ADVANCES IN GENETIC AND STATISTICAL MODELS TO PREDICT BREEDING VALUES

R. L. Fernando and L. R. Totir

Department of Animal Science, Iowa State University, Ames IA 50011, USA

INTRODUCTION

Even with the tremendous advances that have been made in molecular genetics, it is still not possible to directly observe breeding values. Thus, observable information must be used to statistically predict breeding values. One of the components of this prediction problem is the use of genetic and statistical models to relate observable information to unobservable breeding values. An important aspect of this component is the development of models that are realistic and yet lend themselves to statistical methods of prediction that are computationally feasible.

Until recently, the observable information used to predict breeding values was limited to pedigree relationships and trait phenotypes. Now, molecular-marker genotypes are becoming available as an additional source of information for prediction. In this paper we will briefly review the models that are used to predict breeding values and discuss some of the challenges in adapting these models to accommodate molecular information.

GENETIC MODELS

The simplest model to relate the trait phenotype y_i of animal i to its genotypic value g_i is

$$y_i = g_i + e_i \qquad (1)$$

In (1), g_i is defined as

$$g_i = E(y_i \mid t_i), \qquad (2)$$

and $e_i = y_i - g_i$, where t_i denotes the genotype of i for the trait (Bulmer, 1980). In most situations, the vector $\boldsymbol{g}$ of genotypic values is assumed to follow a multivariate normal distribution. Thus, the distribution of $\boldsymbol{g}$ is completely specified by its mean and covariance matrix.

In the models used to predict breeding values, the systematic components of e_i are further modeled using fixed and random effects (Henderson, 1984). The random effects in the model and the residual are often assumed to be normal. Thus, from (1), y_i is also normal. However, the phenotypes for some traits such as litter size or length of herd life are not well approximated by a normal distribution. These traits are analyzed using generalized linear mixed model theory, where a link function is used to relate the location parameters of the distribution for the trait phenotypes to the systematic component of the linear model (Ducrocq and Casella, 1996; Tempelman and Gianola, 1999). Consequently, the relationship between

trait phenotypes and $\boldsymbol{g}$ is more complex in these models. The relationship between marker genotypes and $\boldsymbol{g}$, however, is identical in both types of models. Thus, in this paper we will discuss only the model for $\boldsymbol{g}$.

Modeling the genetic mean under additive inheritance. Under additive inheritance, genetic groups (Thompson, 1979; Quaas, 1988; Westell, 1988) are used to model differences in means of founder animals that belong to different populations, and when pedigree information is missing, genetic groups are also used to model differences between individuals that belong to different generations.

Use of marker information. To see how genetic-marker information can be used as a source of information for $\boldsymbol{g}$, suppose genotypes are available for a marker locus that is closely linked to a QTL (MQTL). The genotypic value can now be modeled as

$$g_i = v_i^m + v_i^p + u_i \qquad (3)$$

where v_i^m and v_i^p are the additive effects of the maternal and paternal alleles at the MQTL, and u_i is the additive effect of the remaining QTL (RQTL). It is assumed that the marker is not linked to any of the RQTL. Let r be the recombination rate between the MQTL and the marker locus, and let σ_Q^2 be the additive variance at the MQTL.

Given a sufficient number of generations of random mating, even closely linked loci can become statistically independent. In the genetics literature, this independence is called gametic phase equilibrium or linkage equilibrium. Thus, if the marker locus and the MQTL are in gametic phase equilibrium, the marker genotypes do not contain any information on the mean of v_i^m and v_i^p. However, even when pure breeds are in gametic phase equilibrium, differences in allele frequencies between breeds result in gametic phase disequilibrium in crossbred animals. Note that any allele in a crossbred animal originates in a purebred founder. Thus, following Wang *et al.* (1998), the conditional mean for v^m_i, for example, can be written as

$$E\left(v_i^m\right) = \sum_l E\left(v_i^m \mid Q_i^m \leftarrow B_l, \boldsymbol{M}\right)\Pr(Q_i^m \leftarrow B_l \mid \boldsymbol{M})$$

$$= \sum_l \mu_l \Pr\left(Q_i^m \leftarrow B_l \mid \boldsymbol{M}\right), \qquad (4)$$

where $Q^m{}_i \leftarrow B_l$ denotes that the maternal MQTL allele of animal i originates in breed l, μ_l is the mean of $v^x{}_i$ in breed l, and $\Pr(Q^m{}_i \leftarrow B_l|\boldsymbol{M})$ is the conditional probability that $Q^m{}_i$ originates in breed l given marker information. This probability can be thought of as the breed B_l composition of MQTL allele $Q^m{}_i$, and it can be obtained recursively as

