A STRATEGY TO IMPROVE THE COMPUTATIONAL EFFICIENCY OF MARKER ASSISTED GENETIC EVALUATION UNDER FINITE LOCUS MODELS

L. R. Totir[1], R. L. Fernando[1], J.C.M. Dekkers[1] and S. A. Fernandez[2]

[1] Department of Animal Science, Iowa State University, Ames, IA 50011, USA, [2] Department of Statistics, The Ohio State University, Columbus, OH 43210, USA

INTRODUCTION

BLUP methodology can be used for marker assisted genetic evaluation (MAGE) (Fernando and Grossman, 1989). MAGE is most useful for traits that have low heritability. Lowly heritable traits are also known to have non additive inheritance. Although BLUP is computationally efficient under additive inheritance, it is inefficient under non additive inheritance.

Regardless of the mode of inheritance, under a finite locus model, the best predictor (BP), which is the conditional mean of the genotypic value given trait and marker data, can be estimated by Markov chain Monte Carlo (MCMC) (Fernando and Grossman 1996, Goddard, 1998, Stricker and Fernando, 1998).

The single site Gibbs sampler has been widely used in genetic analyses. However, it is known that this sampler may yield unreliable results. For example, when a marker locus has more than two alleles, single site Gibbs may not result in a chain that is irreducible. Even when the generated chain is irreducible, mixing might be very slow. These problems can be overcome by sampling genotypes jointly from the entire pedigree. ESIP, an MCMC sampler that combines two peeling techniques (the Elston-Stewart algorithm and iterative peeling) can be used to generate joint genotype samples from the entire pedigree (Fernandez *et al.*, 2001).

The computational efficiency of ESIP is determined by two distinct processes: peeling and sampling. Peeling is used to calculate genotype probabilities, which are then used in reverse peeling to sample genotypes jointly from the entire pedigree (Fernandez *et al.*, 2001). Once the pedigree has been peeled, the computing time required for sampling genotypes by reverse peeling is comparable to the computing time for single site Gibbs. The computing time required for peeling depends on the complexity of the pedigree and the number of missing genotypes. For the situation considered below, the time required for peeling was about ten times that for sampling. For a single locus model, the same genotype probabilities are used repeatedly in reverse peeling. Thus, peeling needs to be done only once, and as a result the computing time required for peeling has negligible effect on the computational efficiency of ESIP. For multilocus models, in order to preserve a linear relationship between computational efficiency and the number of loci in the model, genotypes are sampled one locus at a time, conditional on the current genotype configuration at the other loci. Thus, whenever the sampler moves to a new locus, genotype probabilities must be recomputed by peeling. For a given locus, after peeling, k samples can be obtained before moving to the next locus. As the size of

k increases the computational efficiency of the sampler increases, but it also results in increased dependence between samples (poor mixing).

The effect of the number of samples (k) per peeling on the computational efficiency of ESIP is evaluated in this paper.

MATERIALS AND METHODS

For a simple pedigree with 14 individuals and no loops, data were simulated using a finite locus model that contained one quantitative trait locus flanked by two markers (MQTL), and 100 remaining QTL (RQTL) unlinked to the flanking markers. All QTL had allele frequencies of 0.5. The additive effect of the MQTL was 2 while the dominance effect was 0. Each RQTL had an additive effect of 0.2828, and 50 of the RQTL had a dominance effect of 0.2828, while the other 50 had a dominance effect of - 0.2828. An environmental variance of 63 was used. Thus, the simulated trait had a narrow sense heritability of 0.08 and a broad sense heritability of 0.11. The marker data were simulated assuming 12 alleles at each marker locus, and a recombination rate of 0.05 between each marker and the MQTL.

The simulated data were analyzed using a finite locus model that contained one MQTL and two RQTL. Parameters for the MQTL were set equal to those used for simulation. For the two RQTL, the parameters were derived so that they yielded the same first and second moments as the 100 RQTL of the simulation. Thus, the additive effect for each RQTL was 2, and one RQTL had a dominance effect of 2 while the other had a dominance effect of - 2. In the analysis, an environmental variance of 63 was used as well.

The conditional mean of the vector of genotypic values (**v**) given the vector of trait data (**y**) and the matrix of marker data (**M**) can be written as

$$E(\mathbf{v} \mid \mathbf{y}, \mathbf{M}) = \mathbf{1}\eta + \sum_{G} \mathbf{v}_{\mathbf{G}} \Pr(\mathbf{G} \mid \mathbf{y}, \mathbf{M})$$

where **1** is an 14x1 vector of ones; η is a fixed effect common to all individuals; and $\mathbf{v}_{\mathbf{G}}$ is the 14x1 vector of genotypic values that corresponds to the genotypic configuration **G**. A given genotypic configuration **G** can be represented as a 14x3 matrix, where each column of **G** is equivalent to the 14x1 vector of genotypes for a given QTL (MQTL or RQTL). BP's were estimated by averaging joint genotype samples generated by several ESIP samplers.

We use the notation ESIP(1:k) for an ESIP sampler with k samples per peeling. Samples generated by ESIP(1:1) are independent and thus, were used to calculate reference genetic evaluations for all members of the pedigree. These reference evaluations were estimated by averaging 1,000,000 ESIP(1:1) samples.

Figure 1 shows the behavior of ESIP(1:1) in terms of the mean sum of squares of the difference between the reference genetic evaluations and genetic evaluations computed at different stages of a chain of 1,000,000 samples. ESIP(1:1) reached a high level of accuracy in a short time, and then the accuracy improved slowly throughout the run.

