© INRA, EDP Sciences, 2006
DOI: 10.1051/gse:2006032

Original article

# Improved techniques for sampling complex pedigrees with the Gibbs sampler

K. Joseph ABRAHAM[a*], Liviu R. TOTIR[b], Rohan L. FERNANDO[b,c]

[a] 1301 Agronomy Hall, Iowa State University, Ames, IA 50011, USA
[b] Department of Animal Science, Iowa State University, Ames, IA 50011, USA
[c] Lawrence H. Baker Center for Bioinformatics and Biological Statistics,
Iowa State University, Ames, IA 50011, USA

(Received 5 January 2006; accepted 3 September 2006)

**Abstract –** Markov chain Monte Carlo (MCMC) methods have been widely used to overcome computational problems in linkage and segregation analyses. Many variants of this approach exist and are practiced; among the most popular is the Gibbs sampler. The Gibbs sampler is simple to implement but has (in its simplest form) mixing and reducibility problems; furthermore in order to initiate a Gibbs sampling chain we need a starting genotypic or allelic configuration which is consistent with the marker data in the pedigree and which has suitable weight in the joint distribution. We outline a procedure for finding such a configuration in pedigrees which have too many loci to allow for exact peeling. We also explain how this technique could be used to implement a blocking Gibbs sampler.

**Gibbs sampler / Markov chain Monte Carlo / pedigree peeling / Elston Stewart algorithm**

## 1. INTRODUCTION

The calculation of the likelihood plays an important role in the analysis of genetic data, for example in linkage analysis. Apart from the likelihood, other probability functions such as marginal distributions for certain genotypes of certain individuals are also needed for example in genetic counseling. In many instances, the likelihood, which is proportional to the probability of the observed phenotypic data, can be written as a product of probabilities summed over all possible genotype configurations.

For a trait with $m$ alleles in a population with $N$ individuals, the number of genotypes to be summed over in evaluating likelihoods could be as large as $\{m(m+1)/2\}^N$, clearly a huge number even when $m$ is 2 and $N$ is moderately

[*] Corresponding author: abraham@iastate.edu

large. This number may often be reduced by a large factor if some of the individuals in the population are genotyped [13], but even in such cases is still dauntingly large. Nonetheless, if the inheritance is monogenic and if the pedigree has no loops, the sum over genotypes can be computed easily along the lines of the Elston-Stewart algorithm [4], which is often referred to as peeling. If the pedigree is not too large (about 100 members) and does not have too many loops, extensions of the Elston-Stewart algorithm have been developed for evaluating the likelihood [2,9,11,12,16–18]. This situation arises in human genetics, however, in animal pedigrees the number of individuals may easily reach several hundred and interbreeding loops are common. Furthermore, even if the pedigree structure is not too complex, exact peeling may not be possible if there are many loci present, which is often the case in human genetics.

If exact peeling over all genotypic configurations is not possible, one alternative is to use MCMC procedures to sample genotypic configurations according to the posterior distribution. Among the many samplers in use, the single Gibbs sampler site is possibly the easiest to implement. However, the single Gibbs sampler site frequently has problems with reducibility [1], and poor mixing [8]. The mixing and reducibility problems are not as severe in more complex variants of the Gibbs sampler such as the Blocking Gibbs sampler, as will be discussed later. Quite apart from these problems, a valid starting configuration is needed to initiate the Gibbs sampler, *i.e.* a configuration of genotypes which is consistent with the pedigree and marker information is needed. In order to avoid convergence problems, the starting configuration should also be one whose statistical weight is not too small; this requirement can be hard to fulfill if there are many tightly linked loci in the data set. One way of obtaining a starting configuration when exact peeling over all loci and all individuals is not possible, is to peel as much as possible and then condition on suitably chosen genotypes, as has been implemented for a single locus in Heath [7]. In Heath [7] the framework employed is genotypic sampling which is difficult to extend to multilocus data sets.