$$\Pr\left(Q_i^m \leftarrow B_l \mid \boldsymbol{M}\right) = \Pr(Q_i^m \leftarrow Q_d^m \mid \boldsymbol{M}) \Pr(Q_d^m \leftarrow B_l \mid \boldsymbol{M}) + \Pr\left(Q_i^m \leftarrow Q_d^p \mid \boldsymbol{M}\right)\Pr(Q_d^p \leftarrow B_l \mid \boldsymbol{M}) \qquad (5)$$

where $\Pr(Q^x_i \leftarrow Q^m_d|\boldsymbol{M})$ denotes the conditional probability that Q^x_i descended from the maternal allele of d, and $\Pr(Q^x_i \leftarrow Q^p_d|\boldsymbol{M})$ denotes the conditional probability that Q^x_i descended from the paternal allele of d. In the following, we will refer to these probabilities as the maternal and paternal PDQ's for allele Q^x_i.

When marker genotypes are complete, computing PDQ's is straightforward (Hoeschele, 1993; van Arendonk *et al.,* 1994; Wang *et al.,* 1995). However, when some genotypes are missing, computing PDQ's exactly will require peeling, and this can be computationally infeasible for large pedigrees with many loops (Fernández *et al.*, 2001). Then, PDQ's for i may be approximated by ignoring animals that are more than a specified distance in the pedigree from i. Another alternative is to use a Markov chain Monte Carlo (MCMC) method to estimate the PDQ's (Grignola *et al*., 1996). However, in this case, the single-site Gibbs sampler should not be used for marker loci with more than two alleles, because it may not result in a chain that is irreducible (Thomas and Cortessis, 1992; Sheehan and Thomas, 1993). It has been shown that ESIP, an algorithm for sampling jointly all the missing genotypes at a locus, is guaranteed to give an irreducible Markov chain. Further, in a comparison with other samplers, ESIP was found to be the most efficient (Fernandez *et al.*, 2001).

Now, consider the situation where the marker and the MQTL are in disequilibrium in the pure breeds. For this situation, the formulae of Wang *et al.* (1998) can be extended as:

$$E\left(v_i^m \mid \boldsymbol{M}\right) = \sum_l \sum_k \mu_{lk} \Pr\left(Q_i^m \leftarrow B_{lk} \mid \boldsymbol{M}\right), \qquad (6)$$

where μ_{lk} is the mean of the maternal MQTL allele in a founder belonging to breed B_l with maternal marker allele M_k, and $\Pr(Q^m_i \leftarrow B_{lk}|M)$ is the probability that Q^m_i can be traced to a founder belonging to breed B_l with maternal marker allele M_k. We will refer to this probability as the breed B_{lk} composition of MQTL allele Q^m_i, and it can be computed recursively using the PDQ's as

$$\Pr\left(Q_i^m \leftarrow B_{lk} \mid \boldsymbol{M}\right) = \Pr\left(Q_i^m \leftarrow Q_d^m \mid \boldsymbol{M}\right)\Pr\left(Q_d^m \leftarrow B_{lk} \mid \boldsymbol{M}\right) + \Pr\left(Q_i^m \leftarrow Q_d^p \mid \boldsymbol{M}\right)\Pr\left(Q_d^p \leftarrow B_{lk} \mid \boldsymbol{M}\right) \qquad (7)$$

Modeling genetic covariances under additive inheritance. Under additive inheritance, the genetic covariance matrix can be computed using the well known tabular method (Emik and Terrill, 1949) for pure-bred pedigrees, and using an extension of this method for multi-breed

pedigrees (Elzo, 1990; Lo *et al*., 1993). More importantly, efficient algorithms have been developed to invert these matrices, where computing time is linearly related to the size of the matrix (Henderson, 1976; Elzo, 1990; Lo *et al*., 1993). Thus, under additive inheritance, Henderson's mixed model equations can be used to predict breeding values for very large pedigrees.

