J. Dairy Sci. 97:3943–3952
http://dx.doi.org/10.3168/jds.2013-7752

Using recursion to compute the inverse of the genomic relationship matrix

I. Misztal,*[1] A. Legarra,† and I. Aguilar‡
*Department of Animal and Dairy Science, University of Georgia, Athens 30602-2771
†INRA, UR631-SAGA, BP 52627, 31326 Castanet-Tolosan Cedex, France
‡Instituto Nacional de Investigación Agropecuaria, Las Brujas 90200, Uruguay

ABSTRACT

Computing the inverse of the genomic relationship matrix using recursion was investigated. A traditional algorithm to invert the numerator relationship matrix is based on the observation that the conditional expectation for an additive effect of 1 animal given the effects of all other animals depends on the effects of its sire and dam only, each with a coefficient of 0.5. With genomic relationships, such an expectation depends on all other genotyped animals, and the coefficients do not have any set value. For each animal, the coefficients plus the conditional variance can be called a genomic recursion. If such recursions are known, the mixed model equations can be solved without explicitly creating the inverse of the genomic relationship matrix. Several algorithms were developed to create genomic recursions. In an algorithm with sequential updates, genomic recursions are created animal by animal. That algorithm can also be used to update a known inverse of a genomic relationship matrix for additional genotypes. In an algorithm with forward updates, a newly computed recursion is immediately applied to update recursions for remaining animals. The computing costs for both algorithms depend on the sparsity pattern of the genomic recursions, but are lower or equal than for regular inversion. An algorithm for proven and young animals assumes that the genomic recursions for young animals contain coefficients only for proven animals. Such an algorithm generates exact genomic EBV in genomic BLUP and is an approximation in single-step genomic BLUP. That algorithm has a cubic cost for the number of proven animals and a linear cost for the number of young animals. The genomic recursions can provide new insight into genomic evaluation and possibly reduce costs of genetic predictions with extremely large numbers of genotypes.

Key words: genomic relationship matrix, recursion, genomic selection, single-step BLUP, preconditioned conjugate gradient (PCG) algorithm

INTRODUCTION

When only a fraction of animals are genotyped, a genomic relationship matrix **G** can be combined with a numerator relationship matrix **A** into a genomic-pedigree relationship matrix **H** (Legarra et al., 2009). Such a matrix is complicated, but has a simple inverse (Aguilar et al., 2010; Christensen and Lund, 2010). When the inverse of **H** is used with BLUP, the method is called single-step genomic BLUP (**ssGBLUP**). Advantages of ssGBLUP include simplicity of use (yet another BLUP), relatively high accuracy (Chen et al., 2011; Christensen et al., 2012; Gray et al., 2012), known and explicit control of biases because of different base populations in **A** and **G** as opposed to unknown properties of multistep methods (Tsuruta et al., 2011; Vitezica et al., 2011), and possible accounting for selection bias for genotyped animals (Patry and Ducrocq, 2011; VanRaden, 2012). Accuracy of ssGBLUP can be further improved by using a weighted **G** (Wang et al., 2012), which mimics Bayesian regressions.

The most expensive operation with ssGBLUP, as proposed by Aguilar et al. (2010) and Christensen and Lund (2010), is creating and then inverting **G**. Both operations have an approximately cubic cost with the number of genotypes. With efficient computing algorithms, both operations are feasible for up to 100,000 genotypes (Aguilar et al., 2011; Masuda and Suzuki, 2013). However, the US dairy industry has already collected over 400,000 Holstein genotypes; over 80% of genotypes are for animals without a BLUP evaluation, with a very slow increase in the number of genotypes for proven bulls (Council on Dairy Cattle Breeding, 2013).

Several approaches that do not require the inverse of **G** ($\mathbf{G}^{-1}$) have been proposed for ssGBLUP. Misztal et al. (2009) presented unsymmetric equations where only **H** was required. However, creating **H** directly is complicated. Legarra and Ducrocq (2012) presented

Received November 22, 2013.
Accepted February 10, 2014.
[1]Corresponding author: ignacy@uga.edu

different unsymmetric equations where inverses that were difficult to obtain were not required. Unsymmetric equations exhibited declining convergence with a larger number of genotypes (Aguilar et al., 2013), although they may be useful when a suitable preconditioner is found. Fernando et al. (2013) proposed a method where genotypes of nongenotyped animals were imputed and the final set of equations included SNP effects for all animals plus extra polygenic terms. However, the imputation is expensive; the volume of the imputed data are extremely large for big populations (up to dozens of millions of individuals for dairy cattle), and existing software is not applicable.

