Genet. Res., Camb. (2001), **78**, *pp.* 281–288. *With 3 figures.* © 2001 Cambridge University Press
DOI: 10.1017/S0016672301005316 Printed in the United Kingdom

Estimating genotypes with independently sampled descent graphs

JOHN M. HENSHALL*, BRUCE TIER AND RICHARD J. KERR
Animal Genetics and Breeding Unit†, University of New England, Armidale, NSW 2351, Australia

(*Received 14 August 2000 and in revised form 31 January and 8 May 2001*)

Summary

A method for estimating genotypic and identity-by-descent probabilities in complex pedigrees is described. The method consists of an algorithm for drawing independent genotype samples which are consistent with the pedigree and observed genotype. The probability distribution function for samples obtained using the algorithm can be evaluated up to a normalizing constant, and combined with the likelihood to produce a weight for each sample. Importance sampling is then used to estimate genotypic and identity-by-descent probabilities. On small but complex pedigrees, the genotypic probability estimates are demonstrated to be empirically unbiased. On large complex pedigrees, while the algorithm for obtaining genotype samples is feasible, importance sampling may require an infeasible number of samples to estimate genotypic probabilities with accuracy.

1. Introduction

The estimation of genotypic probabilities and identity-by-descent (IBD) probabilities is of interest to geneticists studying both human and animal pedigrees. In human populations, the gene of interest might be recessive, with genotypes only observable on individuals carrying two copies of the deleterious allele. In livestock populations, undesirable recessive alleles are also of interest, as are areas of the genome which are associated with traits of economic importance. In livestock species, it is increasingly common for the genotype of some animals in the pedigree to be determined using a test (genotyping), and the ability to infer the genotypes of related individuals has the potential to significantly reduce costs.

Estimates of IBD probabilities are of interest to geneticists when knowledge of genotypic probability is insufficient. An example is where no test exists for a gene of importance but there is a test for a linked marker. In this case the actual genotype at the marker locus is not important as, unless there is reason to expect linkage disequilibrium, marker genotype does not determine genotype for the gene of interest. However, by estimating IBD at the marker locus, IBD at the gene locus can be inferred, with appropriate adjustments for recombination.

Inferring genotypic probabilities in complex pedigrees (those with marriage or inbreeding loops) is now a standard procedure using algorithms based on the method of 'peeling' (Elston & Stewart, 1971; van Arendonk *et al*., 1989; Fernando *et al*., 1993; Stricker *et al*., 1995; Janss *et al*., 1995*b*; Kerr & Kinghorn, 1996). This process is often referred to as segregation analysis. Peeling can be used to produce unbiased genotypic probabilities for small complex pedigrees, but peeling may be infeasible for large complex pedigrees. In this case the pedigree may be simplified by cutting loops, or an iterative peeling algorithm may be used, but some bias may be introduced into genotypic or IBD probability estimates (Fernando *et al*., 1993). Lacking exact methods for large complex pedigrees, potential biases resulting from the use of pedigree simplification or iterative peeling are commonly ignored.

For very large complex pedigrees with multiple alleles at each locus, where exact peeling is infeasible, Markov Chain Monte Carlo (MCMC) approaches are often advocated (Guo & Thompson, 1994; Janss *et al*., 1995*a*). Many of these MCMC algorithms work

* Corresponding author. Current address: Animal Genetics and Breeding Unit, University of New England, Armidale, NSW, 2351, Australia. Tel: +61 +2 6773 3979. Fax: +61 +2 6773 3266. e-mail: jhenshal@metz.une.edu.au
† AGBU is a joint institute of NSW Agriculture and The University of New England.

on sampling in the space of descent graphs. A descent graph for the locus of interest is a model which specifies which allele (grandpaternal or grandmaternal) was inherited by each individual from each parent. A legal descent graph (LDG) is one which is consistent with the observed genotypic data. By linking all gametes of known allelic type to a base gamete, a descent graph determines the allelic type of base gametes. Prior belief about allele frequencies in the base population can then be used to determine the likelihood of any particular descent graph (Sobel & Lange, 1996). For a multilocus descent graph, the likelihood is also a function of observed and expected numbers of recombinations between adjacent loci (Sobel & Lange, 1996).

