Theor Appl Genet (1993) 87:89–93

An efficient algorithm to compute the posterior genotypic distribution for every member of a pedigree without loops

R. L. Fernando [1], C. Stricker [2], R. C. Elston [2]

[1] Department of Animal Sciences, University of Illinois at Urbana-Champaign, 1207 West Gregory Drive, Urbana, IL 61801, USA
[2] Department of Biometry and Genetics, Louisiana State University, Medical Center, 1901 Perdido Street, New Orleans 70112-1393, USA

Received: 21 October 1992 / Accepted: 12 March 1993

Abstract. This paper describes a non-iterative, recursive method to compute the likelihood for a pedigree without loops, and hence an efficient way to compute genotype probabilities for every member of the pedigree. The method can be used with multiple mates and large sibships. Scaling is used in calculations to avoid numerical problems in working with large pedigrees.

Key words: Genotype probabilities – Likelihood – Recursive algorithm

Introduction

Most traits of agricultural importance are assumed to be controlled by genes at a large number of loci. These are called polygenic traits, and their analysis is usually based on the assumption of a normal distribution for the genotypic value. This assumption leads to methods of analysis for these traits that have appealing statistical properties and are also computationally efficient, even with large complex pedigrees (Henderson 1973; Henderson 1984; Wiggans et al. 1988).

Traits determined by genes at a single locus or a few loci are called monogenic or oligogenic traits, respectively. Methods for calculating the likelihood of pedigrees for monogenic and oligogenic traits have been extensively discussed in the human genetics literature (Elston and Stewart 1971; Lange and Elston 1975; Cannings et al. 1976; Cannings et al. 1978; Lange and Boehnke 1983) and have only recently been discussed in the animal genetics literature (Elston 1990). Methods have also been developed for calculating genotype probabilities for a future member of human pedigrees (Murphy and Mutalik 1969; Heuch and Li 1972; Lalouel 1980).

Van Arendonk et al. (1989) presented an iterative algorithm to calculate the genotype probabilities of all members in an animal pedigree. Some limitations in their algorithm have been removed by Janss et al. (1992). The objective of this paper is to present a more efficient non-iterative, recursive algorithm to calculate for an oligogenic trait the genotype probabilities of all members in an animal or human pedigree without loops. The principles used in this development were introduced by Murphy and Mutalik (1969), Elston and Stewart (1971), and Heuch and Li (1972). The iterative algorithms by Van Arendonk et al. (1989) and by Janss et al. (1992) are based on these same principles.

Definition and notation

A pedigree can be represented diagramatically (Fig. 1) with lines connecting mates with each other and offspring with parents. Each member of the pedigree has either two parents or no parents in the pedigree; those with no parents are called founders (members 1, 2, 4, 5, 6, 7, 8, 11, 14, and 23 in Fig. 1). Relative to any member i in the pedigree, the remaining members can be divided into two groups – (1) those anterior to it and, (2) those posterior to it. The members anterior to any member i are defined to be those connected to i through its parents and fullsibs (including the parents and fullsibs themselves). For example, the members anterior to 12 are: 1, 2, 3, 4, 5, 6, 9, 10, 15, and 16 (Fig. 1). The members posterior to a pedigree member i are defined to be those connected to i through its mates and offspring (including the mates and offspring themselves). For example, the members posterior to 12

Communicated by E. J. Eisen
Correspondence to: R. L. Fernando

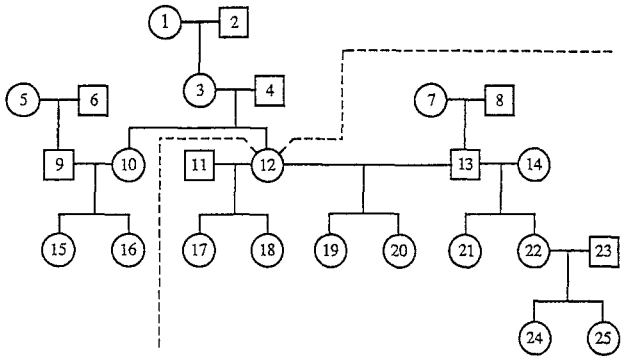

Fig. 1. Diagramatic representation of pedigree with lines connecting mates with each other and offspring with parents. All members of the pedigree *above and to the left of the dotted line* are anterior to 12, all those *below and to the right of the dotted line* are posterior to 12

are: 7, 8, 11, 13, 14, 17, 18, 19, 20, 21, 22, 23, 24, and 25 (Fig. 1). If the anterior members for *any* member i include any of its mates, the pedigree is said to have a loop.