The cost of creating $\mathbf{A}$ by a tabular method (Emik and Terrill, 1949) is quadratic, and the cost to invert it directly is cubic. However, Henderson (1976) developed an algorithm based on recursion to obtain the inverse of $\mathbf{A}$ ($\mathbf{A}^{-1}$) directly at linear cost. Subsequently, $\mathbf{A}^{-1}$ can be computed for millions of animals in seconds. Faux et al. (2012) used Henderson's ideas and conditioned animals on a small number of relatives. However, the cost of their algorithm was higher than that by regular inversion. The purpose of our study was to determine whether recursion is useful in obtaining $\mathbf{G}^{-1}$ at a reasonable cost for a large number of genotypes.

MATERIALS AND METHODS

The method of Henderson (1976) to create $\mathbf{A}^{-1}$ directly depends on the recursion

$$u_{\mathrm{i}} = 0.5\left(u_{s_i} + u_{d_i}\right) + \varphi_i,$$

$$u_i = 0.5(u_{si} + u_{di}) + \varphi_i,$$

where u_i is the animal effect for animal i; s_i and d_i refer to the sire and dam of animal i, respectively; φ_i is Mendelian sampling; and founders of the pedigree are assumed to be unrelated. In matrix notation and with a genetic variance of σ_a^2 of 1 to simplify notation,

$$\mathbf{u} = \mathbf{P}\mathbf{u} + \Phi,\ \mathrm{var}(\Phi) = \mathbf{M},$$

and

$$\mathbf{A}^{-1} = (\mathbf{I} - \mathbf{P})'\mathbf{M}^{-1}(\mathbf{I} - \mathbf{P}) = \mathbf{T}'\mathbf{M}^{-1}\mathbf{T},$$

where $\mathbf{P}$ relates animals to parents; $\mathbf{T}$ is a triangular matrix if animals are ordered from oldest to youngest; and $\mathbf{M}$ is a diagonal matrix. Subsequently, $\mathbf{A}^{-1}$ can be created as a sum of outer products

$$\mathbf{A}^{-1} = \sum_i (\mathbf{t}'_{i,1:n}\,\mathbf{t}_{i,1:n}/m_i),$$

where $\mathbf{t}_{i,1:n}$ contains no more than 3 nonzero elements. Ignoring inbreeding, the value of m_i is (4 − number of known parents)/4. Henderson's rules are simple: when $\mathbf{u} \sim N(0, \mathbf{A}\sigma_a^2)$, animals are ordered from the oldest to the youngest, and all animals (including base animals) are included in the pedigree, $u_i \left| u_1, u_2, \ldots, u_{i-1}, u_{i+1}, \ldots, u_n = u_i \right| u_{s_i}, u_{d_i}$, or the conditional for an animal includes only its parents but not the rest of individuals in the pedigree. For instance, when only animals with records are included in $\mathbf{A}$ or older animals are conditioned on the younger animals, the conditional of u_i may involve more than 2 animals.

With genomic relationships, $\mathbf{u} \sim N(0, \mathbf{G}\sigma_a^2)$. The joint distribution of $u_1, \ldots, u_n$ can be written as

$$p(u_1, \ldots, u_n) = p(u_1)p(u_2|u_1)p(u_3|u_2, u_1)\ldots p(u_n|u_1, u_2, \ldots, u_{n-1}).$$

This decomposition is general and does not involve any particular ordering of individuals. Each of the conditional distributions can be written as

$$p\left(u_i|u_1, u_2, \ldots, u_{i-1}\right) \sim N \left[\mathbf{g}_{i,1:i-1}(\mathbf{G}_{1:i-1,1:i-1})^{-1}\mathbf{u}_{1:i-1}, g_{i,i} - \mathbf{g}_{i,1:i-1}(\mathbf{G}_{1:i-1,1:i-1})^{-1}\mathbf{g}'_{i,1:i-1}\right]$$

with $\mathbf{g}_{i,1:i-1}$ part of the ith row of $\mathbf{G}$, and with the following recursion equation:

$$u_i \mid u_1 \ldots u_{i-1} = \sum_{j=1}^{i-1} p_{ij}u_j + \varepsilon_i,$$

and

$$\mathbf{p}_{i,1:i-1} = \mathbf{g}_{i,1:i-1}(\mathbf{G}_{1:i-1:i-1})^{-1},\ \mathbf{M}_{i,i} = m_i = \mathrm{var}(\varepsilon_i) = \mathrm{g}_{i,i} - \mathbf{p}_{i,1:i-1}\mathbf{g}'_{i,1:i-1}. \quad [1]$$

Mimicking the developments of Henderson (1976) and Quaas (1988),

$$\mathbf{G}^{-1} = (\mathbf{I} - \mathbf{P})'\mathbf{M}^{-1}(\mathbf{I} - \mathbf{P}) = \mathbf{T}'\mathbf{M}^{-1}\mathbf{T},$$

where $\mathbf{T}$ is a triangular matrix as a result of the recursions of u_i on individuals $u_1 \ldots u_{i-1}$. Then $\mathbf{G}^{-1}$ can be created as a sum of outer products as