As a starting point, many MCMC algorithms require a descent graph which is consistent with the observed genotypic data (Sobel & Lange, 1996). The choice of starting LDG may be critical to the success of the MCMC algorithm because, while it is theoretically possible for the Markov chain to move from any legal graph to any other legal graph, in practice some transitions may be extremely unlikely (Sobel & Lange, 1996). This may prevent the chain from moving from a local optimum to a far more likely global optimum.

The problem of obtaining a LDG is closely related to the problem of estimating IBD, and is no easy matter for a large complex pedigree. The space of descent graphs is so large, and the proportion of descent states which are legal so small, that the chance of arriving at a LDG through the use of simple methods (such as gene dropping) is negligible.

An iterative, elimination approach was suggested by Sobel & Lange (1996). The genotype elimination algorithm of Lange & Goradia (1987) is used to generate an ordered legal genotype sample (in which the origin of each allele, maternal or paternal, is specified). An individual with multiple possible genotypes is then chosen at random, and a genotype assigned at random from those feasible. The genotype elimination algorithm is then applied to the newly constrained pedigree. This is repeated until all individuals have only a single legal ordered genotype. A LDG can easily be derived from the ordered legal genotype sample.

In the algorithm of Sobel & Lange (1996), a single genotype is selected from those currently feasible for the chosen individual at each exclusion step. No account is taken of the relative genotypic probabilities, but selecting a genotype of low probability reduces the chance of the algorithm arriving at a LDG. Heath (1998) proposed including a peeling step so that the genotype could be drawn from the approximate distribution of genotypes.

If it were possible to sample LDGs directly from the equilibrium distribution of genotypes, then not only would such samples be ideal for use as starting values in multi-locus MCMC algorithms, but successive independent samples could be used to obtain a description of the equilibrium distribution. Estimating genotypic probabilities and IBD probabilities using this method would avoid the problems with biased estimates due to poor starting values. This approach was attempted by Henshall *et al.* (1999), who proposed the sampling of the origin (grandpaternal or grandmaternal), or inheritance state, of each gamete instead of sampling the genotype of an individual as in the genotype elimination algorithms of Sobel & Lange (1996) and Heath (1998). The genotype elimination through inheritance constraint (GEIC) algorithm (Henshall *et al.*, 1999) also included a weight function, in an attempt to ensure that the mean of the samples was an unbiased estimator of genotypic and IBD probability. Samples were weighted according to the frequency with which the algorithm was likely to find them. However, their algorithm failed to account correctly for the effect of sampling base genotypes, and was unable to incorporate other than uniform prior allele frequencies in the base population.

In the GEIC algorithm of Henshall *et al.* (1999), a gamete is chosen at random, and a random inheritance state assigned (unless of course only one inheritance state is feasible). The consequences of this action are then explored using the genotype elimination algorithm (Lange & Goradia, 1987), modified to take account of inheritance state constraints. This process is repeated until no unconstrained inheritance states remain. However, with random choice of gametes to sample, the proportion of infeasible solutions can be high. Also, the algorithm is slower than necessary, because it fails to take account of the fact that, for many gametes, the inheritance constraint chosen is irrelevant. It is therefore unnecessary to sample inheritance constraints for all gametes. While it can be determined in advance that some gametes need never be sampled, whether or not sampling is required for other gametes will depend on the sampled inheritance state of other gametes.

In this article, an improved GEIC algorithm is proposed. Compared with the original algorithm, the improved algorithm for sampling descent graphs is faster and produces fewer infeasible samples. This allows the algorithm to be used with larger complex pedigrees. Allelic types for base gametes are then sampled to produce a descent state sample. The likelihood of this sample is computed using prior allele frequencies for the base population. Importance sampling is then used to produce genotypic and IBD probability estimates. Thus the improved algorithm correctly accounts for the sampling of genotypes for base individuals, and so produces unbiased estimates. The unbiasedness of the improved method is demonstrated using small datasets. The method's ability to produce legal descent states in large complex pedigrees

is also demonstrated, and the limitations of importance sampling for such pedigrees discussed.

2. Method and application

(i) *The sampling algorithm*

At any locus, individuals inherit either a grandpaternal or a grandmaternal gamete from each parent. There are two constraints which can apply to a gamete. Firstly, genotypic information from the individual, or phenotypic information coupled with a penetrance function, might restrict its allelic type to a single allele if the individual is homozygous, to one of two alleles if the individual is heterozygous, or possibly to a subset of available alleles via the penetrance function. We will refer to such gametes as informative gametes. The second restriction relates to the grandpaternal or grandmaternal origin, or inheritance state of the gamete. For some gametes, observed genotypic information is sufficient to allow the inheritance state to be determined. These gametes have constrained inheritance states. The inheritance states of all other gametes are unconstrained by the genotypie data.