Let y_i be the phenotypic value of i and $g(y_i|u_i)$ the conditional probability of y_i, or probability density, if the phenotype is continuous (this distinction will be omitted in the rest of the paper), given i has genotype u_i. When y_i is missing, $g(y_i|u_i)$ is set to one. The anterior probability for a member i, denoted $a_i(u_i)$, is defined to be the joint probability of pedigree members anterior to i having the observed phenotypes and i having genotype u_i. The posterior probability through mate j, denoted $p_{ij}(u_i)$, is defined to be the conditional probability of phenotypes of pedigree members posterior to i through its mate j and through its offspring from mate j, given i has genotype u_i.

Let S_i be the set of pedigree members that are mates of i, and let C_{ij} be the set of children of parents i and j.

Recursive calculations

Genotype probabilities and likelihood

The conditional probability that pedigree member i has genotype u_i given all the phenotypic data ($\boldsymbol{y}$) can be calculated as

$$\Pr(u_i|\boldsymbol{y}) = a_i(u_i)\, g(y_i|u_i) \prod_{j\in S_i} p_{ij}(u_i)/L \tag{1}$$

where

$$L = \sum_{u_i} a_i(u_i)\, g(y_i|u_i) \prod_{j\in S_i} p_{ij}(u_i). \tag{2}$$

The first factor in the numerator of (1), $a_i(u_i)$, is the joint probability of phenotypes of members anterior to i and of genoype u_i for i. The second factor, $g(y_i|u_i)$, is the conditional probability that i has phenotype y_i given it has genotype u_i. The third factor, $\prod p_{ij}(u_i)$, is the conditional probability of phenotypes of members posterior to i, given i has genotype u_i. The product of these probabilities gives the joint probability of the phenotypes of all members of the pedigree and of genotype u_i for i. Summing this joint probability over all genotypes u_i for i as in (2) gives the probability for the observed pedigree data. Expressed as a function of unknown parameters, L is the likelihood for the pedigree.

To calculate the genotype probabilities using (1), the anterior probabilities $a_i(u_i)$ and posterior probabilities $p_{ij}(u_i)$ for j in S_i are needed. The calculation of each of these probabilities is described below. For simplicity of exposition we shall assume that mating is at random, but only minor modifications are necessary to allow for a joint distribution of mating types.

Anterior probabilities

If i is a founder, it has no anterior members. Thus, the anterior probabilities for i are given by the population genotype probabilities appropriate for i. If i is not inbred, these probabilities will be in Hardy-Weinberg equilibrium proportions; otherwise these probabilities will reflect the degree of inbreeding. If i is not a founder, it is connected to its anterior members only through its parents and full sibs. Thus, to recursively calculate the anterior probabilities for i, first calculate the following:

1) anterior probabilities $a_m(u_m)$ and $a_f(u_f)$ for m and f, the parents of i;
2) posterior probabilities $p_{mj}(u_m)$ for m through all mates j of m, except $j=f$;
3) posterior probabilities $p_{fj}(u_f)$ for f through all mates j of f, except $j=m$;
4) posterior probabilities $p_{jk}(u_j)$ for all offspring j of m and f, except $j=i$, through mates k of j.

For a pedigree without loops, the calculation of these four sets of probabilities will not require the calculation of $a_i(u_i)$, the anterior probabilities for i.

Once these are calculated, $a_i(u_i)$ can be calculated as

$$\begin{aligned} a_i(u_i) = \sum_{u_m} \Bigg\{ & a_m(u_m)\, g(y_m|u_m) \prod_{\substack{j\in S_m \\ j\neq f}} p_{mj}(u_m) \\ & \times \sum_{u_f} \Bigg\{ a_f(u_f)\, g(y_f|u_f) \prod_{\substack{j\in S_f \\ j\neq m}} p_{fj}(u_f) \\ & \times \mathrm{tr}(u_i|u_m, u_f) \\ & \times \prod_{\substack{j\in C_{mf} \\ j\neq i}} \left[\sum_{u_j} \mathrm{tr}(u_j|u_m, u_f)\, g(y_j|u_j) \prod_{k\in S_j} p_{jk}(u_j) \right] \Bigg\} \Bigg\}, \end{aligned} \tag{3}$$

where $\mathrm{tr}(u_i|u_m, u_f)$ is the conditional probability that i has genotype u_i given that i's parents, m and f, have genotypes u_m and u_f. The product of the factors $a_m(u_m)$, $g(y_m|u_m)$, and $\prod p_{mj}(u_m)$ on the first line of (3) gives the joint probability of phenotypes of members anterior to i through parent m and of genotype u_m for m.

