BLUPF90 suite of programs for animal breeding with focus on genomics

I. Aguilar[1], S. Tsuruta[2], Y. Masuda[2], D.A.L. Lourenco[2], A. Legarra[3] & I. Misztal[2]

[1] Instituto Nacional de Investigación Agropecuaria, Ruta 48 km10, 90200, Canelones, Uruguay
iaguilar@inia.org.uy (Corresponding Author)
[2] University of Georgia, Department of Animal and Dairy Science, 30602 Athens, GA, USA
[3] Institut National de la Recherche Agronomique, UMR1388 GenPhySE, Castanet Tolosan, France 31326

Summary

The BLUPF90 suite is a collection of software for mixed-model analysis with focus on breeding and genetics applications. Solving of mixed model equations and variance component estimation are supported for general multiple trait, multiple effect models, with different model design per trait and correlated random effects. Genomic analyses using single-step GBLUP are fully integrated in all programs with efficient optimizations for large scale genetic evaluations. The state of the art of the BLUPF90 suite with a focus on genomic prediction using single-step genomic BLUP is presented.

Keywords: genetic evaluation, variance components, genomic prediction, software

Introduction

The BLUPF90 suite is a collection of software for mixed-model computations with a focus on breeding and genetics applications. It was originally developed to support a Fortran 90/95 programming course in computational techniques in animal breeding that was taught at the University of Georgia in 1999 by I. Misztal. Programming examples from that course lead to an idea of a general yet simple BLUP program, to calculate solutions of mixed model equations, that could be easily modified to accommodate and test new methodologies in animal breeding. The resulting program was called BLUPF90 and supports general multiple trait, multiple effect models, with different model design per trait, allows missing traits, several correlated random effects, such as direct and maternal genetic effects, random regression models, dominance effects and flexibility to handle several pedigree files or different covariance structures defined by the user. A general description of the software and its philosophy was presented previously (Misztal 1999).

The original BLUPF90 program evolved to allow the estimation of variance components (REML, Gibbs sampler), support for threshold models, computations of solutions and approximations of accuracy for large scale genetic evaluations, but instead of creating a single big program, several programs were developed that are generally known as the BLUPF90 suite of programs (Misztal *et al.,* 2002).

The original simplicity and flexibility of the BLUPF90 program allowed a simple but efficient algorithm program to be incorporated to handle genomic information, known as single-step genomic BLUP (ssGBLUP) (Aguilar *et al.,* 2010). This uses simple modifications to incorporate new relationship matrices and nearly all models supported by programs of the BLUPF90 suite became ready to incorporate genomic information (Aguilar *et al.,* 2011).

The objective of this work is to present the state of the art of the BLUPF90 suite with focus in genomic prediction using single-step genomic BLUP.

Description of programs

A full picture of some of the available programs is presented in Figure 1.

Data preparation

RENUMF90, on the top of the diagram, uses a specific parameter file that reads data (records, pedigree and genotypes, possibly in alphanumeric format) and prepares a renumbered data file, a renumbered pedigree, a file with cross-references to the genotype file and the parameter file to be used in the remaining programs (variance component and BLUP predictions). Some of the current features of RENUMF90 include trace back of a given number of generations of animals with either phenotype and/or genotype information, inclusion of inbreeding in the pedigree file to create the inverse of the numerator relationship matrix accounting for inbreeding, handling of unknown parent groups, merging of effects (e.g. herd-year-season). It also generates numerical statistics of data, cross-references (i.e. alphanumeric identities to renumbered levels) and tables with frequencies for each effect.

Once data is renumbered, all BLUPF90 programs are controlled by a unique parameter file that specifies: the input data file; model design; (optionally) covariance structures: pedigree, genotype or user-defined files; and (initial) values of variance components for BLUP or variance component estimation. Extra options to control specific features of each program can be added by optional parameters.