In this investigation, we discuss in detail an alternative procedure which uses allelic, as opposed to genotypic, variables to handle multilocus data sets. Our procedure also relies on peeling and conditioning to find not only a valid starting configuration but also a starting configuration with reasonable statistical weight in cases where there are too many loci for the pedigree to be peeled, necessitating the use of samplers. This situation can be expected to arise in human genetics where marker maps are dense. After presenting numerical results for the situation just described, we will also discuss the extensions of this idea to situations where exact peeling cannot be implemented not just due to the large number of loci but also because of the presence of a large number of loops between individuals. A central concept in our discussion is the notion that certain marginal distributions can be calculated accurately and with relative ease by truncating the full pedigree. These marginal distributions involve variables which are located at some distance from where we truncate the pedigree. This observation relies on the Markov property of probability functions of interest as well as the notion of distance in graph theory. Once we have a reliable estimate for marginal probabilities for these variables, we can sample and condition on these variables which in turn facilitates peeling and conditioning on the full pedigree. Since our initial sampling and conditioning was to a good approximation from the joint distribution of the full pedigree, our subsequent conditioning can also be expected to be from the joint distribution of the full pedigree. By this divide and conquer scheme, we are in a position to reliably sample from the joint distribution for the full pedigree without having to peel the entire pedigree.

## 2. MATERIALS AND METHODS

A vital preliminary step in our discussion is the notion that a pedigree can be represented by an undirected graph with weights associated with each vertex. If we first consider just a single locus, then each vertex in the graph represents an individual, and the edges linking vertices will depend on familial relationship. Thus there will be edges linking any non founding individual to its parents and to its offspring as well as spouses. Once we have a theoretical graph representation of our pedigree, we can discuss the notion of distance along a graph which will play a crucial role. We adopt the definition of distance between two vertices as being the shortest number of edges that need to be traversed to move from one vertex to the next. Following this definition of distance, there is no distance between a node and itself and there is a separation of one unit of distance between any individual and its spouses, offspring or parents. There are two units of distances separating grandchildren and their grandparents, assuming no inbreeding between generations.

Next we note that the calculation of probability functions such as likelihoods involves products of conditional probabilities such as transmission probabilities which involve vertices that are just one unit of distance apart, founder probabilities and penetrance functions which involve just one vertex. This is the Markov property alluded to earlier. It is reasonable to assume on the basis of this Markov property that any alteration to the pedigree will be only weakly felt when calculating probability functions for vertices which are far

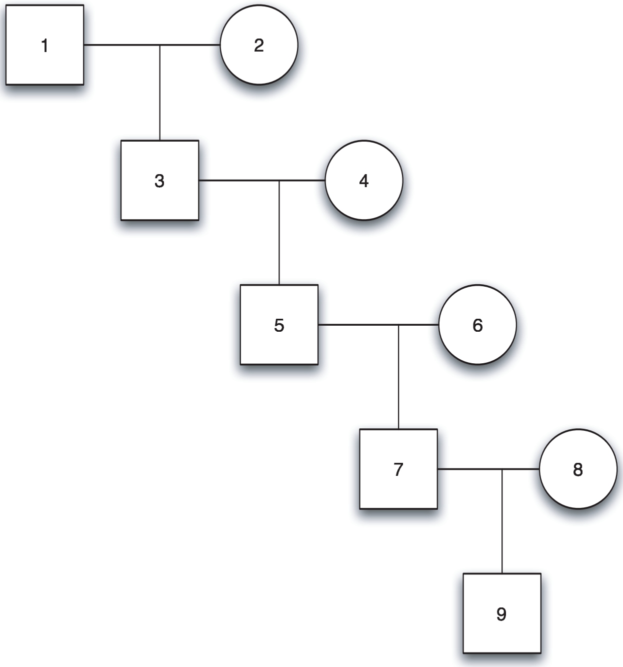

**Figure 1.** A trial pedigree.

away from the location where the pedigree is altered. This gives us a rule of thumb to assess which marginal distributions may be altered when truncating a pedigree at a specified location. To make the notion of "far away" more concrete, let us consider the pedigree in Figure 1 with 9 individuals and assume just one locus with three alleles.

It is easy to see that individuals 8 and 9 are 4 units of distance away from individuals 1 and 2. We simulate marker information for the entire pedigree keeping individuals 1, 2, 3, 7 and 8 ungenotyped. As a result of individuals 1, 3 and 7 being ungenotyped, the genotype frequencies for individual 1 are potentially quite sensitive to the marker information at individual 9. Conditional on our simulated marker information we can calculate genotype probabilities for individuals 1 and 2. Then we remove individuals 8 and 9 from the pedigree and recalculate the same genotype probabilities as before. Since we have altered the pedigree 4 distance units away from the vertices we are interested in, we expect both sets of genotype probabilities to be very similar. This is indeed the case as is apparent from the results in Table I where we display the marginal allele frequencies for individual 1 with and without truncating the pedigree. The numbers in parentheses indicate the values obtained from

**Table I.** Comparing allele frequencies in full and truncated pedigrees.

| Allele | Maternal allele frequency | Paternal allele frequency |
|---|---|---|
| 1 | 0.2496 (0.2422) | 0.2516 (0.2478) |
| 2 | 0.2564 (0.2633) | 0.2570 (0.2633) |
| 3 | 0.4940 (0.4945) | 0.4914 (0.5042) |

the full pedigree; in each case the pedigree was peeled exactly (*i.e.* making no modifications to the pedigree) and 10 000 samples were generated.