To produce a sample LDG, it is sufficient to construct a set of paths connecting every informative gamete in the pedigree to a base gamete, such that Mendelian inheritance rules are not violated. Each path will consist of inheritance states which are either constrained by genotypic data or constrained through sampling. The inheritance states of uninformative gametes which do not lie on any of these paths are not constrained, and can be assigned a random value.

The modified GEIC algorithm to sample a single LDG and associated genotype sample is as follows (an example of steps 1 to 7 of the algorithm is shown in the Appendix):

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. This set will be referred to 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 LDG. With base gametes uniquely determined (steps 6 and 7) and a LDG, the allelic type of all gametes in the pedigree is also uniquely determined, and comprise a legal descent state.

(ii) *Probability distribution junction of the samples*

As this density will be used in importance sampling it will be denoted $I(x)$. With the algorithm described above, LDGs and states may not be sampled with equal probability. The variation comes from two sources. First, when some regions of the space of descent graphs are illegal, some legal samples may require fewer elimination steps (steps 3c and 4) than others. Samples requiring fewer elimination steps are found more frequently than samples requiring more elimination steps. Second, when prior allele frequencies for the base population are other than uniform, steps 6 and 7 introduce variation in the probability of legal descent states being sampled.

Variation in the number of elimination steps required to produce a sample can be incorporated in

https://doi.org/10.1017/S0016672301005316 Published online by Cambridge University Press

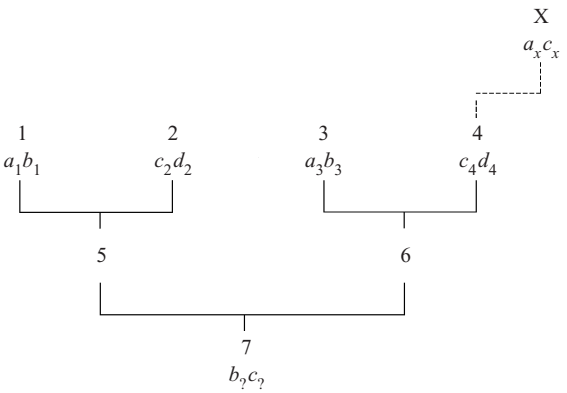

Fig. 1. A small pedigree, with alleles a, b, c and d, with subscripts indicating individual of origin. For this pedigree, the number of descent graphs consistent with individual 7 having inherited allele c from individual 4 is reduced by 1 if individual X is included, while the number of descent graphs consistent with individual 7 having inherited allele c from individual 2 is unchanged.

the probability distribution function for the sampling algorithm by noting that the search process is a series of binary choices between grandpaternal and grandmaternal inheritance state. Therefore a sample requiring s steps will be found twice as often as a sample requiring $s+1$ steps. The probability of finding a particular descent graph is then proportional to 2^{-s_i} where s_i is the number of steps used in obtaining descent graph x_i. Thus

$$I(x_i) \propto I^*(x_i) = 2^{-s_i} \prod_{j=1}^{n_i} p_{ij} \tag{1}$$

where n_i is the number of implementations of step 6 or 7 in obtaining sample x_i and p_{ij} is the prior allele frequency for the allele sampled in each of those steps.

(iii) *Likelihood of a descent state*

The likelihood of a descent state is affected by sampled and expected allele frequency in the base population, and, for multilocus descent graphs, sampled and expected recombination (Sobel & Lange, 1996). This likelihood will be denoted $g(x)$.

It is important to note that, even for a single locus pedigree with genotypes observed on all base individuals, all LDGs (as distinct from descent states) may not be equally likely. This is demonstrated in the pedigree in Fig. 1, a simple pedigree with eight gametes in the base population (individuals 1, 2, 3 and 4), having alleles a, b, c and d, with subscripts indicating the base individual of origin. The genotypes given are unordered; thus for individual 1, allele a_1 could be either maternal or paternal in origin. Individual 7 can have inherited either alleles b_1 and c_4, or alleles b_3 and c_2. It is clear that these two possibilities are equally likely. If we introduce individual X to be a parent of individual 4, with observed genotype ac, then the genotypes $(b_1, c_4(=c_X))$ and (b_3, c_2) for individual 7 should still be equally likely. However, the number of LDGs consistent with genotype (b_3, c_2) is one more than the number of LDGs consistent with genotype (b_1, c_4).

