Computation of identity by descent probabilities conditional on DNA markers via a Monte Carlo Markov Chain method

Miguel Pérez-Enciso, Luis Varona, Max Rothschild

Genet Sel Evol 32 (2000) 467–482

Original article

Computation of identity by descent probabilities conditional on DNA markers *via* a Monte Carlo Markov Chain method

Miguel PÉREZ-ENCISO[a,b,c,*] Luis VARONA[a],
Max F. ROTHSCHILD[b]

[a] Area de Producció Animal, Centre UdL-IRTA, 25198 Lleida, Spain
[b] Department of Animal Science, Iowa State University, Ames, 50011-3150 Iowa, USA
[c] Present address Station d'amélioration génétique des animaux, Institut national de la recherche agronomique, 31326 Castanet Tolosan Cedex, France

(Received 11 January 2000; accepted 05 June 2000)

Abstract – The accurate estimation of the probability of identity by descent (IBD) at loci or genome positions of interest is paramount to the genetic study of quantitative and disease resistance traits We present a Monte Carlo Markov Chain method to compute IBD probabilities between individuals conditional on DNA markers and on pedigree information. The IBDs can be obtained in a completely general pedigree at any genome position of interest, and all marker and pedigree information available is used The method can be split into two steps at each iteration First, phases are sampled using current genotypic configurations of relatives and second, crossover events are simulated conditional on phases Internal track is kept of all founder origins and crossovers such that the IBD probabilities averaged over replicates are rapidly obtained We illustrate the method with some examples First, we show that all pedigree information should be used to obtain line origin probabilities in F2 crosses Second, the distribution of genetic relationships between half and full sibs is analysed in both simulated data and in real data from an F2 cross in pigs

DNA markers / identity by descent probability / Monte Carlo Markov Chain

Résumé – Calcul de probabilités d'identité par descendance sachant les marqueurs moléculaires d'ADN *via* une méthode MCMC. L'estimation précise des probabilités d'identité par descendance (IBD) est fondamentale pour l'analyse génétique de caractères quantitatifs et de susceptibilités aux maladies On présente une méthode de Chaîne de Markov Monte Carlo (MCMC) pour obtenir les probabilités IBD entre individus sachant les marqueurs d'ADN et le pedigree Les IBDs peuvent être calculés pour un pedigree général sur une position quelconque du génome Toute l'information des marqueurs et du pedigree est utilisée La méthode consiste en deux

* Correspondence and reprints
E-mail: mperez@toulouse inra fr

étapes · 1) les phases sont échantillonnées en utilisant les configurations génotypiques des individus apparentés ; 2) les crossing-overs sont simulés sachant les phases. On trace l'origine des fondateurs et des positions des crossing-overs Donc les probabilités IBD sont calculées immédiatement On présente quelques applications D'abord, on montre que toute l'information doit être prise en compte pour calculer les probabilités sur croisements F2 Ensuite, on étudie la distribution de coefficients de parenté entre frères avec données réelles et simulées

chaîne de Markov Monte Carlo / marqueurs moléculaires / probabilité d'identité par descendance

1. INTRODUCTION

The accurate estimation of the probability of alleles being identical by descent (IBD) at loci or genome positions of interest is paramount to the genetic study of quantitative and disease traits. Two alleles are said to be identical by descent if they are replicates of the same allele that were inherited through a common ancestor. DNA markers make it possible to estimate the probability of IBD between relatives at any position of the genome. Nonetheless, there are no general and easy-to-use methods aimed at this purpose that can be applied to any experimental design.

The analytical methods available have the main advantage that they are fast to compute but they do not necessarily use all pedigree and genotypic information, and are usually limited to a number of simple relationships like full or half sibs or assume that phases are known. Haseman and Elston [9] first studied the estimation of IBD probabilities between full sibs in their seminal paper. Chevalet *et al.* [1] obtained the conditional probabilities of loci linked to a marker, but their approach is in practice restricted to rather small pedigrees because the numerical complexity increases exponentially with pedigree size. Fernando and Grossman's [3] method is limited to one marker and one QTL, and the generalization of Goddard [5] assumes that phases are known. This latter author did not consider double recombinants between markers. Haley *et al* [8] only used information from parents and grandparents, discarding full and half sib genotypes to estimate breed origin in F2 populations. Their method is not aimed at obtaining the IBD probabilities between any pair of relatives and is restricted to F2 designs. Kruglyak *et al.* [14] implemented a general and exact approach for calculating the inheritance distribution pattern, but the method is limited to small (< 12 founder individuals) pedigrees given the combinatorial problems handled.

Alternatively, Monte Carlo Markov Chain (MCMC) methods rely on drawing successive random samples of the variables of interest. MCMC methods have received much attention in genetics, especially in humans [15,19,20] but they have been more concerned with the issue of reconstructing genotypes when these are missing than with estimating genetic relationships. Grignola *et al.* [6] presents a method for the obtention of genetic relationships using marker information that sample phases using an MCMC strategy. Then, the authors apply the strategy of Goddard [5].

We present a stochastic method to compute genetic relationships conditional on marker and on pedigree information. Thus the aim of the work presented

here is distinct from that of using DNA markers to infer genetic relationships in the absence of pedigree information, *e.g.* [16]. The method developed here can be applied to a completely general pedigree and uses all marker and pedigree information available. We follow the approach of Guo [7], who considered the genome as a continuum, rather than a collection of ordered loci. This author, however, only provided analytical results and for the simplest relationships.

2. MATERIALS AND METHODS

2.1. Method

We assume that recombination fractions (δ) between markers have been previously estimated and that genotypes are known for all individuals of interest, although a limited generalisation to missing genotypes is also described. Thus, the random variables are the phase (*i.e.*, whether a given allele is of paternal or maternal origin) and the recombination events that may have occurred to produce the observed genotypes. The method can be split into three steps, where steps 2 and 3 are repeated for a predetermined number of iterations:

2.1.1. Initialisation

A compatible phase is randomly sampled by giving equal probability to both phases, whenever the phase is not completely determined by the parents' genotypes.