The effects of truncating the pedigree are indeed small, as expected. In this very simple instance the distance between vertices could be determined visually, in more complex instances the distance between vertices can be efficiently obtained using the breadth first algorithm [3].

Now we consider the more realistic situation of a pedigree with $L$ loci with varying map distances between the loci and varying numbers of alleles at each locus. In this case the vertices in the graph do not correspond to individuals but rather to allelestate nodes and alleleorigin nodes [5, 6]. Allelestate nodes are nodes which store information on the permissible alleles at a given locus, while alleleorigin nodes contain segregation information. (Note that our allelestate nodes and alleleorigin nodes correspond to the genetic loci and selector variables of [5] respectively). The edges between the nodes are filled by determining which nodes arise in some function appearing in the likelihood calculation. For example, the maternal and paternal allelestate nodes of a given individual at the same locus both arise in the calculation of penetrance, thus there is an edge linking these two nodes and these two nodes are one unit of distance apart. For any non-founder individual at any locus there is a transmission probability which involves the maternal allelestate, the maternal allelestate and the paternal allelestate of the mother along with the maternal alleleorigin for that locus. All these nodes are therefore connected to one another by edges and the maternal alleleorigin node is one unit of distance removed from the maternal allelestate node for a fixed locus for any non-founding individual. Once again the separation between any two nodes is determined from the number of edges in the shortest path connecting the two nodes. It is easy to see that there is at least one unit of distance between any two nodes corresponding to adjacent loci, since recombination probabilities involve alleleorigins of a given individual at two adjacent loci. If the nodes correspond to different individuals, then the separation will be larger. Similarly, if the loci corresponding to two nodes are $i$ and $i + j$, then the two nodes are at least $j$ units of distance apart.

We now assume we can (or are willing) to exactly peel over the first $m$ loci where $m < L$. Thus we truncate the pedigree keeping just the first $m$ loci. Since we peel over $m$ loci, we can by reverse sampling condition on all $m$ loci, or any subset $s$ of the $m$ loci, which we have just peeled. If all the nodes in the subset $s$ are far away from where we have truncated the pedigree, we can sample and condition on these nodes. This sample is, (from our previous example) to a good approximation, a sample from the marginal over $s$ of the joint distribution of the untruncated pedigree. Once we have a subset $s$ conditioned, it may be possible to exactly peel over the rest of the pedigree conditional on our initial sample. If peeling over the rest of the pedigree after conditioning on the nodes in $s$ can be carried out with no approximation we can by reverse peeling, sample all the loci in the problem to obtain a starting configuration for a Gibbs sampler. Since the initial sampling (*i.e.* over the subset $s$) is to a good approximation from the marginal of the full distribution over all $L$ loci, then our sample will also be drawn to a good approximation from the full distribution over all $L$ loci and will thus be a good starting configuration for the Gibbs sampler. We have implicitly assumed that all loci can be sampled once we have just one initial sample over just a few loci, *i.e.* we are assuming we can obtain a sample over all loci in just two stages. Later on we will discuss the consequences of relaxing this assumption.