In the algorithm described above, variation in the likelihood of descent graphs is correctly accounted for in the formula for the sampling distribution of descent states (equation 1), specifically by the term $\prod_{j=1}^{n_i} p_{ij}$. However, the above example illustrates the importance of obtaining a descent state sample, even if the purpose is to estimate IBD probabilities rather than genotypic probabilities. MCMC descent graph sampling algorithms also need to take account of this variation in the likelihood of descent graphs; for example, the algorithm of Sobel & Lange (1996) requires that lists of 'founder tree graphs' be maintained for this purpose.

(iv) *Calculating IBD and genotypic probabilities*

With the method described above, each sampled descent state has associated with it a value $I^*(x_i)$, which is proportional to the probability distribution function $I(x_i)$, and a likelihood $g(x_i)$. Importance sampling can be used to obtain estimates of IBD or genotypic probabilities. With importance sampling, an estimate of genotype probability or IBD probability $\hat{J}(y)$ is given by

$$\hat{J}(y) = \frac{\sum f(y \mid x_i) w(x_i)}{\sum w_i}$$

where $f(y \mid x_i)$ is the observed genotype or IBD for sample x_i and $w(x_i) = g(x_i)/I^*(x_i)$ is a weight function (see, for example, Geweke (1989) or Tanner (1993)). As noted by Geweke (1989), it is not necessary to normalize the importance sampling density $I^*(x)$.

Geweke (1989) provides expressions for the Monte Carlo standard error of $\hat{J}(y)$ and for the number of effective samples; however, these are subject to conditions which may not apply to all pedigrees. Accordingly, to evaluate the accuracy of sampled probabilities, the number of effective samples will be approximated by

$$m = \frac{\sum w_i}{w_{\max}} \tag{2}$$

where $w_{\max}$ is the maximum weight observed. This effective number of samples will be less than the total number of samples if $(g(x_i)/I(x_i))$ is not constant for all i.

https://doi.org/10.1017/S0016672301005316 Published online by Cambridge University Press

Table 1. *Pedigree A: for each analysis, genotype was available for only four randomly chosen individuals*

ID	Father	Mother	Genotype
1	0	0	AD
2	0	0	BD
3	0	0	AC
4	2	1	AB
5	2	3	CD
6	2	4	BB
7	5	4	BC
8	5	3	CC
9	7	6	BC
10	7	8	BC
11	10	9	CC

(v) *Test data*

Pedigree A was a simulated 11 individual pedigree, with a number of inbreeding loops. A marker locus with 4 alleles (A, B, C and D) was simulated, with genotype assigned to all individuals (Table 1). From this pedigree, 100 datasets were generated. In each of these datasets four individuals were randomly chosen to be 'genotyped'. The genotypes of the other seven individuals genotype were treated as unknown.

For datasets from pedigree A, it was possible to compare the estimated genotypic probabilities with exact probabilities obtained using the software package MENDEL (Lange *et al*., 1988). The accuracy of the genotypic probability estimates was assessed using the summary statistic $\chi^2 = \sum_{E_{kl}\neq 0}((O_{kl}-E_{kl})^2/E_{kl})$, where k relates to the individual, l is the genotype (e.g. AA, AB, AC, ...), E_{kl} is the expected number of samples 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.

Four larger pedigrees were also simulated. Dataset B is modelled on the simulated pedigrees described by Heath (1998). Twenty generations were simulated, with 16 individuals selected to produce 80 offspring each generation. A 16 allele marker locus was simulated, with genotype records made available on the individuals in the first two generations and the last two generations. Dataset C is similar to dataset B, differing only in that in each generation 32 individuals were selected to produce 160 offspring. Datasets D and E differed from datasets B and C respectively only in that genotype records were not made available for the first two generations, that is, genotype records were only available for individuals in the last two generations.

3. Results

(i) *Small dataset*

For datasets from pedigree A with 11 individuals, 10000 LDG samples were obtained using the modified GEIC algorithm. Prior allele frequencies for alleles A, B, C and D were assumed to be 0·5, 0·25, 0·2 and 0·05 in the base population. The effective number of samples ranged from 112 to 10000, with a mean of 2703. The results were compared with exact probabilities obtained using MENDEL (Lange *et al*., 1988). Fig. 2 plots the values of the summary statistic against the degrees of freedom. In only 1% of replicates is the test statistic significant at the 5% level, if a χ^2 distribution is assumed. The results for the worst replicate, with a test statistic of 61·4, 41 degrees of freedom and 7562 effective samples, are displayed in Table 2. For this dataset, genotypes were known for individuals 2, 3, 5 and 7. Although significant at the 5% level, the genotypic probability estimates in Table 2 appear sufficiently accurate for most purposes.