2.1.2. Phase sampling

If appropriate, the phase for each marker and individual is sampled conditional on the current genotype and phase of the parents, spouse and offspring of the individual. Each of the two phase probabilities is assigned 0.5 *a priori* The first marker where the sire of the individual is heterozygous is identified to the left and to the right of the current marker. The alleles are checked with the individual's genotype to modify the phase probabilities. The same strategy is applied using the current phases for the dam, and the offspring. Finally a phase is sampled at random using the joint probability A detailed description of how probabilities are calculated is given in the Appendix. The phases are first updated for all animals within each marker, then the next marker is chosen. Nonetheless, it is also possible to first update all markers for a single individual.

2.1.3. Crossover sampling

Recombination events are sampled conditional on current phases This is done for each haplotype in turn, considering the sire's origin haplotype and the dam's origin haplotype separately because they are independent conditional on phases. The markers with heterozygous genotypes in the sire (dam) are identified. The number of recombination events is generated as follows. First, note that any number of recombination events may have occurred between the telomere and the first informative marker. Second, an even number of crossovers (including zero) must have occurred between consecutive informative markers if both alleles in the offspring come from the same haplotype in the

parent, and an odd number of crossovers must have occurred if each allele in the offspring comes from different haplotypes, *i.e.* at least one crossover must have occurred during meiosis For the telomeric regions, the number of crossovers is sampled from a Poisson distribution according to the distance between the telomere and the first informative marker. For the within marker crossovers, the number of recombinants follows a censored Poisson distribution, according to the above rule, *i.e.*, the probabilities of odd (even) number of crossovers are set to zero and those of the even (odd) events are rescaled to add up to one. If no marker is informative, the number of crossovers can be either odd or even, and it is sampled simply according to the total chromosome length. The location of the crossovers is assigned at random within the appropriate marker interval. Technically, step two is a Gibbs sampling scheme, whereas step three is a composition sampling [18] A composition sampling scheme means that first a "nuisance" variable is sampled (the crossover locations), and a second variable (the relationship coefficients) are sampled conditional on the first variable. Here note that r is unambiguously determined given the crossover locations.

Figure 1 illustrates the method in a pedigree consisting of sire, dam, and two offspring O1 and O2 Thus, there were four founder haplotypes SS, SD, DS, and DD Three markers were located on positions 20, 40 and 60 cM. Only markers 1 and 3 were relevant for generating crossovers in the sire's origin haplotype, whereas only marker 2 was relevant on the dam's side. The OS1 haplotype was 113, thus there must be an even number of crossovers located between positions 20 and 60 cM (a zero recombinant is pictured in Fig 1); OS2 haplotype was 213, *i.e.*, an even number of crossovers must have occurred in interval 20–60 cM (a crossover in position 25 is shown). OD1 and OD2 crossovers were generated in a similar fashion. Note that in the next MCMC iteration the phase sampled for the sire could be, *e.g.*, $\frac{1\,1\,1}{2\,1\,3}$ and then OS1 would be a recombinant haplotype and OS2, a non-recombinant haplotype

By repeating the process just described all haplotypes in the population are simulated. Internal track is kept of all crossovers within each iteration such that it is possible to trace the founder origin of any descendant at any point of the genome with complete certainty This allows us to obtain any required relationship in a straightforward manner The method assumes that all founder alleles are distinct by descent, although they can, of course, be identical by state. (The program can be modified easily to force that some founder alleles are IBD). The appropriate statistics (*e.g.*, additive or dominance relationship coefficient, inbreeding coefficients) are computed at every iteration Suppose that we are interested in the additive genetic relationship (r) at position x. That position is inspected at each iteration for all pairs of individuals and both haplotypes If all pairs of haplotypes have the same founder origin, the four alleles are IBD and $r = 2$ (equivalent to complete inbreeding), if only one allele of each individual is identical $r = 0.5$. Computing relationships along a given segment between two positions are obtained in a similar way; if crossovers have occurred within the interval of interest, r is weighted appropriately according to the percentage of distance shared. Consider again Figure 1 OS1 and OS2 are IBD between positions 25 and 60 cM, whereas OD1 and OD2 are IBD between 0 and 45 cM. The additive relationship between O1 and O2 is

