OPTIMIZING GENOMIC PREDICTION:

STRATEGIES TO OBTAIN INVERSE OF LARGE RELATIONSHIP MATRICES

P. Faux [1], I. Misztal [2] and N. Gengler [1,3]

[1]Animal Science Unit, Gembloux Agro-Bio Tech, University of Liège, B-5030 Gembloux, Belgium
[2]Department of Animal and Dairy Science, University of Georgia, Athens 30602, GA, USA
[3]National Fund for Scientific Research, B-1000 Brussels, Belgium

CONTEXT: Genomic prediction of breeding values

- Genomic prediction (GBLUP) aims **to link the phenotypic variability to the genome-wide variability**, which is detected by thousands of Single Nucleotide Polymorphisms (**SNP**)
- An innovative and simple procedure (**single-step GBLUP**, *Misztal et al., 2009*) performs genomic prediction by **using all 3 sources of information at the same time** (phenotypes, genotypes and pedigrees)
- Key point of **ssGBLUP**: use of **an enhanced relationship matrix** (**H**, *Legarra et al., 2009*) as covariance matrix between random genetic effects, instead of **pedigree-based relationship matrix** (**A**) **in mixed model equations (MME)**

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z}+\alpha\mathbf{H}^{-1} \end{bmatrix} \qquad \mathbf{H}^{-1}=\mathbf{A}^{-1}+\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1}-\mathbf{A}_{22}^{-1} \end{bmatrix}$$

Left-hand side of MME — *Structure of inverse of **H***

- In order to get $\mathbf{H}^{-1}$, **3 matrices have to be inverted**:
  - **A** → Sparse inverse **easily computed** using particular rules *(see below)*
  - **G** → **Genomic relationship matrix** between genotyped animals; **inverse is dense**
  - $\mathbf{A}_{22}$ → **Pedigree-based relationship matrix between genotyped animals**; inverse may be sparse in some cases *(e.g. weakly-related animals)*

METHODS: Creation of a sparse triangular decomposition of the inverse

- **Direct creation of inverse of A** *(Henderson, 1976)* :

$$\mathbf{A}^{-1}=(\mathbf{T}^{-1})'\mathbf{D}^{-1}\mathbf{T}^{-1}$$

  - $\mathbf{T}^{-1}$ is an identity matrix $\mathbf{I}(n)$ that is filled in its lower part with max. $(2n\text{-}3)$ elements equal to -0.5
  - $\mathbf{D}^{-1}$ is a diagonal matrix with max. 3 different values: 2, 4/3 or 1
  - → **Based on this a priori knowledge of values, any element of $\mathbf{A}^{-1}$ can directly be computed without computation of the whole matrix**

- **Rules are modified** to fit to the case of **G** by **the creation of a sparse approximation of** $\mathbf{T}^{-1}$ ($\mathbf{T}^{*-1}$):

$$\mathbf{G}^{*-1}=(\mathbf{T}^{*-1})'\mathbf{D}^{-1}\mathbf{T}^{*-1}$$

  - For a given animal, elements of $\mathbf{T}^{*-1}$ are determined by **linear regression of relationships coefficients of closely-related animals** (i.e. having a relationship with this animal lower than ***p***, an arbitrary threshold) **on the relationships coefficients they have with this animal**
  - Computing the reverse equation, $\mathbf{D}=\mathbf{T}^{*-1}\mathbf{G}(\mathbf{T}^{*-1})'$, **we obtain a "close-to-diagonal" matrix D**
  - To approximate inverse of **D**, it is either submitted **to a new round** (with a new **p**) **into the same process** of approximation, or computed as **diagonal matrix made of inverse of its diagonal elements**

→ Because $\mathbf{T}^{*-1}$ is sparse, **creation of $\mathbf{G}^{*-1}$ involves less computations**