(ii) *Large datasets*

The modified GEIC algorithm described in this article was used to produce 1000 LDG samples from datasets B, C, D and E. Table 3 contains the percentage of legal samples obtained, and lists for comparison the relevant results published by Heath (1998). For dataset B, legal samples took on average less than 4 seconds on a Pentium II 350 MHz computer. This appears to compare favourably with the 1 to 2 minutes reported by Heath but, as the computers used differ, no direct comparison is available.

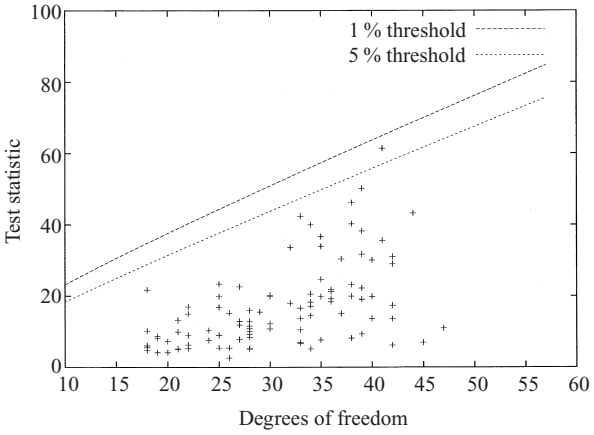

Fig. 2. Test statistics obtained from analysis of 100 datasets sampled from pedigree A, with weights applied. In each analysis, 10000 legal samples were drawn. Also shown are the 5% and 1% significance thresholds for a χ^2 distribution.

https://doi.org/10.1017/S0016672301005316 Published online by Cambridge University Press

Table 2. *Genotypic probability estimates for the worst analysis from pedigree A, with exact probabilities obtained through peeling in brackets*

ID	Allele	A	B	C	D
1	A	0·173 (0·167)	0·326 (0·333)	0·131 (0·133)	0·036 (0·033)
	B	–	0·121 (0·125)	0·133 (0·133)	0·034 (0·033)
	C	–	–	0·030 (0·027)	0·013 (0·013)
	D	–	–	–	0·003 (0·002)
4	B	0·340 (0·333)	0·328 (0·333)	0·138 (0·133)	0·193 (0·200)
6	B	0·085 (0·083)	0·336 (0·333)	0·037 (0·033)	0·376 (0·383)
	D	0·082 (0·083)	–	0·033 (0·033)	0·050 (0·050)
8	A	–	–	0·247 (0·250)	0·249 (0·250)
	C	–	–	0·255 (0·250)	0·249 (0·250)
9	B	0·046 (0·042)	0·298 (0·292)	0·312 (0·308)	0·141 (0·150)
	C	0·039 (0·042)	–	0·017 (0·017)	0·147 (0·150)
10	B	0·121 (0·125)	–	–	0·121 (0·125)
	C	0·122 (0·125)	0·249 (0·250)	0·259 (0·250)	0·129 (0·125)
11	A	0·005 (0·005)	0·076 (0·078)	0·055 (0·054)	0·021 (0·024)
	B	–	0·128 (0·135)	0·349 (0·338)	0·104 (0·105)
	C	–	–	0·134 (0·133)	0·112 (0·108)
	D	–	–	–	0·017 (0·019)

Genotypes are unordered; rows and columns do not indicate parental origin. There were four alleles in the pedigree A, B, C and D, with prior probabilities 0·5, 0·25, 0·2 and 0·05 respectively. Estimates for individuals 2, 3, 5 and 7 are omitted as genotypes are known with certainty to be AB, CD, BB and BC respectively.

Table 3. *Percentage of samples drawn which were legal using the modified GEIC algorithm, for datasets B, C, D and E. The equivalent results from Heath are also shown*

	Heath's results		
Dataset	Method 1	Method 2	Modified GEIC
B	68·5	16·5	99·0
C	na	na	80·8
D	na	na	98·9
E	na	na	79·1

In Heath's method 1 the individual to be constrained was chosen to be the individual with the fewest feasible genotypes, while in Heath's method 2 the individual to be constrained was chosen at random.