BLUPF90

Solving of mixed model equations to get BLUPs and BLUEs was the original purpose of the BLUPF90 programs. These are stored in memory and solutions can be obtained by direct inversion or by iterative methods. The default method is the preconditioned conjugate gradient algorithm (Tsuruta *et al.,* 2001). Pedigree additive relationships (with or without unknown parent groups) are supported, and these can be combined with genomic information by adding extra relationship matrices following the theory of ssGBLUP (Aguilar *et al.,* 2010, Christenssen and Lund, 2010). Other (co)variance structures can be defined including: parental dominance, metafounders (Legarra *et al.,* 2015), relationships derived for honey bee production (Bienefeld *et al.,* 2007) and also user defined matrices, e.g. for genomic epistasis (Vitezica *et al.,* 2017) or additive-dominance with inbreeding (Fernández *et al.,* 2017) or an user defined matrix.

Prediction error (co)variances can be obtained from the sparse inverse of the MME, to derive accuracy for animal models or for random regression models. BLUPF90 supports heterogeneous residual variances as used in GIBBS3F90 where the user can supply residual variance components as a file.

Variance components

Programs to estimate variances components implement REML and Gibbs sampling methods (Thompson *et al.,* 2005). There are two programs available for REML: one that implements an EM-REML algorithm (REMLF90) which is very reliable but slow to converge and another that implements the Average Information REML (AIREMF90) which is much faster. In the case of the AIREMLF90, standard errors of any function of variance components (heritability, genetic correlations) can be obtained following Meyer and Houle

(2013). A model with support for continuous heterogeneous residuals is implemented in AIREMLF90 following Druet *et al.,* 2003. Using optimized techniques for sparse matrices operations implemented in YAMS (see below) decreases computing time by one order of magnitude.

Gibbs sampling programs decrease memory requirements for estimation of variance components in comparison with REML programs. They are highly optimized for storage of mixed model equations and for block sampling for multiple trait (GIBBS1F90), and multiple traits and random correlated effects (GIBBS2F90), and heterogeneous residual variances defined by classes (GIBBS3F90). The program THRGIBBS1F90 estimates variance components for multiple trait threshold-linear traits, and THRGIBBS3F90 adds support for heterogeneous residual variances.

Gibbs sampling programs are capable of any number of nonzero elements in MME as long as the memory is available, which it is beneficial for genomic models. In the current implementations of REMLF90 and AIREMLF90 this is a limitation, but it will be removed in the near future.

Large scale analyses

BLUP90IOD is used to compute solutions for large scale genetic evaluations that use iteration on data with the preconditioner conjugate algorithm solver (Tsuruta *et al.,* 2001). Modified versions provide support for heterogeneous residual variance, multiple breed evaluation (Legarra *et al.,* 2007), optimized preconditioners for random regression models (Aguilar *et al.,* 2010), and threshold-linear models.

ssGBLUP genomic evaluations were originally implemented storing $\mathbf{G}^{-1}$-$\mathbf{A}_{22}^{-1}$ in core (Aguilar *et al.,* 2010; Aguilar *et al.,* 2011), however for large number of genotyped individuals, i.e. more than 150,000 (Aguilar *et al.,* 2014) the computing time was a limiting factor. The APY method presented by Misztal *et al.,* (2014) which computes a sparse inverse of $\mathbf{G}^{-1}$ based on a core set of animals, coupled with an efficient sparse implementation of $\mathbf{A}_{22}^{-1}$ (Masuda *et al.,* 2016a) is now successfully implemented for national genetic evaluations with large number of genotyped animals (Lourenco *et al.,* 2015; Masuda *et al.,* 2016b). A complete formula for unknown parent groups in ssGBLUP using QP-transformation (Misztal *et al.,* 2013) is also implemented in BLUP90IOD.

Genomic information

PREGSF90 is an interface to process the genomic information for the BLUPF90 family of programs. Originally developed to help the implementation of ssGBLUP, it implements a set of quality control on genotypes, and provides several outputs to detect possible errors with genotypes (Aguilar *et al.,* 2014). Using POSTGSF90 solutions from ssGBLUP are used to backsolve estimates for SNP effects, which can be used to predict interim direct genomic values for newly genotyped individuals using PREDF90.