REFERENCES

- Henderson, C.R. 1976. **A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values**. Biometrics 32:69-83
- Legarra, A., I. Aguilar, and I. Misztal. 2009. **A relationship matrix including full pedigree and genomic information**. J. Dairy Sci. 92:4656-4663
- Misztal, I., A. Legarra, and I. Aguilar. 2009. **Computing procedures for genetic evaluation including phenotypic, full pedigree and genomic information**. J. Dairy Sci. 92:4648-4655

OBJECTIVE: Approximation of inverse of large relationship matrices

- **G** and $\mathbf{A}_{22}$ inverted so far **using dense matrix inversion algorithms** *(e.g. Gaussian elimination)*
- Increase of number of genotyped animals = **cubic increase of computation cost of these inverses**

→ Therefore, **an algorithm for a direct approximation** of the inverses of **G** and $\mathbf{A}_{22}$ is required

RESULTS: Different tests on 2 sets

- **1,718 dairy bulls**: Several rounds of approximation of inverse of **G**, assessments of **closeness of approximation** and **sparsity**, **correlation between MME's solutions for genotyped animals** computed either approximate or real inverse (**TABLE 1**)
- **4,536 broiler chickens**: Several rounds of approximation of inverse of $\mathbf{A}_{22}$, same criterions as for dairy bulls and **measure of elapsed CPU time for inversion** (**TABLE 2**)

TABLE 1: Approximation of inverse of G on 5 rounds and correlation (*r*) between solutions for genotyped animals

Round	p	MSD[1]	% of 0[2]	r
1	0.21	$66.53 * 10^{-4}$	92.35	0.79
2	0.017	$34.87 * 10^{-4}$	10.78	0.91
3	0.009	$16.81 * 10^{-4}$	2.28	0.97
4	0.005	$5.97 * 10^{-4}$	0.86	0.99
5	0.003	$1.82 * 10^{-4}$	0.49	1.00

TABLE 2: Approximation of inverse of $\mathbf{A}_{22}$ on 2 rounds, correlation (*r*) between solutions for genotyped animals and elapsed CPU time for inversion

Round	p	MSD[1]	% of 0[2]	r	CPU time[3] (s)
1	0.13	$1.63 * 10^{-5}$	98.52	0.77	170.27
2	0.005	$2.92 * 10^{-6}$	96.19	>0.99	371.47

[1]Stands for "Mean Square Difference", i.e. averaged square difference between elements of real and approximated inverses
[2] Refers to the proportion (in %) of non-zeros elements in the lower triangular part of the final T^{*-1}
[3]CPU times have to be compared with elapsed CPU time (1805 s) for real inversion of $\mathbf{A}_{22}$ using a non-optimized gaussian elimination algorithm

CONCLUSIONS: A new insight

- For both $\mathbf{A}_{22}$ and **G**, the algorithm tends to **provide real inverse: decreasing MSD** and **increasing *r*** show that approximated inverse **comes closer to real inverse throughout rounds**
- As **the triangular decomposition stays highly sparse for** $\mathbf{A}_{22}$, the algorithm is **suitable for inversion** of this matrix
- For the case of **G**, this algorithm brings a new insight by computing a covariance matrix of MME for genetic random effects that **blends both genomic and pedigree** informations

ACKNOWLEDGEMENTS
This research is part of the AFR PhD grant PHD-09-118 (Fonds National de la Recherche, Luxembourg) project NextGenGES done under the Private-Public-Partnership agreement between CONVIS (Ettelbruck, Luxembourg) and ULg-GxABT (formerly FUSAGx). Nicolas Gengler (Honorary Senior Research Associate) acknowledges the National Fund for Scientific Research (FNRS) for the support. The authors gratefully acknowledge support by the Holstein Association USA Inc. (Brattleboro, VT), the Cooperative Dairy DNA Repository (Beltsville, MD) and Cobb-Vantress (Siloam Springs, AR).

Contact: pierre.faux@ulg.ac.be

Inverse of large relationship matrices