fig. 1). The test statistics obtained for the 1000 analyses of pedigree A are plotted against the approximate degrees of freedom, for both IDGS and LJDGS. The test statistics show chance deviations from expectation. Although the distributions of test statistics are similar for the two methods, there appear to be fewer extreme test statistics for LJDGS.

With large complex pedigrees LJDGS performs much better than IDGS (Table 2). Of the 9000 samples, the number accepted is far higher with LJDGS than IDGS, and this is reflected in the much higher effective number of samples for LJDGS. The effective number of samples for IDGS is so low that genotypic probability estimates would be expected to be of very low accuracy, while with LJDGS one would expect reasonable estimates. However, for LJDGS samples are not independent, so it is possible that the effective number of samples is an overestimate.

Most of the measures of mixing indicate that LJDGS shows better mixing than IDGS. The first measure of mixing, symmetry in the base individuals, provides no evidence of mixing problems with LJDGS. However, reasonably high levels of symmetry also occur for IDGS for pedigree B0, despite only 2·5 effective samples, suggesting that base symmetry is not in itself an indicator of good mixing. The second measure of mixing, the percentage of genotypes in which at least one zero probability was observed while the maximum probability observed was greater than 0·00, 0·01, 0·05 or 0·10, suggests that mixing has been good with LJDGS for all pedigrees. Again, however, no problem with mixing has been observed for IDGS for one pedigree, in this case B50. This casts some doubt on the worth of this statistic as an indicator of

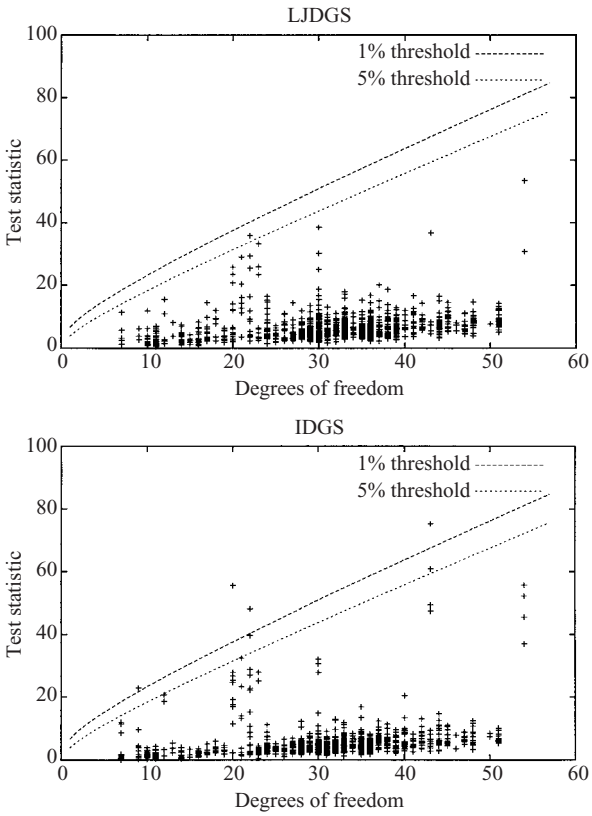

Fig. 1. Distribution of the test statistic obtained for 1000 replicates of pedigree A analysed using the descent graph sampler with independent samples (IDGS) and the long jumping descent graph sampler with correlated samples (LJDGS).

mixing. The third measure of mixing, the standard deviation of genotype probabilities across replicates, suggests that LJDGS mixes better than IDGS. Standard deviations of less than 5% suggest that genotypic probability estimates are accurate to around 5%, which may be acceptable for some applications. A relatively small percentage of genotypic probabilities had standard deviations above 5%, especially for pedigrees B25 and B50.

The fourth measure of mixing, the significance of the between-sample variation in gamete inheritance, suggests that IDGS is superior to LJDGS. This is because, with such a low effective number of samples within-replicate variation is assumed to be high, making between-replicate variation insignificant. For LJDGS, this statistic clearly indicates that mixing has not been ideal, especially for pedigree B25, where for almost one-quarter of gametes the between-replicate variance is significant at the 1% level. Alternatively or as well, this may be an indication that the number of effective samples has been overestimated.

