Theor Appl Genet (1989) 78: 735–740

Method to estimate genotype probabilities at individual loci in farm livestock

J. A. M. van Arendonk [1]*, C. Smith [2] and B. W. Kennedy [2]

[1] Department of Animal Breeding, Wageningen Agricultural University, P.O. Box 338, NL-6700 AH Wageningen, The Netherlands
[2] Centre for Genetic Improvement of Livestock, Department of Animal and Poultry Science, University of Guelph, Guelph, Ontario N1G 2W1, Canada

Received September 2, 1988; Accepted June 27, 1989
Communicated by E. J. Eisen

Summary. A Bayesian method to estimate genotype probabilities at a single locus using information on the individual and all its relatives and their mates has been developed. The method uses data over several generations, can deal with large numbers of individuals in large livestock families and allows for missing information. It can be extended to multiple alleles and can be used for autosomal or sex-linked loci. The allele frequencies and the form of expression (dominance, penetrance) must be specified. An algorithm using the method and involving an iterative procedure has been developed to calculate the genotype probabilities for practical use in livestock breeding. The method and algorithm were used to determine the accuracy of estimating genotype probabilities of sires for a female sex-limited trait, such as genetic variants of milk proteins. Data were similated and genotype probabilities estimated for 100 sires (20 replicates) with 3, 6 and 12 female offspring per sire, for different population frequencies, for additive and dominance gene action and for variable genotypic expression. Such simulation is useful in the design of testing systems for the use of information on specific genetic loci in selection.

Key words: Individual loci – Estimation – Genetic counseling – Genotype probabilities – Accuracy

Introduction

Animal geneticists have relied largely on quantitative variation in the genetic improvement of livestock, and have developed sophisticated statistical methods to exploit all the information in selection (Henderson 1984; Kennedy et al. 1988). In the future, with extensive genetic polymorphism being found by molecular genetic techniques, they will need to develop new methods to deal with information on individual loci (and linked marker loci) and to integrate these into their evaluation systems. Much use can be made of methods developed in human genetics for analyzing modes of inheritance and for determining genotype probabilities for genetic counseling purposes. These methods were first outlined by Elston and Stewart (1971) and are reviewed by Elston and Rao (1978) and recently by Elston (1987). Several computer packages are available for dealing with human pedigrees. In livestock, pedigrees are usually much larger due to generally higher reproductive rates, especially in males. For example, a dairy sire used in artificial insemination may have thousands of daughters in a large number of herds. For this reason, the application of the detailed methods and programs in human genetics to livestock data will often be difficult or inappropriate.

Some work in animal genetics has dealt with methods to determine the mode of inheritance of putative Mendelian traits. For example, Smith and Bampton (1977), Carden et al. (1983) and Southwood et al. (1988) used various likelihood procedures to study the inheritance of reaction to halothane anaesthesia in pigs. However, these applications to livestock data did not allow for missing phenotypes and used only part of the data on related animals.

The objective of this paper is to derive a method and a computer algorithm to estimate genotype probabilities at a single locus for a known mode of inheritance. The method can deal with extensive data over several generations, and simultaneously derives the genotype probabilities for all individuals. Its effectiveness is evaluated for a female sex-limited trait under different population allele

* Prepared during a leave at Centre for Genetic Improvement of Livestock, Guelph, Canada

frequencies and forms of expression with differing numbers of female offspring per sire.

Estimation of genotype probabilities

A method to estimate genotype probabilities at a single locus with two alleles, A and a, will be derived. It can be extended simply to multiple alleles. The basic ideas were taken from Elston and Stewart (1971) and Heuch and Li (1972). The mode of inheritance is assumed known. The method allows for individuals with missing phenotypes, can be applied to sex-limited traits and can be used for both autosomal and sex-linked loci.

Genotype probabilities of an individual are estimated using all the information in its family. The information for an individual can be divided into three categories: (a) parents and collateral relatives; (b) the individual; (c) offspring and their relatives (Fig. 1). The genotype probabilities are calculated using a Bayesian probability approach. The prior probability of an individual having a given genotype is calculated from information on parents and collateral relatives (a) using Mendelian segregation probabilities. The posterior probability is calculated by incorporating the information on the individual (b) and on offspring and their relatives (c). This probability will be referred to as (estimated) genotype probability. For computational reasons, information (c) is condensed into the so-called progeny probabilities. The different probabilities and their method of calculation are given below.