Use of marker information. Even when the marker locus and the MQTL are in gametic phase equilibrium, markers provide information on covariances between relatives (Chevalet *et al.,* 1984; Fernando and Grossman, 1989). Thus, the conditional covariance matrix $\boldsymbol{G}_{g|r,m}$ of $\boldsymbol{g}$ given relationship and marker information is different from this conditional covariance matrix $\boldsymbol{G}_{g|r}$ given only relationship information. Using $\boldsymbol{G}_{g|r\,m}$ in Henderson's mixed model equations would give marker assisted BLUP (MABLUP) for $\boldsymbol{g}$. This, however, requires obtaining the inverse of $\boldsymbol{G}_{g|r,m}$, which is not sparse. On the other hand, the covariance matrix $\boldsymbol{G}_{v|r\,m}$ of the effects of MQTL alleles is sparse, and it can be inverted efficiently (Fernando and Grossman, 1989). Thus, to use Henderson's mixed model equations for BLUP, the model should include the effects of each animal's maternal and paternal MQTL alleles in addition to the effect of the RQTL. For a pedigree with n animals, this results in $2n$ additional equations. However, because the inverse of the gametic covariance matrix is sparse, the mixed model equations can be solved by iteration on data. Then, the BLUP of g_i is obtained as

$$\hat{g}_i = \hat{v}_i^m + \hat{v}_i^p + \hat{u}_i \qquad (8)$$

where $\hat{v}_i^m$, $\hat{v}_i^p$ and $\hat{u}_i$ are the BLUP's of the additive effects of the MQTL alleles and the RQTL.

As mentioned earlier, even when pure breeds are in gametic phase equilibrium, differences in allele frequencies between breeds result in gametic phase disequilibrium in crossbred animals. Wang *et al.* (1998) gave formulae to compute the variance of MQTL alleles in such crossbred animals when the pure breeds are in equilibrium. As shown here, these formulae can be extended to compute the variance of MQTL alleles when the pure breeds are in gametic phase disequilibrium.

Recall that any MQTL allele in a crossbred individual can be traced to a purebred founder. So, conditional on marker information, the variance of v^m_i, for example, can be written as

$$Var\left(v_i^m \mid \boldsymbol{M}\right) = E\left[Var\left(v_i^m \mid Q_i^m \leftarrow B_{lk}, \boldsymbol{M}\right)\right] + Var\left[E\left(v_i^m \mid Q_i^m \leftarrow B_{lk}, \boldsymbol{M}\right)\right] \qquad (9)$$

The first term of (9) can be written in terms of the conditional breed compositions as

$$E\left[Var\left(v_i^m \mid Q_i^m \leftarrow B_{lk}, \boldsymbol{M}\right)\right] = \sum_l \sum_k \sigma^2_{lk} \Pr\left(Q_i^m \leftarrow B_{lk} \mid \boldsymbol{M}\right), \qquad (10)$$

where $\sigma^2_{lk} = Var(v^m_i \mid Q^m_i \leftarrow B_{lk}, \boldsymbol{M})$. The second term of (9) can also be written in terms of the conditional breed compositions as

$$Var\left[E\left(v_i^m \mid Q_i^m \leftarrow B_{lk}, \boldsymbol{M}\right)\right] = \sum_l \sum_k \left(\mu_{lk} - \bar{\mu}_i^m\right)^2 \Pr\left(Q_i^m \leftarrow B_{lk} \mid \boldsymbol{M}\right), \qquad (11)$$

where $\bar{\mu}_i^m$ is the conditional mean of v^m_i given $\boldsymbol{M}$:

$$\bar{\mu}_i^m = \sum_l \sum_k \mu_{lk} \Pr\left(Q_i^m \leftarrow B_{lk} \mid \boldsymbol{M}\right). \qquad (12)$$

Substituting (10) and (11) in (9) gives

$$Var\left(v_i^m \mid \boldsymbol{M}\right) = \sum_l \sum_k \left[\sigma^2_{kl} + \left(\mu_{lk} - \bar{\mu}_i^m\right)^2\right] \Pr\left(Q_i^m \leftarrow B_{lk} \mid \boldsymbol{M}\right) \qquad (13)$$

Now, consider computing the covariance between v^m_i and v^x_j, where x is m or p and j is not a direct descendant of individual i. Note that Q^m_i is either Q^m_d, the maternal allele, or Q^p_d, the paternal allele, of d the mother of i. Thus, the conditional covariance between v^m_i and v^x_j can be computed recursively as

$$\begin{aligned} Cov\left(v_i^m, v_j^x \mid \boldsymbol{M}\right) = {} & Cov\left(v_d^m, v_j^x \mid \boldsymbol{M}\right) \Pr\left(Q_i^m \leftarrow Q_d^m \mid \boldsymbol{M}\right) \\ & + Cov\left(v_d^p, v_j^x \mid \boldsymbol{M}\right) \Pr(Q_i^m \leftarrow Q_d^p \mid \boldsymbol{M}) \end{aligned} \qquad (14)$$

This is identical to the formula used when the marker locus and the MQTL are in equilibrium. However, because the conditional variances under disequilibrium are different from the variances under equilibrium and because $Cov(v^m_d, v^x_j|\boldsymbol{M})$ or $Cov(v^p_d, v^x_j|\boldsymbol{M})$ may reduce to a variance, using (14) gives different results under disequilibrium than under equilibrium.