In Table 3 the mean and minimum percentage of the primary descent graph saved is provided for both accepted and rejected samples. It is clear that samples are more likely to be accepted if a large proportion of

Table 2. *Results of five repeated analyses of a pedigree with 1600 individuals*

	B0		B25		B50	
Pedigree	IDGS	LJDGS	IDGS	LJDGS	IDGS	LJDGS
N_{ACC}	5	1116	4	719	27	1471
N_{EFF}	1·5	137·9	1·6	96·4	4·7	209·1
SYM	0·96	0·99	0·77	0·96	0·80	0·98
Z_0	19·3	1·9	14·9	5·9	2·1	1·9
Z_{01}	19·2	0·8	14·8	4·5	1·6	0·8
Z_{02}	19·2	0·5	14·6	3·8	1·4	0·6
Z_{05}	19·1	0·3	14·3	2·6	1·1	0·2
Z_{10}	18·9	0·3	14·2	1·9	0·9	0·0
S_{01}	92·0	77·5	94·7	79·3	94·8	72·8
S_{02}	69·0	58·1	89·8	54·6	93·2	19·5
S_{05}	69·0	22·2	85·3	20·1	90·2	0·4
S_{10}	51·6	5·6	76·7	8·3	78·2	0·2
S_{20}	41·1	0·6	69·4	2·8	33·8	0·0
W	0·0	9·8	0·2	23·0	0·0	8·7

A 16-allele locus was simulated, with genotypes available on the last generation and on 0 (B0), 25% (B25) or 50% (B50) of the remaining individuals. Ten thousand samples were drawn using the independent descent graph sampler (IDGS) and the long jumping descent graph sampler, with the first 1000 samples discarded. The number of samples accepted (N_{ACC}) and an approximation of the effective number of samples (N_{EFF}) is provided. For these, along with one of the measures of mixing, base allele symmetry (SYM), larger values are desirable. Three other measures of mixing are provided, for which smaller values are desirable. Mixing in genotype estimates is compared using the percentage of genotypes in which at least one replicate had a zero probability while at least one replicate had a probability greater than 0·00 (Z_0), 0·01 (Z_{01}), 0·02 (Z_{02}), 0·05 (Z_{05}) or 0·10 (Z_{10}), and the percentage of genotypes in which the standard deviation of the probability estimate across replicates exceeded 0·01 (S_{01}), 0·02 (S_{02}), 0·05 (S_{05}), 0·10 (S_{10}) or 0·20 (S_{20}). These percentages are calculated using only genotypes in which there was some variation across replicates. The last measure of mixing (W) is the percentage of gametes for which the variation in origin (grandpaternal or grandmaternal) across replicates is significant at the 1% level, when compared with the within-sample variation.

Table 3. *Summary statistics for accepted and rejected samples for pedigrees B0, B25 and B50*

	ACC	μ_a	μ_r	$\min_a$	$\min_r$
B0	0·12	0·98	0·85	0·58	0·24
B25	0·08	0·98	0·72	0·71	0·22
B50	0·16	0·95	0·74	0·54	0·29

The percentage of samples accepted (ACC) is provided, along with the mean and minimum percentage of the primary descent graph retained for both the accepted samples (μ_a and $\min_a$) and rejected samples (μ_r and $\min_r$).

the current primary descent graph is retained. However, it is also evident that on occasions samples are accepted which retain only two-thirds of the current primary descent graph.

Runs of 10 000 samples took on average 16·6, 3·4 and 4·5 hours for LJDGS on pedigrees B0, B25 and B50 respectively, using a Pentium III Xeon 1·7 GHz processor.

5. Discussion

By using the GEIC (genotype elimination by inheritance constraint) algorithm the method described in this paper generalizes the method of Sobel & Lange (1993) for sampling descent graphs. The GEIC algorithm ensures that all candidate descent graphs are legal, and there is no need to 'tunnel through' illegal descent graphs. As with all descent graph sampling algorithms, LJDGS is suitable for loci with more than two alleles to be evaluated without concern about the Markov chain being reducible.

The results from pedigree B suggest that this method may feasible on moderately sized pedigrees with large numbers of alleles per locus. The variation in mixing, and large variation in time taken for pedigrees B0, B25 and B50, suggest that size and number of alleles are not the only factors affecting feasibility. The proportion and distribution of genotyped individuals in the pedigree is also very important, with pedigrees rich in genotyped individuals more quickly analysed than pedigrees with sparse genotype information, but with pedigrees with an intermediate level of genotype information more likely to produce poor mixing.

The evaluation of the small pedigrees described in Table 1 indicates that the results from this algorithm are unbiased. Nevertheless, we found that the choice of too few gametes for re-sampling could limit mixing. The variation in the MH acceptance rate between pedigrees B0, B25 and B50 suggests that the algorithm may be tuned by varying the method of sampling b, the proportion of the primary descent graph to retain. Here, b was sampled from a $\beta(a,1)$ distribution, where $a=\max(n/20,1)$ and n was the number of individuals. It may be more appropriate to use a function of the number of genotyped individuals as the first parameter in the beta distribution. This tuning may be very important. If b is consistently small, then fewer candidate graphs will be accepted, and the effective number of samples will be reduced. If b is consistently large, then adjacent samples will be more correlated, again reducing the number of effective samples. This could be determined while the algorithm is running.