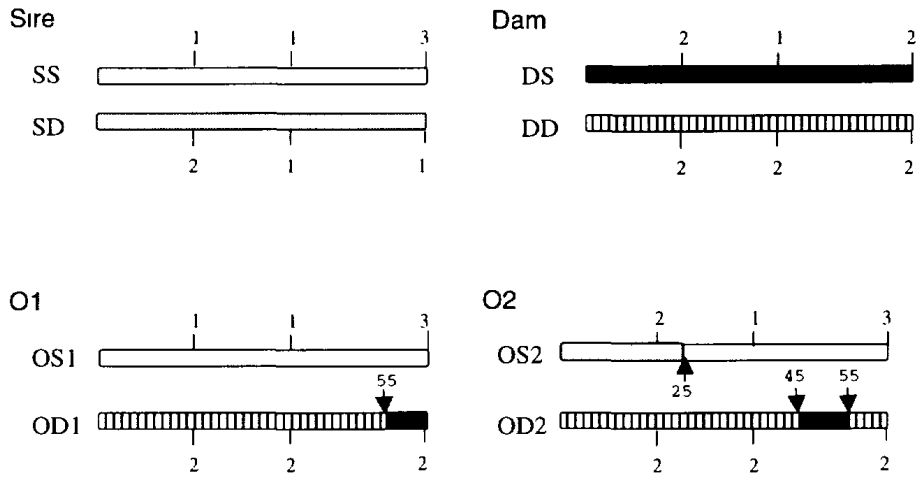

Figure 1. Diagram showing the simulation of crossover events Sire with genotype and current phase $\frac{1\,1\,3}{2\,1\,1}$ was mated to dam with haplotypes $\frac{2\,1\,2}{2\,2\,2}$ produces offspring O1, $\frac{1\,1\,3}{2\,2\,2}$, and O2, $\frac{2\,1\,3}{2\,2\,2}$ All four haplotype origins, SS, SD, DS, and DD, are shown in different patterns The three markers were in positions 20, 40, and 60 cM. Crossovers were generated conditional on genotypes and current phases In the scheme shown here (which would represent one MCMC iterate), O1's sire haplotype (OS1) was non-recombinant, dam's origin chromosome (OD1) resulted from a crossover in position 55, O2 resulted from one crossover on male meiosis at position 25 (OS2), and from two crossovers in positions 45 and 55 for the dam's origin (OD2) The small arrows and numbers in OS2, OD1 and OD2 indicate the crossover positions in cM

thus $r = \frac{1}{2}\left(\frac{60-25}{60} + \frac{45}{60}\right) = \frac{2}{3}$ at the iteration pictured. The dominance coefficient is the probability that both haplotypes have IBD alleles (*e.g.*, [10]), which in Figure 1 occurs between positions 25 and 45 cM, thus $\frac{45-25}{60} = \frac{1}{3}$ is the current dominance relationship. The average of genetic relationships over iterations can be obtained or the distribution can be plotted if the value for each iteration has been saved.

2.1.4. Irreducibility

Note that the Gibbs sampling scheme just described is not reducible when there are no missing genotypes, and thus ends up converging. A Gibbs sampling scheme is said to be irreducible if there is a non-nil (albeit small) probability of changing from any two states [18]. To prove this, first note that the genotypes for which the phase needs to be sampled can be identified by inspecting the genotypes of their parents, and that a phase is sampled in either all or none of the iterations (logically, all founder individual phases are to be sampled). Second, note that despite the fact that phases of the parents and offspring provide information about the most likely phase of an individual, the probability of both phases always remains non-zero, except when phases are determined

from the parents' genotypes or in the trivial case of completely linked markers $\delta = 0$ (but then again phases are not sampled).

2.1.5. Missing genotypes

Two legal descent states may not communicate in the presence of three or more alleles with missing genotypes [15] Unfortunately, at present there is not a general algorithm that guarantees convergence of an MCMC chain in any complex pedigree if some genotypes are missing, despite the significant and recent efforts in strategies like block Gibbs sampling (Jensen and Kong, [11]) We have developed a strategy that can deal with missing genotypes in a limited number of situations. Genotypes cannot be missing in the founder animals and it is not recommended that there are missing genotypes in consecutive generations, *i e.*, if an individual is not genotyped, the algorithm works if its parents and offspring are genotyped or if it is an individual without offspring Running several MCMC chains is highly recommended if this is not the case in order to avoid absorbing states as much as possible During MCMC sampling, the genotypes (and not only phases as when there are no missing genotypes) are sampled conditional on the current phases of the parents, offspring and spouse and the linked markers of the individual. It is ensured that genotypes sampled at every iteration are compatible.

2.1.6. Convergence

An important issue in all MCMC methods is whether convergence has been attained. Many different criteria have been proposed [2], although none of them guarantees that the chain has actually converged. In this sense, it is useful to compute the autocorrelation coefficient (the correlation between every n samples from the chain) because it provides a measure of how many "useful" (independent) samples have been obtained Note that we should expect little autocorrelation with this method. If the markers are informative, phases are not sampled because they are known (except in founders), and the recombination events sampled are completely independent between iterations. If markers are not very informative and a percentage of phases are sampled, a small autocorrelation between successive phases can be expected if the individuals with unknown phases are directly related, but recombination sampling is still independent between samples.

2.2. Applications

In the following we illustrate the method described in three diverse genetic situations Five thousand iterations were run in all cases described.

2.2.1. Line origin probabilities in an F2 cross

The strategy of Haley *et al.* [8] disregards the genotypic information from the F2 full and half sibs, as well as genotypes from the parental and F1 relatives We illustrate the advantages of using all available genotypic information in the pedigree of Figure 2.

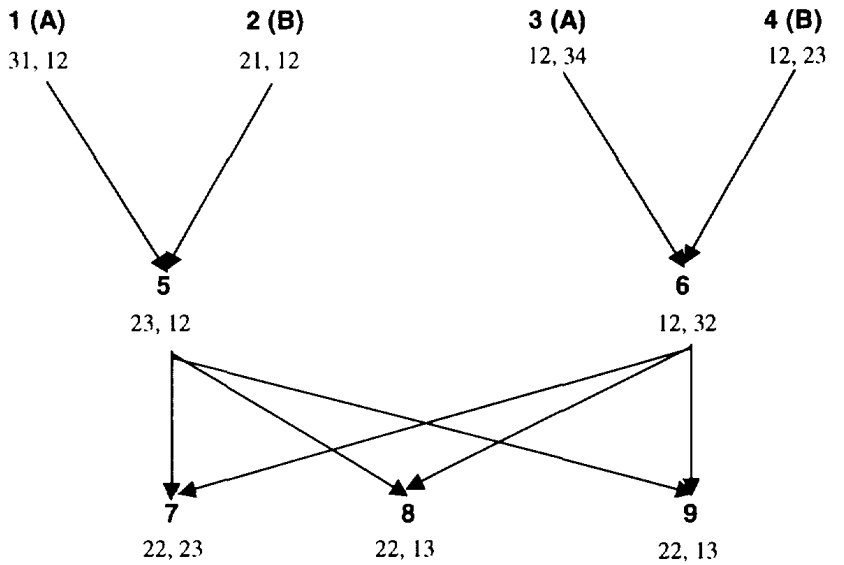

Figure 2. Example pedigree of an F2 cross The individual (ID) number is shown in bold 1 and 3 are parental individuals from breed A; 2 and 4, from breed B; 5 and 7 are the F1 parents; 7, 8, and 9 are F2 individuals. The genotypes are shown below the IDs, genotypes for each marker are separated by commas The first marker is located in position 0 cM, and the second marker, in position 20 cM