Similarly, the product of the factors $a_f(u_f)$, $g(y_f|u_f)$, and $\prod p_{fj}(u_f)$ on the second line gives the joint probability of phenotypes of members anterior to i through parent f and of genotype u_f for f. The last line of (3) gives the joint probability of phenotypes of members anterior to i through its fullsibs, given genotypes u_m and u_f for the parents. The product of the factors on lines 1, 2, and 4, described above, and $\mathrm{tr}(u_i|u_m, u_f)$, defined earlier, gives the joint probability of phenotypes of members anterior to i through its parents and fullsibs and of genotypes for the parents and i. Summing over the genotypes of the parents gives the anterior probability for i.

Posterior probabilities

A pedigree member i is connected to its posterior members only through its mates and offspring. Thus, to recursively calculate the posterior probabilities $p_{ij}(u_i)$ for i through its mate j, first calculate

1) anterior probabilities $a_j(u_j)$ for the mate j of i;
2) posterior probabilities $p_{jk}(u_j)$ for j through all the mates k of j, except $k=i$;
3) posterior probabilities $p_{kl}(u_k)$ for all offspring k of i and j, through the mates l of k.

For a pedigree without loops, calculation of these probabilities does not require the calculation of $p_{ij}(u_i)$, the posterior probability for i through j.

Once these are available, $p_{ij}(u_i)$ can be calculated as

$$p_{ij}(u_i)=\sum_{u_j}\left\{a_j(u_j)\,g(y_j|u_j)\prod_{\substack{k\in S_j\\k\neq i}}p_{jk}(u_j)\times\prod_{k\in C_{ij}}\left[\sum_{u_k}\mathrm{tr}(u_k|u_i,u_j)\,g(y_k|u_k)\prod_{l\in S_k}p_{kl}(u_k)\right]\right\} \tag{4}$$

The product of the factors, $a_j(u_j)$, $g(y_j|u_j)$, and $\prod p_{jk}(u_j)$ on the first line of (4) gives the joint probability of the phenotypes of pedigree members posterior to i through j and of genotype u_j for j. The second line gives the joint probability of the phenotypes of the pedigree members posterior to i through its offspring from mate j, given the genotypes u_i and u_j for i and j, respectively. The product of the factors on line 1 and line 2 gives the joint probability of the genotype for j and of phenotypes of members posterior to i through mate j and through its offspring from mate j, given the genotype for i. Summing over the genotypes of j gives the posterior probability $p_{ij}(u_i)$.

Example

The pedigree given in Fig. 1 is used here to illustrate the recursive calculations. Consider the calculation of genotype probabilities for individual 12.

Table 1. Probabilities required for the recursive calculation of $\Pr(u_{12}|y)$. Column $i+1$ gives the probabilities required for calculation of those in column i. Anterior probabilities for founders are given by the population genotype frequencies, and no further recursion is needed

$\Pr(u_{12}\|y)$	$a_{12}(u_{12})$	$a_4(u_4)$		
		$a_3(u_3)$	$a_1(u_1)$	
			$a_2(u_2)$	
		$p_{10,9}(u_{10})$	$a_9(u_9)$	$a_5(u_5)$
				$a_6(u_6)$
	$p_{12,11}(u_{12})$	$a_{11}(u_{11})$		
	$p_{12,13}(u_{12})$	$a_{13}(u_{13})$	$a_7(u_7)$	
			$a_8(u_8)$	
		$p_{13,14}(u_{13})$	$a_{14}(u_{14})$	
			$p_{22,23}(u_{22})$	$a_{23}(u_{23})$