Multiple traits. The genetic merit of an animal often is a function of several traits, and phenotypic values from these or other related traits may be used jointly to predict breeding values. Theory and principles used to model genetic means for single traits are also applicable here.

Assuming pleiotropy is the basis for the correlation between traits, the covariance between the genotypic values g_i^1 of animal i for trait 1 and g_j^2 of animal j for trait 2 is

$$Cov\left(g_i^1, g_j^2\right) = a_{ij}\sigma_{12} \qquad (15)$$

where a_{ij} is the additive relationship coefficient between animals i and j, and σ_{12} is the genetic covariance between traits 1 and 2. Thus an analysis with k traits involves $k(k + 1)/2$ genetic variance and covariance parameters.

The situation is quite different for modeling covariances of MQTL effects. Consider a MQTL with two alleles. In this case, it can be shown that the genetic variance covariance matrix between traits for the MQTL has rank one. Thus, the correlation between the effects of an MQTL allele on any two traits is one. As a result, for each MQTL allele, only the effect v_{ti}^x on a single trait t should be included in the mixed model equations. The effect $v_{t'i}^x$ on any other trait t' can be written as

$$v_{t'i}^x = \beta_{tt'} v_{ti}^x \qquad (16)$$

where $\beta_{tt'}$ is the regression coefficient of $v_{t'i}^x$ on v_{ti}^x. Thus, to model covariances at an MQTL with two alleles, in an analysis with k traits, only one variance parameter and k - 1 regression parameters are required. For an MQTL with l alleles, the effects on l - 1 traits can be included in the mixed model equations. The effects on the remaining traits can be written in terms of the effects that were included in the mixed model equations. Now, to model covariances at an MQTL with l alleles, in an analysis with k traits, $(l - 1)l/2$ variance covariance parameters are needed for the effects included in the mixed model equations, and $(l - 1)(k - l + 1)$ regression parameters are needed for the remaining effects.

Modeling the genetic mean and covariances under dominance. Lo *et al.* gave theory to compute genotypic means and covariances in a two-breed population. They showed that the genetic mean could be written as a linear combination of five location parameters, and that the genetic covariance is a linear combination of 25 dispersion parameters. Perez-Enciso *et al.* (2001) extended this theory to include information from genetic markers. However, the genotypic variance covariance matrix constructed by this theory is not sparse, and thus, is not useful for predicting breeding values with large pedigrees.

Finite locus models. If a finite locus model is assumed, the conditional mean of the genotypic value could be estimated by MCMC methods (Fernando and Grossman, 1996; Goddard, 1998; Stricker and Fernando, 1998). The feasibility of this method will depend on the number of loci assumed in the analysis. The effect of the number of loci on genetic evaluations was studied using data simulated with a large number of loci. For several small pedigrees, genetic evaluations obtained by BLP were compared with the conditional means estimated by MCMC. For traits with low heritability, conditional means estimated using models with as few as three loci were very close to the BLP evaluations (Totir *et al.* 2001). Strategies to improve the efficiency of the MCMC methods are being investigated.

DISCUSSION

Most of the models used in the prediction of breeding values assume that inheritance is additive. Under additive inheritance, the inverse of the genotypic covariance matrix can be computed efficiently and it is very sparse. Thus, under additive inheritance Henderson's mixed

model equations can be used to obtain BLUP with very large pedigrees. Further, under additive inheritance, marker information can be incorporated to obtain MABLUP. Under non-additive inheritance, however, this is not the case.

It is often stated that the MABLUP approach requires assuming the MQTL effects are normally distributed, and thus each founder MQTL allele is unique. However, even if the MQTL has only two alleles, the gametic covariance matrix used in the MABLUP approach is correct, and BLUP does not require the random effects in the model to be normally distributed. On the other hand, in a population undergoing selection, the usual mixed model equations do not give BLUP unless the trait values and genotypic values are jointly normal. If multivariate normality is a good approximation for $\boldsymbol{y}$ and $\boldsymbol{g}$ when markers are not used in BLUP, this approximation should still hold when markers are used. Also, BLUP of g_i obtained from (8) is identical to that obtained directly by fitting $\boldsymbol{g}$ in the model and using $\boldsymbol{G}_{g|r,\,m}$ in the mixed model equations. Thus, if normality is a good approximation for $\boldsymbol{y}$ and $\boldsymbol{g}$, even in a population undergoing selection, using (8) should yield MABLUP.