2.2.2. Distribution of additive relationship coefficients in half and full sibs

We simulated 100 families of 10 half sibs (100 sires mated to 10 dams each), and 100 families of 10 full sibs in order to study the distribution of r between half and full sibs. Two extreme instances were considered: no marker information available (*i.e.*, simply the same allele was assigned to all markers and individuals) and maximum marker informativity (each founder had two different alleles that were also different from the other founder alleles for all markers). A single chromosome of 100 cM was simulated, there were six markers evenly spaced every 20 cM beginning at 0 cM

2.2.3. F2 reference family in pigs

We computed r for the porcine PIT1 region on chromosome 13 in the F2 reference families described in Yu *et al.* [22]. The genotyped loci were Swr1008, SOO68, Swr398, and Sw1056 plus the PIT1 locus itself. The number of alleles was 4 (PIT1), 7 (Swr398 and Sw1056) and 11 (Swr1008 and SOO68). The region spanned 49 cM [23]. The five largest full sib families from F1 plus the five largest non-inbred families from F2 were selected and the r between all full sibs of the whole segment was computed The families were chosen to be non-inbred in order to ensure that the expected relationship coefficient was 0 5. In total, there were 45 and 476 full sib pairs in the F1 and F2 generations, respectively. For the sake of comparison, 5 full sib families of size 10 were also simulated with a marker spacing identical to the real data.

Table I. Probabilities of line origins in the interval 19–20 cM for the F2 individuals ID in pedigree from Figure 2 calculated under different situations. The results shown in the last two rows were obtained when the genotype for the second marker in individual 7 (alleles 2,3) was missing

Full sibs considered	Missing genotype	ID	p(AA)	p(BB)	p(AB)	p(BA)
No	No	7	0.18	0 01	0 01	0 80
		8	0 18	0 01	0.01	0.80
		9	0 18	0 01	0 01	0 80
Yes	No	7	0 80	0 01	0.01	0 18
		8	0.17	0 00	0 00	0 83
		9	0.17	0 00	0.00	0 83
No	Yes	7	0.08	0 42	0.08	0.42
Yes	Yes	7	0 13	0 15	0 03	0 68

3. RESULTS

3.1. Line origin probabilities in an F2 cross

The main results are given in Table I. The first three rows show the line origin probabilities in the neighborhood of the second marker position (19–20 cM) when the F2 individuals were analysed separately as in Haley *et al.* ([8]; the program of Haley *et al.* provided the same results except for small differences due to sampling). Note that individuals 7, 8 and 9 would have equal regression coefficients using this approach. However, if all individuals are analysed jointly (rows 4–6 in Tab. I) the coefficients for ID 7 change dramatically. The reason is that the most likely phase of ID 5 is different according to whether the genotype from ID 7 or 8 is observed. If the genotype from individual 7 is observed, the most likely (non-recombinant) phase for individual 5 is $\frac{2\,2}{3\,1}$, whereas it is $\frac{2\,1}{3\,2}$ if genotype 8 (or 9) is observed. Now, if both 7 and 8 genotypes are observed, either the sire's gamete that resulted in ID 7 or 8 must be recombinant. Given that 8 and 9 have the same genotype, the most likely phase for 5 is $\frac{2\,1}{3\,2}$, and 7 must result from a recombinant gamete. In consequence, the most likely line origin for allele "2" in the second marker of ID 7 is breed A, rather than B. The effect of including all available information when the last marker of ID 7 is missing (2 –, 2 –) is shown in the last two rows of Table I. Again, it is clear that considering genotypes from IDs 8 and 9 does have a large influence on the predicted line origin probabilities for ID 7.

3.2. Distribution of additive relationship coefficients in half and full sibs

The empirical mean and standard deviations of r between half and full sibs without marker information are shown in the upper two rows of Table II.

Table II. Mean and standard deviations (s d) of the additive relationship coefficient between full and half sibs for several haplotype configurations. Each digit represents an allele for each of the six markers, a dot is used to signify that the genotypes are not known (upper two rows) or that they are not relevant (bottom four rows), the sire's haplotype is $\frac{1\,1\,1\,1\,1\,1}{2\,2\,2\,2\,2\,2}$ The results correspond to a single chromosome 100 cM long

Sib-type	Sibs' haplotypes	Mean	s.d	E(s.d.)*
Half	$\frac{\cdot\quad\quad}{\cdot\quad\quad}$, $\frac{\cdot\ \cdot\quad\quad}{\cdot\ \cdot\ \cdot\ \cdot}$	0.250	0 149	0 153
Full	$\frac{\quad\quad\quad}{\quad\quad\quad}$, $\frac{\cdot\quad\quad\cdot}{\cdot\ \cdot}$	0.500	0 215	0 212
Half	$\frac{1\,1\,1\,1\,1\,1}{\cdot\ \cdot}$, $\frac{1\,1\,1\,1\,1\,1}{\cdot}$	0 490	0.027	–
Half	$\frac{1\,1\,1\,2\,2\,2}{\cdot\ \cdot\ \cdot\ \cdot}$, $\frac{1\,1\,1\,2\,2\,2}{\cdot\quad\cdot}$	0.461	0 028	–
Half	$\frac{1\,1\,1\,1\,1\,1}{\cdot\ \cdot\ \cdot}$, $\frac{1\,1\,1\,2\,2\,2}{\cdot\ \cdot\ \cdot\ \cdot}$	0 250	0.030	–

* Predicted standard deviation of r using Hill's formula (1993, [11])