From (1), calculation of $\Pr(u_{12}|y)$ requires $a_{12}(u_{12})$, $p_{12,11}(u_{12})$, and $p_{12,13}(u_{12})$ (see Table 1). Calculation of $a_{12}(u_{12})$ in turn requires $a_3(u_3)$, $a_4(u_4)$, and $p_{10,9}(u_{10})$. Calculation of $a_3(u_3)$ requires $a_1(u_1)$ and $a_2(u_2)$. Individuals 1 and 2 are founders, and thus their anteriors are given by population genotype frequencies. Now the calculations for $a_3(u_3)$ can be completed. Individual 4 is also a founder so no further recursion is needed to obtain its anterior. Now the calculation of $p_{10,9}(u_{10})$ is undertaken. This requires $a_9(u_9)$. Calculation of $a_9(u_9)$ in turn requires $a_5(u_5)$ and $a_6(u_6)$. Individuals 5 and 6 are founders, and no further recursion is needed to get their anteriors. The calculation of $a_9(u_9)$ can now be completed, and this is used to complete the calculation of $p_{10,9}(u_{10})$. Now that $a_3(u_3)$, $a_4(u_4)$, and $p_{10,9}(u_{10})$ are available, the calculation of $a_{12}(u_{12})$ can be completed. Next, the calculation of $p_{12,11}(u_{12})$ is undertaken. This requires $a_{11}(u_{11})$. Individual 11 is a founder, and no further recursion is needed. The last item required to complete the calculation of $\Pr(u_{12}|y)$ is $p_{12,13}(u_{12})$. Calculation of $p_{12,13}(u_{12})$ requires $a_{13}(u_{13})$ and $p_{13,14}(u_{13})$. Calculation of $a_{13}(u_{13})$ requires $a_7(u_7)$ and $a_8(u_8)$. Individuals 7 and 8 are founders, and no further recursion is needed to get their anteriors. The anteriors for 7 and 8 are now used to complete the calculation of $a_{13}(u_{13})$. Next, the calculation of $p_{13,14}(u_{13})$ is undertaken. This requires $a_{14}(u_{14})$ and $p_{22,23}(u_{22})$. Individual 14 is a founder, and no further recursion is needed to obtain its anterior. Next, calculation of $p_{22,23}(u_{22})$ requires $a_{23}(u_{23})$. Individual 23 is a founder, and its anterior is obtained without further recursion. This is used to complete the calculation of $p_{22,23}(u_{22})$, and this in turn is used together with $a_{14}(u_{14})$ to complete the calculation of $p_{13,14}(u_{13})$. Next, $a_{13}(u_{13})$ and $p_{13,14}(u_{13})$ are used to complete the calculation of $p_{12,13}(u_{12})$. Finally, this is used together with $a_{12}(u_{12})$ and $p_{12,11}(u_{12})$ to calculate $\Pr(u_{12}|y)$.

In the calculation of genotype probabilities for another individual, only the required anteriors and posteriors that have not been already obtained need to be calculated. For example, to calculate $\Pr(u_{13}|y)$, $a_{13}(u_{13})$, $p_{13,12}(u_{13})$, and $p_{13,14}(u_{13})$ are needed. However, $a_{13}(u_{13})$ and $p_{13,14}(u_{13})$ have already been obtained in the calculation of $\Pr(u_{12}|y)$. Thus, only $p_{13,12}(u_{13})$ needs to be calculated. Calculation of $p_{13,12}(u_{13})$ requires $a_{12}(u_{12})$ and $p_{12,11}(u_{12})$. Both of these have already been obtained. Thus, the calcuation of $p_{13,12}(u_{13})$ can be completed, and this in turn can be used together with $a_{13}(u_{13})$ and $p_{13,14}(u_{13})$ to calculate $\Pr(u_{13}|y)$.

Scaling

Calculation of the genotype probabilities and likelihood by direct application of Eqs. (1) and (2) together with (3) and (4) can result in the manipulation of very small or very large numbers. This can lead to numerical problems on computers where numbers are represented with finite precision. This problem can be avoided by scaling $a_i(u_i)$, $p_{ij}(u_i)$, and $g(y_i|u_i)$ such that they always sum to one over the genotypes u_i. Scaling of $g(y_i|u_i)$, for example, is accomplished by dividing $g(y_i|u_i)$ by the scaling factor $\sum_{u_i} g(y_i|u_i)$.

When the scaled $a_i(u_i)$, $p_{ij}(u_i)$, and $g(y_i|u_i)$ are used in (1) and (2), the scaling factors in the numerator and denominator of (1) cancel to give the unscaled value for $\Pr(u_i|y)$. Thus, we do not calculate the unscaled value of the likelihood, which may be outside the range of numbers that can be represented on computers. The unscaled log likelihood can be obtained by accumulating the logs of the scaling factors. Let K be the sum of the log scaling factors used in calculating the likelihood. Then, the unscaled value of the log likelihood is given by $K+\log(L)$, where L is the scaled value of the likelihood.

Time and storage requirements

Formulae (3) an (4) can be used to calculate all the anterior and posterior probabilities by iteration. We first describe how this can be done, and then compare the time required for recursive calculation with that for iterative calculation.

To start the iterative process, set the anterior probabilities of all founders to the population genotype frequencies and set all other anterior and posterior probabilities to one. An iteration consists of the following calculations for every individual in the pedigree:

1) If the individual is not a founder, calculate its anterior using formula (3) with the current values for the required anteriors and posteriors.
2) If the individual is a parent, calculate its posterior through each mate using formula (4) with the current values for the required anteriors and posteriors.