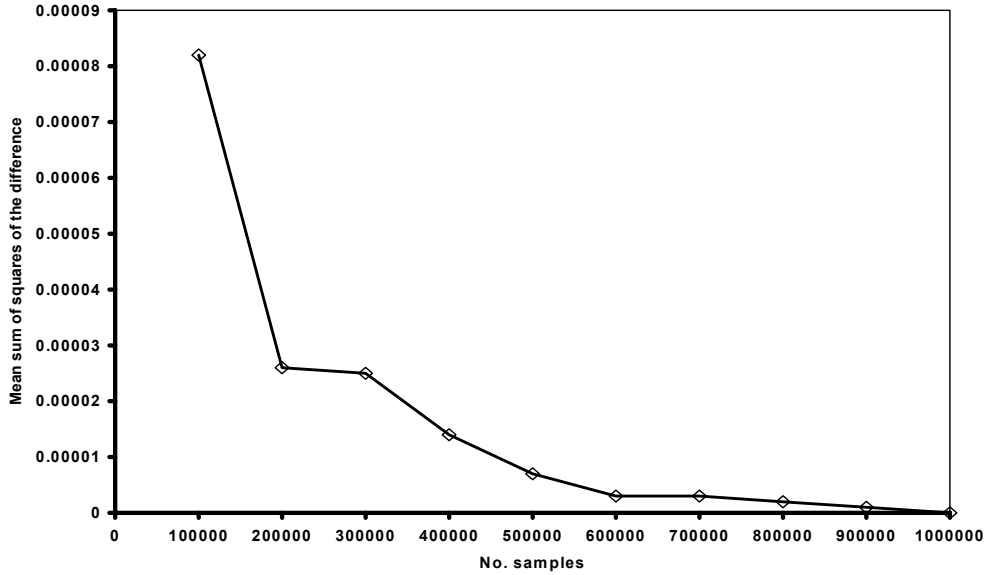

Figure 1. Mean sum of squares of the difference of genetic evaluations obtained after n from those obtained after 10^6 samples for ESIP(1:1).

To have a meaningful comparison between the samplers, genetic evaluations were obtained by running each sampler for the same amount of time. The time allocated to each sampler was equal to the time needed by ESIP(1:1) to generate 100,000 samples. At the end of the allocated time, the absolute difference between the reference genetic evaluations and ESIP(1:k) were scaled by the genetic standard deviation. For each ESIP(1:k), the maximum and the mean of the scaled absolute differences, and the scaled standard deviation (S.S.D) of the difference of genetic evaluations from the reference genetic evaluations were used to summarize the performance of the sampler.

RESULTS AND DISCUSSION

Table 1 summarizes the performance of several ESIP(1:k) samplers. As k increased, the number of samples generated in the allocated time increased as well. The rate of increase, however, decreased rapidly after k=100. In terms of accuracy, ESIP(1:5) performed best. For k larger than 15, increased dependence between samples resulted in reduced accuracy. For example, ESIP(1:100) was approximately two times less accurate than ESIP(1:5), while ESIP(1:1000) was ten times less accurate than ESIP(1:5). ESIP(1:100,000) had the highest dependence between samples and thus, it was the least accurate sampler.

The finite locus model used in this analysis had only three loci, one MQTL and two RQTL. For lowly heritable traits, two RQTL provide a good approximation for the polygenic component (Totir *et al.* 2001). MAGE is known to have the greatest advantage for lowly heritable traits.

When herititability is low, for models that fit a single MQTL, ESIP(1:k) with k between five and 15 will be most efficient.

Table 1. The performance of ESIP(1:k) , for a fixed amount of computing time, in terms of the maximum and mean of the scaled absolute difference, and the scaled standard deviation (S.S.D) of the difference between BP's obtained by ESIP(1:k) and reference BP's.

Sampler	*No samples*	*Maximum*	*Mean*	*S.S.D*
ESIP(1:1)	100,000	0.0059	0.0026	0.0032
ESIP(1:5)	361,433	0.0031	0.0014	0.0017
ESIP(1:10)	535,510	0.0059	0.0016	0.0021
ESIP(1:15)	640,000	0.0035	0.0014	0.0018
ESIP(1:20)	716,940	0.0048	0.0021	0.0023
ESIP(1:100)	950,725	0.0089	0.0027	0.0035
ESIP(1:1000)	1,025,000	0.0293	0.0107	0.0137
ESIP(1:100,000)	1,041,270	0.2339	0.0998	0.1220

For traits with high heritability, a larger number of RQTL must be used to approximate the polygenic component (Totir *et al* 2001). As the number of RQTL in the model increases, the dependence between samples generated by ESIP(1:k) with large k will increase as well. Thus, for traits with high heritability or models with several MQTL, an ESIP(1:k) with k smaller than five might be more efficient.

ACKNOWLEDGEMENTS
Research support was provided by award no. 2002-35205-1156 of the National Research Initiative Competitive Grants Program of the USDA.

REFERENCES
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*, p 19-28. Poult. Breeders Am. and US Poult. Egg Assoc., Tucker, GA.
Goddard, M.E. (1998) *Proc. 6th WCGALP*, **26** : 33-36.
Stricker, C. and Fernando, R.L. (1998) *Proc. 6th WCGALP*, **26** : 25-32.
Totir, L.R., Fernando, R.L. and Fernandez, S.A.(2001) J. Anim. Sci. **79** (Suppl.1) : 191.