Information on the individual

When only the phenotype f on the individual is known, the probability of that individual having genotype u is

$$\text{prob}(u) = \text{prior}(u)\, g(f|u) \Big/ \left\{ \sum_{v=1}^{k} \text{prior}(v)\, g(f|v) \right\} \qquad (1)$$

where prob (u) = probability of the individual having genotype u, prior (u) = prior probability of the individual having genotype u, $g(f|u)$ = conditional probability of phenotype f given genotype u, k = number of possible genotypes (three for a two-allele locus). In the absence of information on parents, prior (u) equals the proportion $\tau(u)$ of individuals in the population that have genotype u. When the population is assumed to be in Hardy-Weinberg equilibrium, the proportions $\tau(u)$ equal p^2, $2pq$ and q^2, where p and q are the frequencies of the A and a alleles, respectively. Deviations from Hardy-Weinberg equilibrium can be taken into account.

The probability $g(f|u)$ describes the relation between phenotypes and genotypes. It thus takes account of dominance, penetrance and the probabilities of misclassification. Table 1 gives the probabilities $g(f|u)$ for two situations: I – no dominance and complete penetrance; and II – complete dominance of A allele and 80% penetrance of the recessive aa genotype. Equation (1) will be illustrated for both situations. The estimated genotype probabilities for an individual with phenotype $f = AA$ in situation I are obviously 1, 0 and 0 for genotype AA, Aa and aa, respectively. For situation II, Hardy-Weinberg equilibrium and a frequency of the A allele of 0.6 are used. In that case, the estimated probabilities for an animal with the dominant phenotype $f = A$- is $0.41 = 1\,(0.6)^2/[1\,(0.6)^2 + 1\,(2)\,(0.6)\,(0.4) + 0.2\,(0.4)^2]$ for genotype AA, and similarly 0.55 and 0.04 for genotypes Aa and aa.

There is no major difficulty in letting $g(f|u)$ be also a function of some other factor, e.g. age or sex (Elston and Stewart 1971). In case of a missing observation of the phenotype, $g(f|u)$ is set equal to unity for all u. Note that Eq. (1) reduces to $\tau(u)$ if the phenotype is missing. Both $\tau(u)$ and $g(f|u)$ are assumed known.

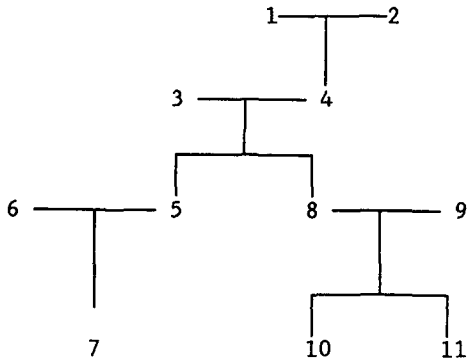

Fig. 1. Information for individual 8 in different categories: (a) parents and collateral relatives (1–7); (b) individual (8); (c) offspring and their relatives (9–11). *Numbers* are used to identify animals

Table 1. Conditional probability given genotype u of observing phenotype f, $g(f|u)$, for I: no dominance and complete penetrance; II: complete dominance and 80% penetrance of recessive genotype

	I			II	
	$f = AA$	Aa	aa	$f = A$-	aa
$u = AA$	1	0	0	1	0
Aa	0	1	0	1	0
aa	0	0	1	0.2	0.8

Table 2. The conditional probability of the individual having genotype u_i given its parents genotype u_s and u_d, pt $(u_i|u_s, u_d)$

	$u_s = AA$			$u_s = Aa$			$u_s = aa$		
$u_i =$	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa
$u_d = AA$	1	0	0	0.5	0.5	0	0	1	0
Aa	0.5	0.5	0	0.25	0.5	0.25	0	0.5	0.5
aa	0	1	0	0	0.5	0.5	0	0	1

Information on parents

When, in addition to the phenotype of the individual, the genotypes of the parents (sire and dam) are known to be u_s and u_d, Eq. (1) can be modified to

$$\text{prob}(u) = \text{pt}(u|u_s, u_d)\, g(f|u) \Big/ \left\{ \sum_{v=1}^{k} \text{pt}(v|u_s, u_d)\, g(f|v) \right\} \qquad (2)$$

where pt $(u|u_s, u_d)$ is the conditional probability of the individual having genotype u given its parent's genotypes are u_s and u_d. These probabilities are based on Mendelian transmission, which is different for an autosomal and a sex-linked locus. Table 2 gives pt $(u|u_s, u_d)$ for an autosomal locus with two alleles.