The formulae presented here for incorporating markers in BLUP assume that the parental origin of marker alleles is known. They can be extended to accommodate unknown origin of the marker alleles (Wang *et al.,* 1995). When marker genotypes are missing, equation (14) does not give exact results. When marker haplotypes are used for BLUP (Goddard, 1992), the linkage phase of the markers needs to be known. Again, when the linkage phase is not know, results are not exact.

The above problems and the difficulties with non-additive inheritance can be overcome by using MCMC methods. However, more research will be needed before MCMC methods can be used for genetic evaluation with large livestock pedigrees.

ACKNOWLEDGMENTS
This work was partially funded by award no. 2002-35205-1156 of the National Research Initiative Competitive Grants Program of the USDA.

REFERENCES

Bulmer, M.G. (1980) "The Mathematical Theory of Quantitative Genetics" Clarendon Press, Oxford.
Chevalet, C., Gillois, M. and Khang, J.V.T. (1984) *Genet. Sel. Evol.* **16** : 431-444.
Ducrocq, V. and Casella, G. (1996) *Genet. Sel. Evol.* **28** : 505-529.
Elzo, M. A. (1990) *J. Anim. Sci.* **68** : 1215-1228.
Emik, L.O. and Terrill, C.E. (1949) *J. Hered.* **40** : 51-55.
Fernandez, S.A. (2001) PhD thesis, Iowa State University.
Fernandez, S.A., Fernando, R.L., Gulbrandtsen, B., Totir, L.R. and Carriquiry, A.L. (2001) *Genet. Sel. Evol*. **33** : 337-367
Fernando, R.L. and Grossman, M. (1989) *Genet. Sel. Evol.* **21** : 467-477.
Fernando, R.L. and Grossman, M. (1996) *Proc. Forty-Fifth Annu. Natl. Breeders Roundtable*, 19-28, Poult. Breeders Am. and US Poult. Egg Assoc., Tucker, GA.
Goddard, M.E. (1992) *Theor. Appl. Genet.* **83** : 878-886.

Goddard, M.E. (1998) *Proc. 6th WCGALP* **XXVI** : 33-36.
Gringola, F.E., Hoeschele, I. and Tier, B. (1996) *Genet. Sel. Evol.* **28** : 479-490.
Henderson, C.R. (1976) *Biometrics* **32** : 69-83.
Henderson, C.R. (1984) <<Applications of Linear Models in Animal Breeding.>> Univ. Guelph, Guelph, Ontario, Canada.
Hoeschele, I. (1993) *J. Dairy Sci.* **76** : 1693-1713.
Lo, L.L., Fernando, R.L., Cantet, R. J.C. and Grossman, M. (1995) *Theor. Appl. Genet.* **90** : 49-62.
Lo, L.L., Fernando, R.L. and Grossman, M. (1993) *Theor. Appl. Genet.* **87** : 423-430.
Pérez-Enciso, M., Fernando, R.L., Bidanel, J. and Roy., P.L. (2001) *Genetics* **159** : 413-422.
Quaas, R.L. (1988) *J. Dairy Sci.* **71** : 1338-1345.
Sheehan, N. and Thomas, A. (1993) *Biometrics* **49** : 163-175.
Stricker, C. and Fernando, R.L. (1998) *Proc. 6th WCGALP* **XXVI :** 25-32.
Tempelman, R.J. and Gianola, D. (1999) *J. Dairy Sci.* **82** : 1834-1847.
Thomas, D.C. and Cortessis, V. (1992) *Hum. Hered.* **42** :63-76.
Thompson, R. (1979) *Biometrics* **35** : 339-353.
Totir, L.R., Fernando, R.L. and Fernandez, S.A. (2001) *J. Anim. Sci.* **79**(Suppl. 1) : 191.
Van Arendonk, J. A.M., Tier, B. and Kinghorn, B. (1994) *Genetics* **137** : 319-329.
Wang, T., Fernando, R.L. and Grossman, M. (1998) *Genetics* **148** : 507-515.
Wang, T., Fernando, R.L., van der Beek, S., Grossman, M. and van Arendonk, J. A. M. (1995) *Genet. Sel. Evol.* **27** : 251-274.
Westell, R.A. (1988) *J. Dairy Sci.* **71** : 1310-1318.