There is an excellent agreement with the theoretical values predicted using Hill's [11] formula. It is also most interesting to compare the plot of the distributions of r for both half and full sibs (Figs. 3a, 3b). The distribution for half sibs shows two distinct peaks at the extreme values of the distribution $r = 0$ and $r = 0.5$. In this case the mean (0.25) does not provide a useful description of the parameter at all The reason for this *a priori* surprising result is that the probability of a non-recombinant gamete is relatively high for a 100 cM chromosome ($P = 0.38$), thus two half sibs share with relatively high probability either all or none of the sire's haplotype, and these are the two modal values. In the case of full sibs (Fig. 3b), the modal value is around the mean because even if the percentage of no recombinant gametes is the same, there are two possibilities of being identical by descent, either by the sire or by the dam. The small bumps in the plots are caused by crossovers, and they smooth out as the number of MCMC iterates tend to infinity.

The last three rows of Table II contain the r's obtained for some particular genotypic configurations. The numbers in the haplotypes show which alleles are shared by both sibs In rows three and four, both sibs share all the alleles. In row three, the sibs share a non-recombinant haplotype, whereas in row four, both sibs share a recombinant haplotype. In the latter case at least one recombinant must have occurred between positions 40 and 60, so that the distribution of r is not so peaked as when sibs share non-recombinant haplotypes (Figs. 4a, 4b). The case shown in the last row of Table II corresponds to sibs sharing only half of the sire's alleles (the expected value). In this case the distribution of r is completely symmetric around 0.25 with a sharp

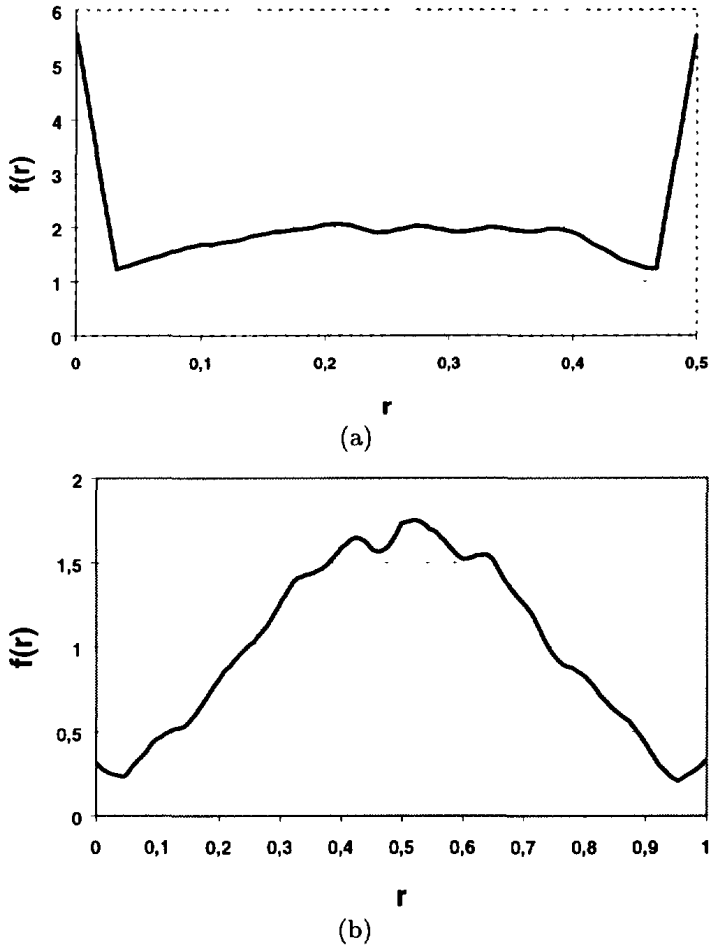

Figure 3. Distribution of the relationship coefficients between half sibs (a) and between full sibs (b) for a 1 M chromosome in the absence of molecular markers

drop off below $r = 0.2$ and above $r = 0.3$. This peculiar pattern is caused because the haplotype 111222 (Tab II) results from an obligatory crossover in a 20 cM space between positions 40 and 60, and from the fact that the location of the crossover is random within these positions. To understand this, consider the chromosome divided in three segments 0–40, 40–60, and 60–100 cM. Almost all half sib pairs will be IBD for the first segment, and non IBD for the last segment. Thus, most relationship coefficients will range between $r = \frac{1}{2}(40/100) = 0.2$ as a minimum and $r = \frac{1}{2}(40 + 20) = 0.3$ Take λ to be the location of the crossover event in cM, which is distributed uniformly between 40 and 60. Thus the relationship coefficient between sibs is $r = 0\,2 + \frac{1}{2}(\lambda/100)$, and is also uniform between 0 2 and 0.3, as observed in Figure 4c