Equation (2) can be used only when the genotypes of the parents are known. The equation can be generalized to allow for the use of the parents' genotype probabilities. To distinguish between the terms relating to the different individuals, a subscript i will be added to the terms referring to individual i for which genotype probabilities are estimated.

The information on the parents of individual i can be combined in the prior probability of i having genotype u_i, $\text{prior}_i(u_i)$, as follows

$$\text{prior}_i(u_i) = \sum_{u_s=1}^{k} \text{prob}_s(u_s, i) \sum_{u_d=1}^{k} \text{prob}_d(u_d, i)\, \text{pt}(u_i|u_s, u_d)/tot \quad (3)$$

where $\text{prob}_s(u_s, i)$ is the probability of the sire having genotype u_s when the information from individual i or its descendants is not used (see below). Similarly, $\text{prob}_d(u_d, i)$ gives the probabilities for the dam. To make the prior probabilities sum to 1, each probability is divided by *tot* which is the sum of all the possible numerator terms (three possible genotypes) for individual i. In the case of a unknown parent, $\text{prob}_s(u_s, i)$ or $\text{prob}_d(u_d, i)$ is set equal to $\tau(u)$. In a case where both parents are missing, it can be shown that Eq. (3) reduces to $\tau(u)$.

Murphy and Mutalik (1969) showed that the genotype probabilities of the parents used in Eq. (3) should not include the information on the individual for which genotype probabilities are estimated or on descendants of that individual. It may be useful to illustrate this. For situation II with a frequency of the A allele of 0.5, consider a sire X with unknown parents and phenotype and its offspring Y with phenotype A-. The prior probabilities for Y are $\tau(u)$ (0.25, 0.50, 0.25) because no information is available on either parent. As will be shown in Eqs. (4) and (5), the estimated genotype probabilities for X using information on Y are 0.40, 0.34 and 0.26 for u being AA, Aa and aa, respectively. Using these probabilities in Eq. (3) will lead to the erroneous prior probabilities for Y of 0.285, 0.5 and 0.215, respectively.

Progeny

The information from the progeny of individual i can be combined in the progeny probability, $\text{prog}_i(u_i)$, which can be calculated as

$$\text{prog}_i(u_i) = \prod_{p=1}^{n_i} \left\{ \sum_{u_m=1}^{k} \text{prob}_m(u_m, p) \cdot \sum_{u_p=1}^{k} g(f_p|u_p)\, \text{prog}_p(u_p)\, \text{po}(u_i|u_m, u_p) \right\} \Big/ tot \quad (4)$$

where n_i = number of offspring of individual i, $\text{prob}_m(u_m, p)$ = probability of mate producing offspring p having genotype u_m when information on p and its descendants is not used but all other information on mate is used, $g(f_p|u_p)$ = conditional probability of phenotype f_p given genotype u_p, $\text{prog}_p(u_p)$ = progeny probability of offspring p having genotype u_p, $\text{po}(u_i|u_m, u_p)$ = conditional probability of individual i having genotype u_i given genotype of mate and offspring, *tot* = sum of all possible numerator terms of individual i.

The conditional probabilities, $\text{po}(u_i|u_m, u_p)$, are given in Table 3 for a single-locus trait with two alleles.