The method for sampling descent graphs, by randomly drawing a number of gametes to be re-sampled, should permit rapid exploration of the parameter space, enhancing mixing. While it has not been shown that mixing is a problem in descent graph sampling methods such as Sobel & Lange (1996), it is difficult to be sure that it is not. These methods are roughly equivalent to LJDGS but with a value of b very close to 1·0, and as mixing is shown to be a problem for LJDGS in pedigree B25, it is likely that mixing will also be a problem for descent graph sampling methods that move through the parameter space making smaller jumps. The methods used here to test mixing are not perfect; for example, the methods which use the similarity between replicates would give a favourable statistic if all replicates were similar, even if this similarity were due solely to the use of similar starting values. This could be a problem with any MCMC method that requires a valid descent graph as a starting value, as the descent graph sampling density of Henshall *et al.* (2001) shows that some descent graphs are thousands of times more likely to be found than others, and this variation is not due to the likelihood of the descent graph. Therefore, there may be a significant chance that sampled descent graphs, drawn for use as 'fresh' starting values, all share some characteristic. It would appear that this is less likely to be a problem with LJDGS, as the descent graph sampling density is explicitly included in the MH step. LJDGS correctly accounts for the density of the starting values, and provides a statistically sound method for combining replicates.

The execution times presented are for development software, and it is likely that significant speedups could be made through enhancements to the GEIC algorithm such as those proposed by Du & Hoeschele (2000). As they are cheap to obtain, it is possible to sample many secondary descent graphs for each primary descent graph, and to use a weighted average for determining genotypic and IBD probabilities. Even with significant speedups it will not be possible to draw as many samples as is possible with other MCMC descent graph sampling algorithms, such as those of Sobel & Lange (1996). However, as adjacent samples obtained with LJDGS should be less correlated than those from other MCMC algorithms, fewer samples need be drawn to get good genotypic probability estimates. A composite method, using a conventional MCMC descent graph sampler to sample in the region of each LJDGS sample, is also possible.

While the results presented here are for single loci, the extension to multiple loci is straightforward, using the likelihoods in Sobel & Lange (1996). As the algorithm operates by re-sampling whole regions of the descent graph together, re-sampling all loci within individuals together promotes efficient mixing. Interestingly, for multilocus graphs the GEIC sampling density has less effect on the probability of accepting a sample, as it becomes a function of powers of $1-r$

and r, where r is the recombination rate, instead of a function of powers of $\frac{1}{2}$. Estimates of haplotype probabilities can be obtained for multilocus data. Descent graph sampling methods for quantitative trait loci (QTL), such as that of Tier & Henshall (2001), can also be combined with LJDGS, to sample QTL linked to markers.

Estimating genotypic and IBD probabilities in large complex pedigrees remains a difficult and complicated process. With the method described here, the output should be critically examined to ensure that the likelihood is not being dominated by the GEIC sampling density, and all available measures of mixing should be considered. For some pedigrees the method may be totally unsuitable, but the diagnostics suggested here should give an indication of whether or not this is the case.

6. Conclusions

By combining the best elements of existing MCMC descent graph sampling algorithms with the best elements of independent descent graph sampling methods, the method proposed here has the potential to be of use in estimating genotypic probabilities and IBD probabilities in moderately large complex pedigrees. Adjacent samples are potentially less correlated than those produced using other MCMC descent graph sampling algorithms, reducing the number of samples required to produce reliable estimates. At the same time, allowing some correlation ensures that enough samples are accepted to make the analysis of large pedigrees feasible.

Appendix. The GEIC algorithm (from Henshall *et al.* 2001)

1. Construct a list of informative gametes, ordered in reverse pedigree order (i.e. progeny before parents).
2. Construct a list of feasible ordered genotypes for each individual, using the genotype elimination algorithm (Lange & Goradia, 1987).
3. For each informative gamete, construct a path from the informative gamete to a base gamete by repeating the following:

   (a) If a path already exists from the gamete to a parent gamete, proceed up the path until a gamete with an unconstrained inheritance state is found.
   (b) If the gamete is a base gamete, proceed to the next informative gamete.
   (c) Otherwise, sample an inheritance state for the gamete, by eliminating one inheritance state at random.
   (d) Construct a new list of feasible ordered genotypes for each individual, using the genotype elimination algorithm, modified to take account of inheritance constraints. Note that there may be no valid genotypes for some individuals, in which case the sample is illegal, and the algorithm has failed.

The sample now consists of a set of constrained inheritance states connecting informative gametes to base gametes. We will refer to this set as the primary descent graph sample.

