Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Predictive Performance Yielded by Approximate Genomic Matrices

E. Manfredi[1], **C. Carre**[1,2], **and M.A. Toro**[3]
[1]UMR1388 INRA/ENVT/ENSAT, Genphyse, Toulouse, France,
[2]Institut Mathematique de Toulouse, France [3]Universidad Politécnica de Madrid, Spain

ABSTRACT: Some methods for genetic prediction require the inverse of large genomic variance-covariance matrices amongst individuals. We studied by simulation the predictive performance of approximate inverses of genomic matrices, relatively to the usual complete genomic matrix. The approximate inverses were easy to compute but accuracies dropped by up to 30% according to the studied genetic scenarios.
Keywords: genetic prediction; genomic selection; genomic matrices

Introduction

Usual approaches for genetic prediction are based on individual or marker models. Individual models may be preferred because they generate n unknown effects (n: number of individuals) instead of m unknown (m: number of markers), and, usually, m > n.

In genomic BLUP, predictions are obtained by solving $\begin{bmatrix} X'X & X'Z \\ Z'Z & Z'Z + \gamma G^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ u \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$

where β are the unknown nuisance parameters and u are the genetic values. X and Z are incidence matrices and G is a genomic variance-covariance matrix amongst individuals built from marker data as $G = MM'k$. M is a matrix of marker genotypes, centered by columns, and k is a function of allele frequencies (e.g., Van Raden (2008)). The inverse of G must be computed using general algorithms which are time consuming. Alternatively, approximated genomic matrices might be easily computed with factorization techniques. Here, we tested by simulation the predictive performance of approximated, easily computed, inverses of genomic matrices.

Materials and Methods

Genomic matrices compared. A conceptual matrix to relate individuals is taken from the traditional infinitesimal model, where the numerator relationship matrix A can be factorized as $A = TDT'$ [I] (Quaas (1976); Henderson (1976)). T is a genetic transmission matrix with form T= $(I-1/2\ P)^{-1}$ and P is, for individuals sorted by increasing age, lower triangular with row i including a "1" in each column corresponding to the father f and the mother m of individual i. Under the infinitesimal additive model, D is a diagonal matrix with diagonal elements equal to the ratio between the variance due to Mendelian sampling, a function of inbreeding coefficients, and the additive genetic variance.

The approximated genomic matrices are based on [I]. Assume that a molecular score S_{ij} can be computed to measure realized inbreeding and co-ancestry relating individuals i and j. These scores can be used to identify parent-progeny pairs with good accuracy (e.g., Rohlfs et al. (2012)) and to provide molecular measures of inbreeding. The general form of the approximated genomic matrix is KΔK'. As detailed below, we tested two approximated genomic matrix by using marker data to build K and Δ. We computed the usual genomic matrix G as the reference.

The genomic matrix G was computed as G=MM'k where M was the column-centered matrix of SNP genotypes coded as 0 (homozygous 11), 1 (heterozygous) and 2 (homozygous 22), and k was twice the cross-product of allele frequencies: $k = 2\sum_j p_j q_j$ (Van Raden (2008)). Two approximated genomic matrices G1 and G2 were computed. The first one was $G1=T\Delta_1 T'$ with T representing genetic transmission between parents and progeny given by a fixed 0.5 coefficient (as in [I]), and Δ_1 is the diagonal of G. The second approximation was $G2 = T_2\Delta_2 T_2'$ where $T_2 = (\Delta_1 - Q)^{-1}$ with Q being a lower triangular matrix with two no null elements in each row: $Q_{if}=G_{if}$ and $Q_{im}=G_{im}$., f and m representing the father and the mother the ith. Individual. Δ_2 is diagonal with ith. diagonal element equal to G_{ii}-G_{if}-G_{im}+G_{ff}/4+G_{mm}/4+G_{mf}/2.

Data. Simulations were performed using the QMSim software (Sargolzaei and Schenkel (2009)). The simulated population had 1 base generation (25 individuals), 3 training generations (120 individuals) and the last generation (40 individuals) taken as prediction target. The phenotypes of base and target individuals were simulated but not used for prediction of phenotypes and genetic values of individuals in the target population. The simulated genome had 2 chromosomes of 1 Morgan each. The number of SNP markers was 2000 per chromosome. The phenotypes had variance 1 and overall heritability (infinitesimal + QTL effects) was 0.3. Two genetic scenarios were replicated 200 times: complete (100%) and intermediate (50%) proportion of genetic variance explained by QTL.

Results and discussion

Correlations between true additive genetic values and their predictions were computed separately for base individuals, individuals with known phenotype (to quantify Goodness of fit GF) and descendants whose genetic values are the prediction target (to quantify Predictive performance PP). The inverses of G1 and G2

were readily computed using algorithms for inverting triangular and diagonal matrices.

As expected, for both scenarios and all compared matrices, GF was higher and less variable than PP (table 1). The average GF yielded by the approximate matrices were slightly lower than the GF of the G matrix when markers explained 100% of total genetic variance (0.72 for G vs 0.64 for G1 and G2). GF were very similar for all genomic matrices when markers explained half of the total genetic variance (PP about 0.65).

The usual genomic matrix yielded higher PP than the approximated methods (Table 1). This was particularly true when QTL explained 100% of total genetic variance (losses of about 30% of PP for G1 and G2) and less marked when markers explained only 50% of total variance (losses smaller than 10% for G1 and G2). Fixing genetic transmission (which is equivalent to annul some co-ancestry coefficients among base individuals) has a detrimental effect on accuracy. Whenever realized co-ancestries among base individuals are well quantified by genetic markers, like in our simulated situation, the approximated methods lose data useful for prediction.

Table 1. Goodness of fit (GF) and Predictive performance (PP) yielded by approximate genomic matrices*

%Genetic Variance	**100**		**50**	
	GF	**PP**	**GF**	**PP**
G	.72±.09	.61±.15	.65±.10	.44±.20
G1	.64±.09	.40±.20	.65±.10	.41±.21
G2	.64±.09	.40±.20	.64±.09	.40±.21

* PP and GF are averages over 200 replicates

No differences were found between the approximated matrices. In G1, marker data are just used to find parent-progeny pairs and to estimate inbreeding coefficients. In G2 marker data are also used to modify coefficients representing genetic transmission and to improve the computation of inbreeding coefficients. According to present results G1 should be preferred because it produces the same accuracy than G2, and it has a simple and clear form.

Conclusion

Approximate genomic matrices have easily computed inverses because they ignore some distant co-ancestry coefficients. When such relationships are well described by markers, it is worthwhile to keep the complete genomic matrix. In other situations, accuracies provided by approximate matrices are comparable to that one yielded by the full genomic matrix.

Literature cited

Henderson, C.R. (1976). Biometrics 32: 69-83.

Quaas, R.L. (1976). Biometrics 32: 949-953.

Rohlfs, R.V., Fullerton, S.M., Weir, B.S. (2012). PLoS Genet 8: e1002469-e1002469.

Sargolzaei, M., and Schenkel, F.S. (2009). Bioinformatics, 25: 680-681. First published January 28, 2009, doi:10.1093/bioinformatics/btp045.

Van Raden, P.M. (2008). J Dairy Sci., 91(11):4414-4423.