Genotype probability

The prior probability [Eq. (3)], the phenotype of the individual and its progeny probability [Eq. (4)] can be combined to estimate the probability of individual i having genotype u_i as follows:

$$\text{prob}_i(u_i) = \text{prior}_i(u_i)\, g(f_i|u_i)\, \text{prog}_i(u_i) \Big/ \left\{ \sum_{v=1}^{k} \text{prior}_i(v)\, g(f_i|v)\, \text{prog}_i(v) \right\} \quad (5)$$

Table 3. The conditional probability of the individual having genotype u_i given the genotype of the mate (u_m) and progeny (u_p), $\text{po}(u_i|u_m, u_p)$

	$u_m = AA$			$u_m = Aa$			$u_m = aa$		
$u_i =$	AA	Aa	aa	AA	Aa	aa	AA	Aa	aa
$u_p = AA$	0.67	0.33	0	0.67	0.33	0	0	0	0
Aa	0	0.33	0.67	0.33	0.33	0.33	0.67	0.33	0
aa	0	0	0	0	0.33	0.67	0	0.33	0.67

Algorithm

In human genetics, non-iterative procedures are available to estimate the genotype probabilities of a particular individual in the pedigree (Elston and Stewart 1971; Cannings et al. 1978). In most animal breeding situations, however, the size of the pedigree is too large to allow application of these procedures, and the genotype probabilities of all individuals are required. Therefore, an iterative algorithm to estimate genotype probabilities for all individuals in the family simultaneously has been developed using Eq. (5). The iterative and non-iterative algorithms, of course, give the same results.

The algorithm involves two steps in every iteration. First, the prior probabilities are calculated for all animals, starting with the oldest animal (individuals are listed by date of birth). Calculation of the prior probabilities involves the prior probabilities, the phenotype and the progeny probabilities of the parents as can be seen from Eqs. (3) and (5). In the first iteration, because progeny probabilities of parents have yet to be estimated, these probabilities cannot be used in estimating prior probabilities for the individual being evaluated. In later iterations, the progeny probabilities of the parents estimated in the previous iteration are used.

In the second step of each iteration, progeny probabilities are calculated starting with the youngest animal. Genotype probabilities of the mates, $\text{prob}_m(u_m, p)$ in Eq. (4), are calculated using the latest prior probabilities, the phenotype and the progeny probabilities from the previous round. In the first round, however, no progeny probabilities of mates are used.

After each iteration, the information on the phenotype of the animal and results of the first and second step are combined to calculate the genotype probabilities. The absolute difference between genotype probabilities in the current and the previous round of iteration is used as a convergence criterion.

Simulation

To evaluate the effectiveness of estimating genotype probabilities, a simulation program was written to generate data for a female sex-limited trait controlled by a single-locus. Data were simulated for a 10-year period with random mating to generate pedigree structure. All dams in the base population were unrelated and had known phenotypes. Every year, one-third of the dams were replaced by young females which originated from dams and sires in the population. At 2 years of age, when the females had their first offspring, the phenotype for the single-locus trait was measured.

Twenty sires were used simultaneously. The sires originated from five sires which had four sons each but no female offspring. A dam was allowed to produce only one son and, to avoid

Table 4. Distribution (%) of estimated frequency of the *A* allele in a sire for different true genotypes and number of offspring in a situation without dominance (population freq. = 0.5)

No. off.	True genotype	Estimated frequency of *A* allele (%)								
		0–5	5–15	15–30	30–45	45–55	55–70	70–85	85–95	95–100
3	*AA*	0	0	0	0	0	5	20	49	26
	Aa	1	5	8	3	65	3	9	5	1
	aa	28	46	24	2	0	0	0	0	0
6	*AA*	0	0	0	0	0	0	2	20	78
	Aa	1	2	2	0	92	0	1	1	1
	aa	77	19	4	0	0	0	0	0	0
12	*AA*	0	0	0	0	0	0	0	0	100
	Aa	0	0	0	0	99	0	0	0	0
	aa	100	0	0	0	0	0	0	0	0

Table 5. Distribution (%) of estimated frequency of the *A* allele in a sire for different true genotypes and number of offspring in a situation with dominance and 80% penetrance (population freq. = 0.5)

No. off.	True genotype	Estimated frequency of *A* allele (%)								
		0–5	5–15	15–30	30–45	45–55	55–70	70–85	85–95	95–100
3	*AA*	0	0	0	0	1	13	57	29	0
	Aa	0	2	12	39	10	11	22	4	0
	aa	3	18	31	34	4	4	6	0	0
6	*AA*	0	0	0	0	0	2	17	65	16
	Aa	1	2	8	33	36	2	7	10	1
	aa	13	20	28	30	6	1	1	1	0
12	*AA*	0	0	0	0	0	0	1	12	87
	Aa	0	2	4	19	69	0	1	1	4
	aa	27	23	22	20	8	0	0	0	0

inbreeding, these dams were unrelated to other animals in the population. The phenotype of the dam of the sire was recorded. Sires were used for 2 consecutive years, during which they all produced either 3, 6 or 12 female progeny with known phenotype.