4. Assign an inheritance state at random to every non-base gamete which is not in the primary descent graph sample. These will be referred to as the secondary descent graph sample.
5. If the sample is legal, each base gamete in the primary descent graph sample will now be constrained to have either a single possible allelic type, or a subset of possible allelic types. Construct a list of those which have more than one possible allelic type.
6. Repeat for each base gamete in the list of base gametes with more than one possible allelic type:

   (a) Constrain the allelic type by assigning at random one of the possible allelic types; reference may be made to prior allele frequencies for base alleles.
   (b) Use the genotype elimination algorithm, modified to take account of inheritance constraints, to determine the consequences of this constraint. This may include removing base gametes from the list to be constrained.

7. Base gametes which are not in the primary descent graph sample can be sampled according to prior allele frequencies for base alleles.
8. Drop down through the pedigree, assigning allelic types to all gametes.

The sampled inheritance states obtained following step 4 comprise a legal descent graph. With base gametes uniquely determined (steps 6 and 7), and a legal descent graph, the allelic type of all gametes in the pedigree is also uniquely determined, and comprise a legal descent state.

References

Du, F.-X. & Hoeschele, I. (2000). A note on algorithms for genotype and allele elimination in complex pedigrees with incomplete genotype data. *Genetics* **156**, 2051–2062.

Elston, R. C. & Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity* **21**, 523–542.

Fernando, R. L., Stricker, C. & Elston, R. C. (1993). An efficient algorithm to compute the posterior genotypic distribution for every member of a pedigree without loops. *Theoretical and Applied Genetics* **87**, 89–93.

Guo, S. W. & Thompson, E. A. (1994). Monte Carlo estimation of mixed models for large complex pedigrees. *Biometrics* **50**, 417–432.

Heath, S. C. (1998). Generating consistent genotypic configurations for multi-allelic loci and large complex pedigrees. *Human Heredity* **48**, 1–11.

Henshall, J. M., Tier, B. & Kerr, R. J. (1999). Inferring genotype probabilities for untyped individuals in complex pedigrees. In *Proceedings of the Thirteenth Conference of the Association for the Advancement of Animal Breeding and Genetics*, pp. 329–332.

Henshall, J. M., Tier, B. & Kerr, R. J. (2001). Estimating genotypes with independently sampled descent graphs. *Genetical Research* **78**, 281–288.

Janss, L. L. G., Thompson, R. & van Arendonk, J. A. M. (1995*a*). Application of Gibbs sampling for inference in a mixed major gene-polygenic inheritance model in animal populations. *Theoretical and Applied Genetics* **91**, 1137–1147.

Janss, L. L. G., van Arendonk, J. A. M. & van der Werf, J. H. J. (1995*b*). Computing approximate monogenic model likelihoods in large pedigrees with loops. *Genetics Selection Evolution* **27**, 567–579.

Kerr, R. J. & Kinghorn, B. P. (1996). An efficient algorithm for segregation analysis in large populations. *Journal of Animal Breeding and Genetics* **113**, 457–469.

Lange, K. & Goradia, T. M. (1987). An algorithm for automatic genotype elimination. *American Journal of Human Genetics* **40**, 250–256.

Lange, K. & Sobel, E. (1991). A random walk method for computing genetic location scores. *American Journal of Human Genetics* **49**, 1320–1334.

Lange, K., Weeks, D. E. & Boehnke, M. (1988). Programs for pedigree analysis: MENDEL, FISHER and dGENE. *Genetic Epidemiology* **5**, 471–472.

Sheehan, N. & Thomas, A. (1993). On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. *Biometrics* **49**, 163–175.

Sobel, E. & Lange, K. (1993). Metropolis sampling in pedigree analysis. *Statistical Methods in Medical Research* **2**, 263–282.

Sobel, E. & Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *American Journal of Human Genetics* **58**, 1323–1337.

Stricker, C., Fernando, R. L. & Elston, R. C. (1995). An algorithm to approximate the likelihood for pedigree data with loops by cutting. *Theoretical and Applied Genetics* **91**, 1054–1063.

Tier, B. & Henshall, J. M. (2001). A sampling algorithm for segregation analysis. *Genetics Selection Evolution* **33**, 587–603.

van Arendonk, J. A. M., Smith, C. & Kennedy, B. W. (1989). Method to estimate genotype probabilities at individual loci in farm livestock. *Theoretical and Applied Genetics* **78**, 735–740.

First published in *Genetical Research*, volume 81, issue 3, 2003.
Published by Cambridge University Press.
Genetical Research online: http://journals.cambridge.org/action/displayJournal?jid=GRH
This article is also available online at: http://dx.doi.org/10.1017/S0016672303006232