## 3. RESULTS

We now assume for the sake of concreteness that the total number of loci $L = 10$ and the number of loci that we peel exactly $m = 8$. Our set $s$ is over the first two loci. We neglect loci 9 and 10 and peel over the first eight loci and consider just the joint marginal distributions over just the first two loci (*i.e.* loci 1 and 2). Both of these loci are at a substantial distance from the location where we have altered the pedigree by neglecting loci 9 and 10. More precisely each of the nodes in the first two loci is a minimum of 6 units of separation from where we modified the pedigree. Based on our earlier reasoning we expect that the joint marginal distribution of the first two loci will be only weakly affected by truncating the pedigree and is a good approximation to the joint marginal distribution for the full pedigree. To test this idea, we consider a real dataset with 555 individuals and marker information at ten highly polymorphic markers [14, 15], which includes a sprinkling of missing marker information. The pedigree was generated by crossing two Berkshire grand sires and nine Yorkshire grand dams and includes 499 F2 progeny from 45 F1 matings. In order to estimate the number of loops in the pedigree we consider the same pedigree with just one biallelic locus with no marker data. We deliberately restrict the amount of memory available in the peeling process requiring a loop to be cut whenever there is a memory bottleneck preventing peeling. From keeping track of the number of loops which get cut we get a lower bound on the number of loops in the pedigree; our results indicate that the pedigree has over a thousand loops. If we greatly relax restrictions on memory usage the entire pedigree can be peeled by brute force using a greedy heuristic [11] to determine the peeling order, thus we are able to obtain the true joint marginal haplotype distribution for the first two loci for each individual in the pedigree. Next we ignore loci 9 and 10 and obtain a joint marginal haplotype distribution for the first two loci for each individual in the pedigree. Each joint marginal is based on a sample of size 10 000. In order to compare frequencies of different haplotypes in both samples we ignore all haplotypes in each sample where the frequency is 1 or less than 0.001. We observe that in each sample the haplotypes which survive after making these selection criteria are identical. The frequencies of the haplotypes considering all ten loci and considering all eight loci are stored in vectors **V** and **U** such that $\mathbf{V}[i]$ and $\mathbf{U}[i]$ correspond to the frequencies of the same haplotype for the same individual in each sample. We construct the quantities $\mid \mathbf{V} \mid = \sqrt{\sum \mathbf{V}[i]^2}$ and $\mid \mathbf{U} \mid = \sqrt{\sum \mathbf{U}[i]^2}$ where the sum is over all elements of each vector. Then we compute **W** which is given by $\sum (\mathbf{V}[i] - \mathbf{U}[i])^2$. Finally we calculate

$$\mathbf{D} = \frac{\mathbf{W}}{\mid \mathbf{V} \mid\mid \mathbf{U} \mid} \tag{1}$$

which turns out to be 0.000211655. The two joint marginal distributions are thus in good agreement with each other, as expected. Next we construct the quantity

$$\mathbf{S} = \frac{\mid (\mathbf{V}[i] - \mathbf{U}[i]) \mid}{\mathbf{V}[i]} \tag{2}$$

this time keeping only values of $\mathbf{V}[i]$ and $\mathbf{U}[i]$ that are larger than 0.05. We found that the largest value that this quantity takes is 0.08. As a further check we compare the genotype probabilities at locus 2 for all 555 individuals in the pedigree. Since we base our comparison on 1000 samples for any individual we ignore any genotypes which are sampled less than 50 times (*i.e.* corresponding to a probability of less than 5%) as well as genotypes which are completely fixed by marker information. The surviving genotypes in each of the joint marginal distributions for each individual are identical. Next we store the probabilities for the surviving genotypes in vectors $\mathbf{E}[i]$ from sampling eight loci and $\mathbf{T}[i]$ from sampling ten loci. These vectors are constructed so that $\mathbf{E}[j]$

and **T**[*j*] contain the probabilities for the same genotype for the same individual. Once again we calculate quantities analogous to **D** and **S** using the vectors **E** and **T** instead of **U** and **V**; **D** is 0.000551779 while **S** exceeds 0.07 in less than 3% of all cases. Even in cases where **S** exceeds 0.07 the marginal probabilities are consistent with each other assuming normally distributed sampling errors. This indicates that the genotypes for all individuals at locus 2 are indeed well sampled by considering just the first eight loci. Genotype probabilities closer to where we alter the pedigree are not as well reproduced; for example the 1/2 genotype for individual 55 at locus 7 is sampled 120 times when using the complete pedigree, and just 15 times when using just the first eight loci and the same sample sizes. Another instance arises in comparing the 1/2 genotype of individual 354, with the information on all ten marker loci it is sampled 147 times; keeping just eight loci it is sampled 555 times, clearly a glaring discrepancy. These discrepancies are hardly surprising given that locus 7 is just one unit of distance away from locus 8 where the pedigree has been modified in contrast to loci 1 and 2 which are much further away from the location where the pedigree has been truncated. At locus 8 (*i.e.* where we have modified the pedigree) the marginal probabilities for the genotypes of individuals 55 and 354 are very poorly reproduced when the pedigree is truncated, the effects of the propagation of the error induced by truncation to nearby loci is apparent.