This process is repeated, calculating anterior and posterior probabilities in reverse sequence to that in which they were calculated in the previous iteration, until convergence is reached.

The time required to calculate all the anterior and posterior probabilities by recursion is equal to that for one iteration of the above process. Once all the anteriors and posteriors are available, both approaches will require the same amount of time to calculate the genotype probabilities using (1) and (2). Both the recursive and iterative methods require the storage of anterior probabilities for every parent that is not a founder and of posterior probabilities for every parent through each of its mates. Thus, our non-iterative algorithm requires less time and no more storage than does the iterative algorithm.

Discussion

The algorithm presented here for pedigrees without loops is relatively easy to program because of its recursive nature, is efficient in time requirements because it is non-iterative, and requires no more storage space than a comparable iterative algorithm. Although it can be adapted to be used for pedigrees with loops, the results obtained would not be exact. A non-iterative algorithm for pedigrees with loops could be based on the method of likelihood calculation suggested for this situation by Lange and Elston (1975) or Goradia et al. (1992). Such an algorithm would be computationally feasible for pedigrees with a few small loops, such as are generated, for example, by sire-daughter matings, but would quickly become impractical (in terms of both time and storage requirements) as the loops increase in number and size. For such situations, approximate calculations using recursive or iterative procedures may be the best solution. The larger the loops, the slower one might expect the speed of convergence of iterative procedures to be. On the other hand, the effect of cutting a loop, to eliminate it, is smaller the larger the size of the loop, so that slow speed of convergence may not matter in these cases. The best way to calculate genotype probabilities for a pedigree with loops is a topic deserving further study.

Acknowledgements. The authors are grateful to Alexa Sorant for useful discussions on likelihood calculations. This work was supported in part by U.S. Public Health Service research grant GM 28356 from the National Institute of Medical Sciences, resource grant RR 03655 from the Division of Research Resources, and training grant HL 07567 from the National Heart, Lung, and Blood Institute. C. Stricker was supported by the Schweizerischer Nationalfonds, Switzerland.

References

Cannings C, Thompon EA, Skolnick MH (1976) The recursive derivation of likelihoods on complex pedigrees. Adv Appl Prob 8:622–625

Cannings C, Thompson EA, Skolnick MH (1978) Probability functions on complex pedigrees. Adv Appl Prob 10:26–61

Elston RC (1990) Models for discrimination between alternative modes of inheritance. In: Gianola D, Hammond K, (eds) Advances in statistical methods for genetic improvement of livestock. Springer, Berlin Heidelberg New York, pp 41–55

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21:523–542

Goradia TM, Lange K, Miller PL, Nadkarni PM (1992) Fast computation of genetic likelihoods on human pedigree data. Hum Hered 42:42–62

Henderson CR (1973) Sire evaluation and genetic trends. In: Anim Breed Genet Symp in Honor of Dr JL Lush. Am Soc Anim Sci Am Dairy Sci Assoc, Champaign, Ill., pp 10–41

Henderson CR (1984) Applications of linear models in animal breeding. University of Guelph, Guelph, Ontario

Heuch I, Li FHF, (1972) Pedig – a computer program for calculation of genotype probabilities using phenotype information. Clin Genet 3:501–504

Janss LLG, Van der Werf JHJ, Van Arendonk JAM (1992) Detection of a major gene using segregation analysis in data from several generations. In: Proc Eur Assoc Anim Prod

Lalouel JM (1980) Probability calculations in pedigrees under complex modes of inheritance. Hum Hered 30:320–323

Lange K, Boehnke M (1983) Extensions to pedigree analysis. V. Optimal calculation of Mendelian likelihoods. Hum. Hered 33:291–301

Lange K, Elston RC (1975) Extension to pedigree analysis. I. Likelihood calculations for simple and complex pedigrees. Hum Hered 25:95–105

Murphy EA, Mutalik GS (1969) The application of Bayesian methods in genetic counselling. Hum Hered 19:126–151

van Arendonk JAM, Smith C, Kennedy BW (1989) Method to estimate genotype probabilities at individual loci in farm livestock. Theor Appl Genet 78:735–740

Wiggans GR, Misztal I, Van Vleck LD (1988) Implementation of an animal model for genetic evaluation of dairy cattle in United States. In: Schmidt GH (ed) Proc Anim Model Workshop. Am Dairy Sci Assoc, Champaign, Ill., pp 54–69