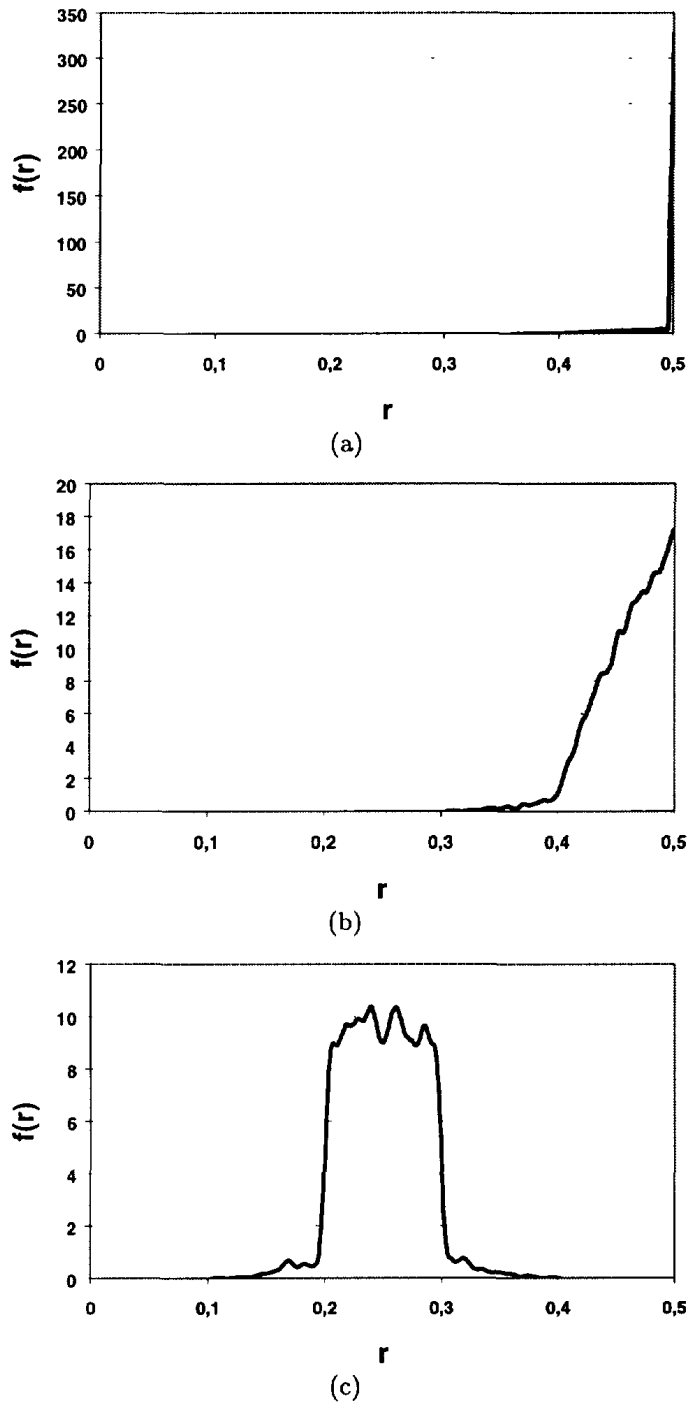

Figure 4. Distribution of the relationship coefficients between half sibs with molecular markers· (a) sibs share a complete non-recombinant haplotype; (b) sibs share a complete recombinant haplotype; (c) sibs share half of a recombinant haplotype See Table II for details

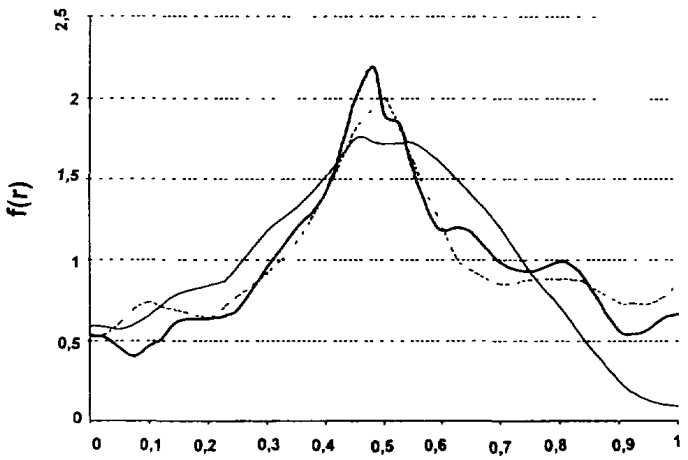

Figure 5. Plot of average relationship coefficients between full sibs around the PIT1 region in a pig F2 reference family [22], F1 (thin line), F2 (thick line), and simulated (dashed line)

3.3. Pig reference families. Convergence issues

Figure 5 is a plot of the distribution of average additive relationship coefficients between all full sibs in the pig reference family. The dashed line corresponds to results obtained by simulation. The good agreement with the data from the F2 generation rather than with the F1 is apparent. But there were only 45 full sib pairs in the F1 *vs.* 476 in the F2, and it may be reasonable that discrepancies may be due to the small number of data in the F1. Strong discrepancies between simulated and observed values would be a symptom of segregation bias, or strong crossover interference.

We studied the autocorrelation coefficient between MCMC samples in a number of situations and we found, as expected, very little autocorrelation between successive samples. For instance, the autocorrelation coefficients between consecutive samples for r between full sibs 149 and 150 was 0.06, and 0.01 between every 100 samples. These autocorrelations were negligible, as usual Gibbs sampling schemes may generate samples with an autocorrelation coefficient of 0.6 even for lag-300 samples [21]. The successive sample values for $r_{149,150}$ are plotted in Figure 6. A simple visual inspection of the plots confirms that there was no apparent trend in the sampling, and basically the same average value for $r_{149,150}$ was obtained with the first half as with the last half of the samples.

4. DISCUSSION

The production of a massive number of genotypes per individual at a reasonable cost will be a feasible task in the near future by using DNA chip or microarray technology. Statistical methods to analyse this information in a flexible and general way are thus much needed. MCMC methods like the one described here offer a variety of advantages over analytical approaches.