Sparse-dense matrix efficient methods

A key feature of the BLUF90 programs was a sparse matrix module that allowed efficient programming of sparse matrix computations. Sparse matrix operations use the module FSPAKF90 (Misztal & Perez-Enciso 1998), an interface to FSPAK (Perez-Enciso *et al.,* 1994). For sparse matrices, FSPAK is very efficient. With the incorporation of genomic information with single-step GBLUP, large blocks of dense matrices deteriorate the performance of the FSPAK subroutines. Masuda *et al.,* (2015) implemented a new module in

the BLUPF90 suite (YAMS) that detects such dense blocks in the mixed model equations and rearrange computations using dense operations with optimized and parallelized subroutines. This reduces drastically the computing time for variance component estimation using REML, or to get exact accuracies by inversion.

Availability

All but the large-data programs are available online as well as full wiki page with documentation for each program and examples (http://nce.ads.uga.edu/software/). The programs are free for research use but their use should be acknowledged in publication.

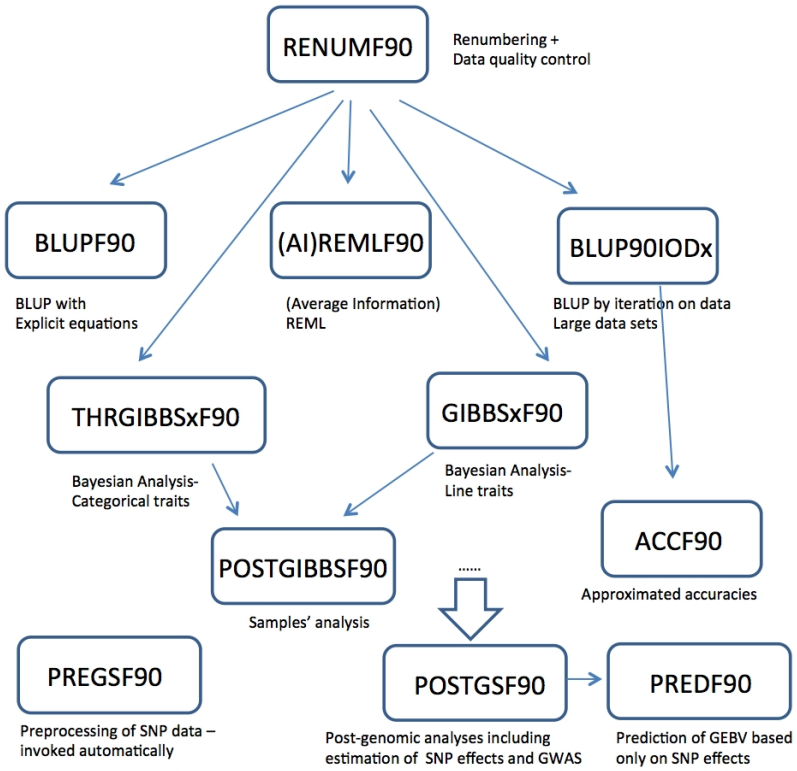

Figure 1. Main programs of BLUPF90 suite and their relationship

List of References

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., Lawlor, T. J. (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score J. Dairy Sci. 93: 743–752

Aguilar, I., Misztal, I., Legarra, A., Tsuruta, S. (2011). Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation J Anim Breed Genet 128: 422-8

Aguilar, I., Misztal, I., Tsuruta, S., Legarra, A., Wang, H. (2014). PREGSF90 – POSTGSF90: Computational Tools for the Implementation of Single-step Genomic Selection and Genome-wide Association with Ungenotyped Individuals in BLUPF90 Programs. Proc. 10th World Cong. Gen. Appl. Livest. Prod.. Vancouver, Canada.

Aguilar, I., Tsuruta, S., Misztal, I. (2010). Computing options for multiple-trait test-day

random regression models while accounting for heat tolerance J. Anim. Breed. Genet. 127: 235–241

Bienefeld, K., Ehrhardt, K., Reinhardt, F. (2007). Genetic evaluation in the honey bee considering queen and worker effects – A BLUP-Animal Model approach Apidologie 38: 77-85

Christensen, O. and M. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42:2.