4. Discussion

When applied to small complex pedigrees, for which exact genotypic probabilities can be obtained using peeling, the difference between the exact genotypic probabilities and those estimated with the modified GEIC algorithm is statistically significant no more frequently than would be expected by chance. This suggests that the LDG samples obtained using the modified GEIC algorithm are close to the equilibrium distribution. This should also be the case for larger pedigrees, provided that the use of importance sampling remains valid.

With importance sampling, genotypic and IBD probabilities are estimated as the weighted means of a number of independent samples. The weights have two components: the posterior density and the importance sampling density. It is important that the importance sampling density mimics the posterior density for the method to be effective (Geweke, 1989). For a descent state sampled using the method described here there are two parts to the importance sampling density. The first of these, due to variation in the probability with which a descent graph sample is found, is determined by the algorithm and the data, and can potentially diverge significantly from the posterior density. The second part, due to sampling base alleles, is chosen to be in accord with the posterior density.

In practice, as noted by many authors including Hastings (1970) and Geweke (1989), poor behaviour occurs when a small number of samples have associated with them very large weights. For very large pedigrees, where significant variation in the number of steps required to obtain a sample can be expected, there is indeed a risk that the estimates will be dominated by a small number of samples. This is exactly what happened when the algorithm was applied to the larger pedigrees B, C, D and E, for which the effective number of samples after 1000 samples was less than 2. As there are a finite number of possible descent graphs for any pedigree, in theory if the algorithm is left running, sufficient effective samples will eventually be obtained. However, the

https://doi.org/10.1017/S0016672301005316 Published online by Cambridge University Press

required number of samples may be very large, making this approach infeasible. For pedigree B, increasing the number of samples to 10000 only changed the number of effective samples to a little over 2.

As samples are independent, some speedup can be achieved by running multiple processors in parallel. Another way of increasing the number of samples obtained is suggested by noting that the main cost of the algorithm is in sampling the primary descent graph. Sampling base genotypes is relatively fast, and sampling secondary descent graphs is very fast, when compared with the process of obtaining the primary descent graph sample. This makes it feasible to obtain a modest number of base genotype samples and many secondary descent graph samples for each primary descent graph sample. However, this does not overcome the fundamental problem of a small number of primary descent graph samples having very high weights associated with them.

A partial solution to the problem of small numbers of effective samples exists if one is prepared to accept some potential bias in estimates obtained from the weighted samples. Inflating the importance sampling density of the samples with the smallest density increases the effective number of samples, but may cause some bias. For example, with the 10000 samples from pedigree B mentioned above, setting the importance sampling density for the 50 samples with smallest importance sampling density to the density of the 51st ranked sample increased the effective number of samples to 89·5. That is, 9950 samples with correct weights relative to each other contributed 39·5 to the effective number of samples, and 50 unweighted samples contributed another 50 to the effective number of samples.

The degree of bias introduced through manipulating the importance sampling density will be pedigree-dependent. Potential exists to minimize the bias through the application of algorithms more sophisticated than setting a floor on the importance sampling density as described above. Whether the bias described here is more significant than the bias caused by the pedigree simplification or iteration in peeling algorithms is unknown, and worthy of further research. For large complex pedigrees, it is also possible that genotypic and IBD probabilities produced by MCMC algorithms have poor characteristics, due to difficulties in traversing the parameter space. Running multiple chains with different starting values is of benefit, but if the chains truly do not communicate, then the sampling density of the starting values is relevant, and the problems described above are equally applicable.

Despite the limitations of the algorithm when applied to the problem of estimating IBD and genotypic probabilities for large pedigrees, it remains a fast and efficient method for producing descent graphs for use as starting values for other algorithms. The algorithm produces a lower proportion of invalid solutions for dataset B than reported by Heath (1998), and appears to be competitive with regard to time per sample.

5. Conclusion

The modified GEIC algorithm is able to obtain LDG samples for large and complex pedigrees. For small pedigrees it has been shown that with importance sampling, these samples have been drawn from a distribution which approximates the equilibrium distribution very well. Where the pedigree size is such that importance sampling is not effective, then the algorithm is ideal for producing legal descent graphs for use as starting values in other algorithms. No pedigree simplification is required for large complex pedigrees, and, where feasible, obtaining genotypic probabilities as the average of a large number of independent samples is preferable to the MCMC approach of exploring the space surrounding a small number of starting values.