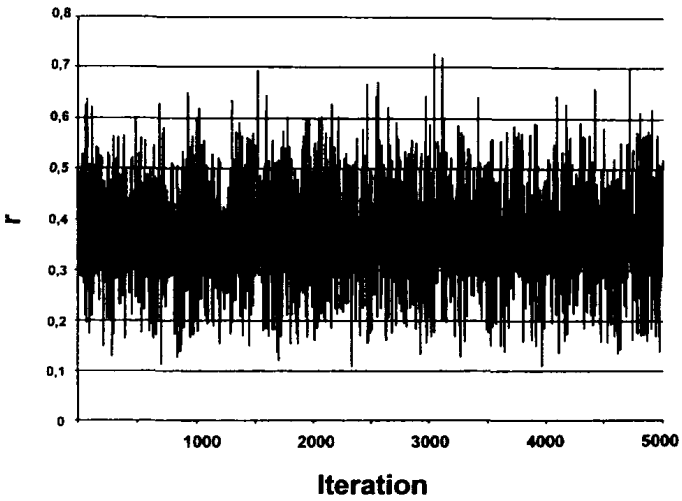

Figure 6. Sample values obtained for the relationship coefficient between full sibs 149 and 150 The average value was 0.369

First, they are more general and flexible than analytical approaches, and are not restricted to a particular population structure. Thus they make use of all available information automatically. The user needs not worry about which genotypic information is used in computing origin probabilities, since the whole genotyped pedigree is taken into account. We are currently working in extending the method to allow for a more general treatment of missing genotypes. Interestingly, the usefulness of considering all pedigree information increases with lowly polymorphic but tightly linked markers, a characteristic of rapidly developing markers like single nucleotide polymorphisms. It must be noticed that the approach depicted here can be applied to any experimental design (F2, backcross, recombinant inbred lines, F2 pseudocross) or natural population as long as genotypes and pedigree information are available.

Second, the approach followed here, whereby crossovers are simulated conditional on current phase is particularly suited to explore any kind of genetic relationship. Dominance relationship coefficients can be computed without any extra theoretical development, as well as the identity coefficients developed by Gillois [4] or the generalized kinship coefficients of Karigl [13]. We have already developed a modification of the algorithm to obtain the Gillois' coefficients (Pérez-Enciso and Fernando, unpublished results). The availability of these coefficients makes the study of more complex genetic effects than the merely additive effects possible. In a similar fashion, it is straightforward to obtain the joint genetic relationship for two distant loci or segments, facilitating the study of epistatic interactions. In this work we have used Haldane's function (one crossover expected every Morgan) but the strategy presented lends itself to model other recombination functions as well. This would permit to locate QTL positions using more realistic recombination functions. MCMC also allows us to infer the phase of the founder individuals. This is particularly useful to detect allele heterogeneity in the parental populations and to trace back the contribution of each founder genome to any descendant.

Third, MCMC methods provide the marginal distribution of the parameter of interest, which can be drawn or wherefrom any statistics (*e.g.*, standard deviation, skewness) can be obtained. No resort to infinite sample theory is required. The distribution of genetic relationship parameters is relevant for discriminating between alternative relationships when the pedigree is not known [20], and also to provide a clear picture of how much deviation from the mean IBD sharing we can expect. We have shown that r can follow extremely complex distribution patterns, and for most situations the normal distribution is inappropriate or even misleading. A somewhat surprising example occurs with half sibs, where a clear bimodal distribution is obtained (Fig. 3a). (Recall nonetheless that the shape will depend on the length of the linkage group.) We have concentrated here on obtaining r for a single chromosome. For the whole genome a number of independent events is added up, and normality is reached rapidly in the absence of markers. Yet, a smooth function cannot be expected with markers (Fig. 4). In practice, one would be interested in obtaining r for the regions where QTLs are located, *i.e.*, only a small fraction of the genome.

Finally, note that each sampling stage depends exclusively on the parents' genotypic configurations Thus, the computer run time and memory requirement of our approach are proportional to N, the number of individuals in the pedigree, provided that only the individual genotypic probabilities are required. And memory and CPU requirements increase approximately $O(N^2)$ in complex pedigrees if the relationship between all individuals is required. Memory requirements also increase linearly with the number of generations and genome length, as the number of crossovers increase.

We have applied the method described here to a variety of QTL detection approaches. In Pérez-Enciso and Varona [17], we computed the relationship matrix conditional on marker information in an F2 pedigree: a different relationship matrix for IBD probabilities of each parental line was obtained. We showed by simulation that the approach is very robust in analysing data of crosses between outbred lines. We are currently analyzing real data sets from an F2 cross in pigs and a half sib design in dairy cattle using the IBD probabilities obtained as described here.

Software availability

FORTRAN77 software is available from the senior author on request. (Note: The software will also be available *via* an anonymous ftp or web page).

ACKNOWLEDGEMENTS

We are grateful to Rohan L. Fernando, Miguel A. Toro, Eduardo Manfredi and Jean Michel Elsen for useful discussions and comments. MPE expresses his appreciation for the financial support received by Cotswold USA during his stay at Iowa State University. The work was supported by CICYT AGF96–2510 (Spain) and BIO4-CT97–962243 (E.U.) grants, by Cotswold USA and the Iowa Agricultural and Home Economics Experiment Station and State of Iowa Hatch Funds. This is paper no. J - 18649.

REFERENCES

[1] Chevalet C , Gillois M., Vu Tien Kang J , Conditional probabilities of identity of genes at a locus linked to a marker, Genet. Sel. Evol 16 (1984) 431–444.

[2] Cowles M K , Carlin B.P , Markov Chain Monte Carlo convergence diagnostics a comparative review, J Am. Stat Assoc. 91 (1996) 883–904

[3] Fernando R L , Grossman M , Marker-assisted selection using best linear unbiased prediction, Genet Sel Evol. 21 (1989) 467–477

[4] Gillois M , La relation d'identité en génétique, Ann Inst. Henri Poincaré B2 (1964) 1–94

[5] Goddard M , A mixed model for analyses of data on multiple genetic markers, Theor Appl Genet 83 (1992) 878–886

[6] Grignola F E , Hoeschele I , Tier B , Mapping quantitative trait loci *via* residual maximum likelihood I. Methodology, Genet. Sel. Evol 28 (1996) 479–490

[7] Guo S , Proportion of genome shared identical by descent by relatives· concept, computation, and applications, Am. J Hum Genet 56 (1995) 1468–1476.

[8] Haley C S , Knott S A., Elsen J M., Mapping quantitative trait loci in crosses between outbred lines using least squares, Genetics 136 (1994) 1195–1207

[9] Haseman J K , Elston R.C , The investigation of linkage between a quantitative trait and a marker locus, Behav Genet 2 (1972) 3–19.