A single autosomal locus model with two alleles (*A*, *a*) was used for inheritance of the trait. Data were simulated for frequencies of the *A* allele of 0.2, 0.5 and 0.8. A population in Hardy-Weinberg equilibrium and equal viability of all genotypes was assumed. The same two alternatives, I and II, as before for dominance and penetrance were considered: I – no dominance and complete penetrance; and II – complete dominance and 80% penetrance of the *aa* genotype (Table 1).

Genotype probabilities of the sires in the simulatd data were estimated using Eq. (5). The gene frequency and the relation between genotype and phenotype were taken to be equal to those used in the simulation of data. Iterations were stopped when the sum of absolute differences between genotype probabilities of sires in the current and the previous round was less than 0.01. Twenty replicates of each alternative were run.

For prediction of changes in frequency of genotypes due to selection, the average estimated allele frequencies of the selected sires as well as of the selected dams are needed. For selection purposes it is, therefore, important to estimate the frequency of the *A* allele in each sire. This frequency, $p(A)$, was calculated from the estimated probabilities of a sire having genotype *AA* and *Aa* as: $p(A) = \text{prob}(AA) + 0.5\ \text{prob}(Aa)$.

Results

No dominance and complete penetrance

Table 4 gives the distribution of the estimated frequency of the *A* allele in sires for different true genotypes and number of offspring for additive gene action (I) and a population frequency of 0.5. With three offspring, 65% of the *Aa* sires had an estimated frequency of the *A* allele between 45% and 55%. Heterozygous sires showed greater variation in estimated frequency than homozygous sires. The variation decreased markedly with increasing number of offspring. There was close to perfect agreement between estimated and true values when sires had 12 offspring.

Dominance and penetrance

The distribution of estimated allele frequencies of heterozygous and homozygous recessive sires showed considerable overlap in the case of complete dominance and 80% penetrance and a population frequency of *A* of 0.5 (Table 5). The estimates of *Aa* and *aa* sires showed much

Table 6. Influence of mode of inheritance and population frequency of *A* allele on distribution (%) of estimated frequency of the *A* allele in a sire having six offspring for different true genotypes (G)

Pop. freq.	G	Estimated frequency *A* allele (%)								
		0–5	5–15	15–30	30–45	45–55	55–70	70–85	85–95	95–100
I: no dominance and complete penetrance										
0.8	*AA*	0	0	0	0	0	0	0	3	97
	Aa	0	1	1	0	95	0	0	1	2
	aa	70	15	14	1	0	0	0	0	0
0.2	*AA*	0	0	0	0	0	1	14	15	70
	Aa	2	1	0	0	95	0	1	1	0
	aa	97	3	0	0	0	0	0	0	0
II: dominance and 80% penetrance										
0.8	*AA*	0	0	0	0	0	1	7	74	17
	Aa	0	0	0	18	24	2	11	40	5
	aa	0	0	6	40	26	5	9	14	0
0.2	*AA*	0	0	0	0	2	2	19	64	13
	Aa	3	6	15	30	38	2	3	3	0
	aa	62	20	12	5	1	0	0	0	0

larger variation than those of *AA* sires. More offspring were needed to distinguish between homozygous recessive and heterozygous sires. This also holds, to a lesser extent, for the distinction between homozygous dominant and heterozygous sires.

Population allele frequency

The influence of population frequency of the *A* allele on the distribution of estimated allele frequency in sires having six offspring is given in Table 6. With additive gene action (I), the estimates of the homozygous sires were affected by the population allele frequency. The population allele frequency had a very limited effect on the estimates of the heterozygous sires in the case of additive gene action.