In order to set up a desirable starting configuration for the Gibbs sampler we first truncate loci 9 and 10 and then draw a sample for the full distribution over the remaining eight loci. We store the sampled information over just the first two loci. This sample is to a good approximation a sample from the joint marginal distribution from the true distribution. Next we use our sample to condition the first two loci and then peel over the remaining eight loci (loci 3 through 10) and draw a sample from the remaining loci. Since our sample over the first two loci was to a good approximation a sample over the true marginal, our sample over all ten loci will also be to a good approximation a sample from the joint true distribution over all ten loci. This gives us a desirable starting point for the Gibbs sampler. What is striking though is the dramatic difference in memory requirements; in peeling all ten loci by brute force using a peeling order determined by a greedy heuristic the largest cutset [2] encountered has a size of just over 260 million, however the size of the largest cutset encountered when peeling the pedigree in two stages using the same greedy heuristic to determine the peeling order was just under two million. Thus conditioning on a few judiciously chosen loci at the start has had the effect of reducing memory requirements by more than two orders of magnitude. Given that the peeling order generated by the greedy heuristic is very likely not optimal in either case it would be unwise to draw any firm conclusions about reduction in memory usage in peeling, however the results presented are encouraging in this regard. In this example we broke up the problem in two steps, *i.e.* we were able to sample across all loci with two judiciously chosen peelings, one from loci 1 to 8 and the other from loci 3 to 10. If we had more than 10 loci, *e.g.* 12 loci we could consider using the second peeling to condition on loci 3 and 4. With loci 1 through 4 conditioned, we could then consider exactly peeling loci 5 to 12 and generating a desirable initial sample. In this manner datasets with rather more loci than can be peeled exactly could presumably be handled.

## 4. DISCUSSION

We demonstrate with one simple and one complex example, how certain marginal probability functions may be accurately estimated from truncated pedigrees. As long as the marginal probabilities involve variables which are many units of distance from where we modify the pedigree the resulting error may be expected to be small. Although we have illustrated the utility of this idea for tackling complications due to many loci, the idea can conceivably be used in other circumstances where exact peeling cannot be implemented. For example, we could consider a situation even with a single locus where there are too many loops to allow exact peeling. In this situation we could also truncate the pedigree keeping just a handful of individuals to begin with and then calculate the marginal distribution for individuals far away from where we have truncated the pedigree. In this case, the distances between nodes (which correspond to individuals in this case) would have to be computed using the breadth first algorithm mentioned earlier to locate the individuals for which marginal distributions could be reliably calculated despite the truncation. Having done so, we could then sample other individuals in the pedigree along the lines suggested above. In situations where there are not only many loops but also many loci, the same idea could in principle apply. In this case though, the subset $m$ may not be that straightforward to determine. Work along these lines is currently in progress.

Our strategy could also be used to implement the blocking Gibbs sampler [10] for sampling the joint genotype distribution of pedigrees with many loci. In the example we have just considered, we could, after obtaining a starting configuration over all ten loci, condition on loci 9 and 10 and then resample loci 1 through 8. With a new sample for loci 1 through 8, we could condition again on loci 1 and 2 and obtain a new sample for loci 3 through 10. This would in turn yield a new sample for loci 9 and 10 which could in turn yield another sample for loci 1 through 8. In this manner we could implement the blocking Gibbs sampler for the pedigree just considered with better mixing and none of the irreducibility problems of the scalar Gibbs sampler. Furthermore, in the scheme just described, we could sample all loci  with approximately the same frequency as required for a successful implementation of the Blocking Gibbs sampler [10].

The scheme we have just outlined presupposes that we can peel over all loci in just two stages, in many cases of interest this may not be the case. Let us assume for concreteness that we have 12 loci to peel over. We might consider the following adaptation of our basic strategy to set up a blocking Gibbs sampler: we partition the pedigree into overlapping blocks with loci 1 through 8 defining one block, loci 3 through 10 defining another block and loci 5 through 12 the third block. We generate an initial sample as follows:

(i) Peel loci 1 through 8 and sample loci 1 and 2. Save sample for loci 1 and 2.
(ii) Use sample from (i) to condition loci 1 and 2.
(iii) Peel loci 3 through 10.
(iv) Sample loci 3 and 4. Save sample on 3 and 4 and use to condition on loci 3 and 4. Loci 1 through 4 are now conditioned.
(v) With loci 1 through 4 conditioned, peel and sample loci 5 through 12. We now have sampled the entire pedigree. Save sample for loci 5 through 12.