[10] Henderson C R., Applications of linear models in animal breeding, University of Guelph Press, Guelph, 1984

[11] Hill W G , Variation in genetic identity within kinships, Heredity 71 (1993) 652–653

[12] 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

[13] Karigl G , A recursive algorithm for the calculation of identity coefficients, Ann Hum Genet 45 (1981) 299–305.

[14] Kruglyak L , Daly M.J., Reeve-Daly M.P., Lander E.S., Parametric and Non-parametric linkage analysis: A unified multipoint approach, Am. J Hum Genet 58 (1996)1347–1363

[15] Lange K , Mathematical and Statistical Methods for Genetic Analysis, Springer, New York, 1997

[16] Lynch M , Ritland K , Estimation of pairwise relatedness with molecular markers, Genetics 152 (1999) 1753–1766.

[17] Pérez-Enciso M , Varona L., Quantitative trait loci mapping in F2 crosses between outbred lines, Genetics 155 (2000) 391–405.

[18] Tanner M , Tools for Statistical Inference, Springer, New York, 1993.

[19] Thompson E A , Monte Carlo likelihood in genetic mapping, Stat. Sci 9 (1994) 355–366

[20] Thompson E A , Inferring gene ancestry: estimating gene descent, Int Stat. Rev 66 (1998) 29–40.

[21] Wang C S , Rutledge J J., Gianola D., Bayesian analysis of mixed linear models *via* Gibbs sampling with an application to litter size in Iberian pigs, Genet Sel Evol 26 (1994) 91–115

[22] Yu T -P , Tuggle C K., Schmitz C B., Rothschild M.F , Association of PIT1 polymorphisms with growth and carcass traits in pigs, J. Anim Sci 73 (1995) 1282–1288

[23] Yu T -P , Wang L , Tuggle C K , Rothschild M F , Mapping genes for fatness and growth on pig chromosome 13: a search in the region close to the PIT1 gene, J Anim Breed Genet 116 (1999) 269–280

APPENDIX

Computation of phase probabilities

Denote by $M_{i,j,1}$ and $M_{i,j,2}$ the unordered alleles 1 and 2 from individual i at marker j, and by $G_{i,j,1}$ and $G_{i,j,2}$, the paternal and maternal alleles, respectively, *i.e.*, G is the ordered M. The sire of individual i is S, D is the dam, and i has n offspring. The conditional probability of $M_{i,j,1}$ being of paternal origin ($M_{i,j,1} = G_{i,j,1}$) is:

$$p_{i,j} = p_{1,i,j}/(p_{1,i,j} + p_{2,i,j}),$$

where

$$p_{1,i,j} = q^{\mathrm{L}}_{1,\mathrm{S},i,j}\, q^{\mathrm{R}}_{1,\mathrm{S},i,j}\, q^{\mathrm{L}}_{1,\mathrm{D},i,j}\, q^{\mathrm{R}}_{1,\mathrm{D},i,j} \prod_{k=1}^{n} q^{\mathrm{L}}_{1,k,i,j}\, q^{\mathrm{R}}_{1,k,i,j},$$

$$\text{and } p_{2,i,j} = q^{\mathrm{L}}_{2,\mathrm{S},i,j}\, q^{\mathrm{R}}_{2,\mathrm{S},i,j}\, q^{\mathrm{L}}_{2,\mathrm{D},i,j}\, q^{\mathrm{R}}_{2,\mathrm{D},i,j} \prod_{k=1}^{n} q^{\mathrm{L}}_{2,k,i,j}\, q^{\mathrm{R}}_{2,k,i,j}.$$

The coefficients q_1 and q_2 express the probabilities of $M_{i,j,1}$ and $M_{i,j,2}$ being of paternal or maternal origin given the current phases in the remaining markers and individuals. Take the closest marker to j, located to the "left" (*i.e.*, $j' < j$) where the sire is heterozygous, thus informative. The recombination fraction between markers j and j' is $\delta_{j,j'}$. There are four mutually exclusive cases, and the corresponding $q^{\mathrm{L}}_{1,\mathrm{S},i,j}$ and $q^{\mathrm{L}}_{2,\mathrm{S},i,j}$ are:

Case	$q^{\mathrm{L}}_{1,\mathrm{S},i,j}$	$q^{\mathrm{L}}_{2,\mathrm{S},i,j}$
$\{G_{\mathrm{S},j',1}, G_{\mathrm{S},j,1}\} = \{G_{i,j',1}, M_{i,j,1}\}$	$(1-\delta_{j,j'})$	$\delta_{j,j'}$
$\{G_{\mathrm{S},j',1}, G_{\mathrm{S},j,1}\} = \{G_{i,j',1}, M_{i,j,2}\}$	$\delta_{j,j'}$	$(1-\delta_{j,j'})$
$\{G_{\mathrm{S},j',2}, G_{\mathrm{S},j,2}\} = \{G_{i,j',1}, M_{i,j,1}\}$	$\delta_{j,j'}$	$(1-\delta_{j,j'})$
$\{G_{\mathrm{S},j',2}, G_{\mathrm{S},j,2}\} = \{G_{i,j',1}, M_{i,j,2}\}$	$(1-\delta_{j,j'})$	$\delta_{j,j'}$

The same procedure is followed for the next marker to the "right" ($j' > j$) to obtain $q^{\mathrm{R}}_{1,\mathrm{S},i,j}$ and $q^{\mathrm{R}}_{2,\mathrm{S},i,j}$, and for the dam and offspring contribution. The q_1 and q_2 are both set to 0.5 if the corresponding information is not available, or if no marker is informative, *e.g.*, $q_{1,\mathrm{S}} = q_{2,\mathrm{S}} = 0.5$ if the sire is unknown.