With dominance and 80% penetrance, the population allele frequency had a big influence on the distribution of estimates of all genotypes (Tables 5 and 6).

Discussion

The method to estimate genotype probabilities has been described for a single-locus trait with two alleles, but extension to a larger number of alleles is straightforward. In the derivation of the method, it has been assumed that there is no inbreeding in the population. Inbreeding will lead to the use of information on a related animal in calculating both the prior and progeny probabilities of an individual, in which case the information on that animal might be weighted incorrectly. This can be avoided by excluding the information on offspring of the parents in calculating the prior probabilities and information on parents of the mates in calculating the progeny probabilities. However, omitting information from the more distant relatives will influence the effectiveness of the method. The consequences of inbreeding in combination with the mode of inheritance of the trait and ways to account for inbreeding need further investigation.

It has been assumed that mode of inheritance and allele frequencies are known. Effectiveness of estimation of genotype probabilities is expected to decrease when poor estimates of population parameters are used. Population allele frequencies mainly affect the prior probabilities of an animal. The influence of the population allele frequencies is expected to be smaller when information on phenotypes of parents can be used in estimating the prior probabilities.

Selection of individuals on allelic frequency is expected to affect the population allele frequency in subsequent generations and to cause deviations from Hardy-Weinberg equilibrium. The base population allele frequency is used to estimate the prior probabilities for animals in the base population. The effect of selection of individuals based on their phenotypes is expected to be accounted for when the information on parents is used in estimating the genotype probabilities. When information on parents is missing, the prior probabilities, $\tau(u)$, can be estimated from the current allele frequencies to account for the change in population allele frequency due to selection. However, these evaluations are likely to be biased if animals are culled for reasons related to their genotype before their phenotype has been assessed, or when animals with certain phenotypes are excluded from the evaluation.

The simulation results showed a very clear influence of the mode of inheritance on accuracy of estimating allele frequency in sires. The combination of dominance with incomplete penetrance resulted in a much lower accuracy of estimation compared with additive gene action. Thus, through these methods, it is possible to estimate the number of progeny required in a complex family history to get an accurate prediction of a sire's genotype. The efficiency can further be improved by making planned matings to animals with a certain phenotype.

The estimated allele frequency in individual animals can be used as a selection criterion when selecting for a single gene. The results given in this study can be used to derive the expected efficiency of a selection strategy. For example, assume that in situation II (dominance) all sires have six offspring with known phenotype. Selection of all sires having an estimated frequency of the dominant *A* allele of 85% or higher results in an average real frequency of *A* in the selected group of 0.88, 0.88 and 0.89 when the population allele frequency is 0.2, 0.5 and 0.8, respectively (Tables 5 and 6). The proportion of sires selected, how- ever, varied from 4% to 67%.

Acknowledgements. Thanks are expressed to the Niels Stensen Foundation, from The Netherlands, and the Canadian Association of Animal Breeders for financial support.

References

Cannings C, Thompson EA, Skolnick MH (1978) Probability functions on complex pedigrees. Adv Appl Probab 10:26–81

Carden AE, Hill WG, Webb AJ (1983) The inheritance of halothane susceptibility in pigs. Génét Sél Evol 15:65–82

Elston RC (1987) Human quantitative genetics. In: Weir BS, Eisen EJ, Goodman MM, Namkoong G (eds) Proc 2nd Int Conf Quant Genet. Sinauer, Sunderland, pp 281–282

Elston RC, Rao DC (1978) Statistical modeling and analysis in human genetics. Annu Rev Biophys Bioeng 7:253–286

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

Henderson CR (1984) Application of linear models in animal breeding. University of Guelph Press, Guelph

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

Kennedy BW, Schaeffer LR, Sorensen DA (1988) Genetic properties of animal models. J Dairy Sci 71(suppl 2):17–26

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

Smith C, Bampton PR (1977) Inheritance of reaction to halothane anaesthesia in pigs. Genet Res 29:287–292

Southwood OI, Simpson SP, Webb AJ (1988) Reaction of halothane anaesthesia among heterozygotes at the halothane locus in British Landrace pigs. Génét Sél Evol 20:357–366