Druet, T., Jaffrézic, F., Boichard, D., Ducrocq, V. (2003). Modeling Lactation Curves and Estimation of Genetic Parameters for First Lactation Test-Day Records of French Holstein Cows J. Dairy Sci. 86: 2480–2490

Fernández, E. N., Legarra, A., Martínez, R., Sánchez, J. P., Baselga, M. (2017). Pedigree-based estimation of covariance between dominance deviations and additive genetic effects in closed rabbit lines considering inbreeding and using a computationally simpler equivalent model J. Anim. Breed. Genet. 134: 184-195

Legarra, A., Bertrand, J. K., Strabel, T., Sapp, R. L., Sanchez, J. P., Misztal, I. (2007). Multi-breed genetic evaluation in a Gelbvieh population J. Anim. Breed. Genet. 124: 286–295

Legarra, A., Christensen, O. F., Vitezica, Z. G., Aguilar, I., Misztal, I. (2015). Ancestral Relationships Using Metafounders: Finite Ancestral Populations and Across Population Relationships Genetics 200: 455-468

Lourenco, D. A. L., Tsuruta, S., Fragomeni, B. O., Masuda, Y., Aguilar, I., Legarra, A., Bertrand, J. K., Amen, T. S., Wang, L., Moser, D. W., Misztal, I. (2015). Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus J. Anim. Sci. 93: 2653-2662

Masuda, Y., Aguilar, I., Tsuruta, S., Misztal, I. (2015). Technical note: Acceleration of sparse operations for average-information REML analyses with supernodal methods and sparse-storage refinements J. Anim. Sci. 93: 4670-4674

Masuda, Y., Misztal, I., Legarra, A., Tsuruta, S., Lourenco, D. A. L., Fragomeni, B. O., Aguilar, I. (2016a). Technical note: Avoiding the direct inversion of the numerator relationship matrix for genotyped animals in single-step genomic best linear unbiased prediction solved with the preconditioned conjugate gradient1 J. Anim. Sci.

Masuda, Y., Misztal, I., Tsuruta, S., Legarra, A., Aguilar, I., Lourenco, D. A. L., Fragomeni, B. O., Lawlor, T. J. (2016b). Implementation of genomic recursions in single-step genomic best linear unbiased predictor for US Holsteins with a large number of genotyped animals J. Dairy Sci. 99: 1968-1974

Meyer, K., Houle, D. (2013). Sampling based approximation of confidence intervals for functions of genetic covariance matrices Proc. Assoc. Advmt. Anim. Breed. Genet 20: 523-526

Misztal, I. (1999). Complex models, more data: Simpler programming? Interbull Bull.: 33

Misztal, I., Legarra, A., Aguilar, I. (2014). Using recursion to compute the inverse of the genomic relationship matrix J. Dairy Sci. 97: 3943-3952

Misztal, I., Perez-Enciso, M. (1998). FSPAK90 - A Fortran90 interface to sparse-matrix package FPSPAK with dynamic memory allocation and sparse matrix structure Proc. 6th World Cong. Gen. Appl. Livest. Prod. 27: 467-468

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T., Lee, D. H. (2002). Proc 7th World Congress on Genetics Applied to Livestock Production. Montpellier, France.

Misztal, I., Vitezica, Z. G., Legarra, A., Aguilar, I., Swan, A. A. (2013). Unknown-parent groups in single-step genomic evaluation J. Anim. Breed. Genet. 130: 252-258

Perez-Enciso, M., Misztal, I., Elzo, M. A. (1994). FSPAK : An interface for public domain sparse matrix subroutines 5th World Congress on Gen. Appl. to Livest. Prod 22: 87-88

Thompson, R., Brotherstone, S., White, I. M. S. (2005). Estimation of quantitative genetic

parameters. Philos Trans R Soc Lond B Biol Sci. 360:1469-1477.
Tsuruta, S., Misztal, I., Stranden, I. (2001). Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed-model equations in animal breeding applications J. Anim. Sci. 79: 1166–1172
Vitezica, Z. G., Legarra, A., Toro, M. A., Varona, L. (2017). Orthogonal Estimates of Variances for Additive, Dominance and Epistatic Effects in Populations Genetics 206: 1297-1307