Once we have this initial sample it is straightforward to sample and update blocks keeping all nodes outside the blocks conditioned. More precisely, loci 1 through 8 are sampled conditional on the previous sample for loci 9 through 12, loci 3 through 10 are sampled conditional on the current sample for loci 1 and 2 and the previous sample for loci 11 and 12, and loci 5 through 12 are sampled conditional on the current sample for loci 1 through 4. This procedure can be repeated as many times as desired, while potentially keeping mixing and irreducibility problems under control.

To conclude, we have described a method to generate a desirable starting configuration for the Gibbs sampler. Our method relies on finding a good approximation to the marginal distribution over a handfull of loci and then conditioning to permit peeling over the remaining loci. We demonstrate how this can be achieved in practice by breaking up a dataset involving ten loci in two, stages, one stage involving a peeling over loci 1 to 8 neglecting loci 9 and 10, followed by a conditioning on the first two loci followed in turn by a peeling from loci 3 through 10. The distance between two vertices in the graph representing the pedigree plays a crucial role in determining which marginals can be reliably computed for a given truncation of the pedigree. Our results indicate that this divide and conquer approach requires much less memory than would be needed to peel across all loci.

## REFERENCES

[1] Cannings C., Sheehan N.A., On a misconception about irreducibility of the single site Gibbs sampler in a pedigree application, Genetics 162 (2002) 993–996.

[2] Cannings C., Thompson E.A., Skolnick M.H., Probability functions on complex pedigrees, Adv. Appl. Prob. 10 (1978) 26–61.

[3] Cormen T.H., Leiserson C.E., Rivest R.L., Stein C., Introduction to Algorithms, 2nd. edition, The MIT Press, McGraww-Hill Book Company, 2001.

[4] Elston R.C., Stewart J., A general model for genetic analysis of pedigree data, Hum. Hered. 21 (1971) 523–542.

[5] Fishelson M., Dovgolevsky N., Geiger D., Maximum likelihood haplotyping for general pedigrees, Hum. Hered. 59 (2005) 41–60.

[6] Friedman N., Geiger D., Lotner N., Likelihood computations with value abstraction, in: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI), 2001.

[7] Heath S., Generating consistent genotypic configurations for multiallelic loci and large complex pedigrees, Hum. Hered. 48 (1998) 1–11.

[8] Janss L.L.G., Thompson R., van Arendonk J.A.M., Application of Gibbs sampling for inference in a mixed major gene polygenic inheritance model in animal populations, Theor. Appl. Gen. 91 (1995) 1137–1147.

[9] Janss L.L.G., van Arendonk J.A.M., van der Werf J.H.J., Computing approximate monogenic model likelihoods in large pedigrees with loops, Genet. Sel. Evol. 27 (1995) 567–579.

[10] Jensen C.S., Kong A., Blocking Gibbs sampling for linkage analysis in large pedigrees with many loops, Am. J. Hum. Genet. 65 (1999) 885–901.

[11] Lange K., Boehnke L., Extensions to pedigree analysis. V. Optimal calculations of Mendelian likelihoods, Hum. Hered. 33 (1983) 291–301.

[12] Lange K., Elston R.C., Extensions to pedigree analysis. I. Likelihood calculations for simple and complex pedigrees, Hum. Hered. 25 (1975) 95–105.

[13] Lange K., Goradia T.M., An algorithm for automatic genotype elimination, Am. J. Hum. Genet. 40 (1987) 250–256.

[14] Malek M., Dekkers J.C.M., Lee H.K., Baas T.J., Rothschild M.F., A molecular genome scan to identify chromosomal regions influencing economic traits in the pig. I. Growth and body composition, Mamm. Genome 12 (2001) 630–636.

[15] Malek M., Dekkers J.C.M., Lee H.K., Baas T.J., Rothschild M.F., A molecular genome scan to identify chromosomal regions influencing economic traits in the pig. II. Meat and muscle composition. Mamm. Genome 12 (2001) 637–645.

[16] Stricker C., Fernando R.L., Elston R.C., An algorithm to approximate the likelihood for pedigree data with loops by cutting, Theor. Appl. Gen. 91 (1995) 1054–1063.

[17] Thomas A., Approximate computations of probability functions for pedigree analysis, IMJ J. Math. Appl. Med. Biol. 3 (1986a) 157–166.

[18] Thomas A., Optimal computations of probability functions for pedigree analysis, IMJ J. Math Appl. Med. Biol. 3 (1986b) 167–178.

To access this journal online:
www.edpsciences.org