Appendix

Using the algorithm to identify a primary descent graph sample for the population in Table 1, with genotype known for individuals 1, 5, 6 and 10

Step 1: A list of informative gametes in reverse order:

10p(B,C), 10m(B,C), 6p(B), 6m(B), 5p(C,D), 5m-(C,D), 1p(A,D), 1m(A,D) where 10p(B,C) and 10m-(B,C) are individual 10's paternal and maternal gametes, each of which has either allele B or allele C.

Step 2: A list of feasible ordered genotypes for each individual, paternal followed by maternal (9 and 11 are ignored as they are not required for the primary set; * indicates any allele):

1 [AD,DA], 2 [BC,CB,BD,DB], 3 [C*,*C,D*,*D], 4 [BA,BD], 5 [CD,DC], 6 [BB], 7 [CA,CD,CB,DB], 8 [C*,DB,DC], 10 [BC,CB].

Step 3: Link informative gametes to founders. The sequence of choices for step 3 is shown by the numbered paths between parents and progeny in Fig. A1.

Gamete 10p(B,C):

Step 3(c): Choose 10p to be from 7m. (*i*)
Step 3(d): 7 is constrained to [CB,DB], 8 to [C*,DC] and 10 to [BC].

https://doi.org/10.1017/S0016672301005316 Published online by Cambridge University Press

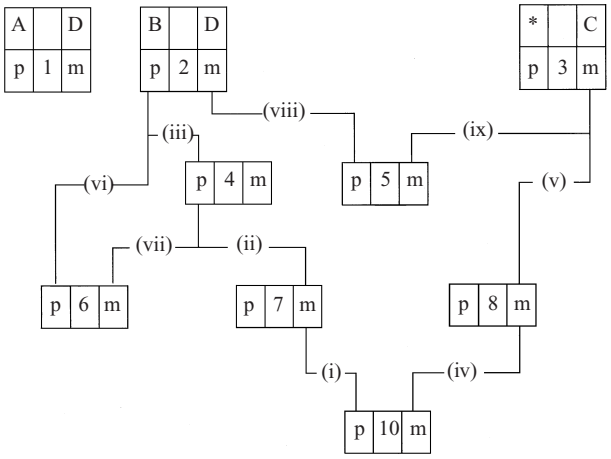

Fig. A1. A primary descent graph sample for the pedigree in Table 1, with genotype known on individuals 1, 5, 6 and 10. Paternal (p) and maternal (m) origin is indicated, along with the allele (A, B, C, D or unsampled*) for base individuals. The numerals (i to ix) relate to the steps of the algorithm, given in the Appendix.

Step 3(a): 7m must be from 4p. (*ii*)
Step 3(c): Choose 4p to be from 2p. (*iii*)
Step 3(d): 2 is now constrained to [BC,BD].
Step 3(b): Next informative gamete is 10m(C)
Step 3(c): Choose 10m to be from 8m. (*iv*)
Step 3(d): 3 is constrained to [C*,*C], and 8 to [CC,DC].
Step 3(c): Choose 8m to be from 3m. (*v*)
Step 3(d): 3 is constrained to [*C].
Step 3(b): Next informative gamete is 6p(B)
Step 3(a): 6p must be from 2p(B). (*vi*)
Step 3(b): Next informative gamete is 6m(B)
Step 3(a): 6m must be from 4p(B). (*vii*)
Step 3(b): Next informative gamete is 5p(C,D)
Step 3(a): 5p must be from 2m. (*viii*)
Step 3(b): Next informative gamete is 5m(C,D)
Step 3(c): Choose 5m to be from 3m(C). (*ix*)
Step 3(d): 2 is constrained to [BD], and 5 to [DC].

Now all informative gametes have been linked to founders.

Step 4: Not required for primary descent graph.

Step 5: A list of base gametes with more than one possible allelic type.

1p(A,D), 1m(A,D)

Step 6: Choose 1p(A) with consequence 1m(D).

Step 7: In this descent graph sample, gamete 3p is unobserved and therefore can be sampled without checking for consequences.

References

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.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* **57**, 1317–1339.

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

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

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.

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

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.

Tanner, M. A. (1993) *Tools for Statistical Inference*, 2nd edition. New York: Springer

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.

https://doi.org/10.1017/S0016672301005316 Published online by Cambridge University Press