PEDIG : A FORTRAN PACKAGE FOR PEDIGREE ANALYSIS SUITED FOR LARGE POPULATIONS

D. Boichard

Institut National de la Recherche Agronomique,
Station de Génétique Quantitative et Appliquée, 78352 Jouy en Josas Cedex, France

This fortran 77 software, suited to the analysis of very large populations (up to several tens of million individuals), is a set of independent programs to calculate probabilities of gene origin, relationship and inbreeding coefficients, and to characterize the quality of pedigree information.

GENERAL PRESENTATION OF THE PROGRAMS

Inbreeding. Two programs are available to compute inbreeding coefficients of all individuals in a population. The first one, *meuw.f*, uses the method described by Meuwissen and Luo (1992). This method is very fast, at least for pedigrees with less than 15 generations. A second program, *vanrad.f*, relies upon VanRaden's method (1992), derived from the tabular method. It is about three times slower than *meuw.f* but it makes it possible to compute inbreeding coefficients for a given number of generations considered and, therefore, to distinguish close from far inbreeding. It can also assume inbred or related founders, which could be very useful when the pedigree information is heterogeneous.

Probabilities of gene origin. The program *prob_orig.f* computes the probabilities of gene origin for a reference population, *i.e.* a group of individuals defined for instance according to their sex and birth period. These probabilities are combined altogether to estimate an effective number of founders. The output file provides the non zero contributions of each ancestor and each founder. This program also iteratively detects the ancestors with the greatest marginal contribution, as defined by Boichard *et al.* (1997), and combines these contributions into an effective number of ancestors.

The program *segreg.f* simulates the segregation of founder alleles in the population, estimates founder gene frequencies by gene counting in the reference population and combines these frequencies to estimate the effective number of remaining genomes in the reference population. The program *etr.f* computes the contributions of the different origins of a crossbred population, as well as their evolution over time. It uses a classification of the founders in the different origins provided in the pedigree file. Default values of genetic composition could also be defined according to sex and birth year. An output file provides the genetic composition of each individual.

Relationship Coefficients Four programs are provided, either to compute a relationship matrix or to get an estimate of the average relationship level in a population. The program *parente.f* computes the complete relationship matrix for one or two small groups of individuals (few thousands) with the tabular method. If two groups are defined, the output distinguishes the within and between groups relationship matrices.

To estimate the average relationship level in a large population, the program *par.f* computes relationship coefficients on repeated random samples of either males, females, complete herds, or couples.

The program *par2.f* estimates the average relationship between a set of individuals on one hand, and a reference group on the other hand. The list of individuals of interest is provided in a second file, in addition to the pedigree file. The population of reference could be defined either from the pedigree file by selecting a subgroup according to its sex and birth year, or in the third file. A progeny is generated from each mating and its inbreeding coefficient is estimated by Meuwissen's method.

Characterization of the pedigree information. The program *ngen.f* computes the average number of known ancestors, the percentage of known ancestors per generation, and the number of equivalent complete generations traced. These statistics are provided by sex and birth year. The percentage of known ancestors at the first 8 generations is stored in an output file for each individual. The program *intgen.f* computes generation intervals by sex and birth year of the progeny. It uses the birth year information in the pedigree file.

USER'S GUIDE

All these programs read the same pedigree file, which could contain the following information, even if not all these informations are used by all programs : individual identification, from 1 to N; sire and dam Id; birth year; sex; group of origin; herd Id. It is advised to prepare one unique file including these 7 informations in this order. Such a complete file could be used by all programs. However, if one is interested only to inbreeding coefficients (for instance), only the first three informations are necessary.

Although all algorithms used require a preliminary sort of individuals in such way that parents appear before progeny, this constraint is not necessary in the pedigree file. Indeed, all programs perform this ordering internally, and this step is very fast. In contrast, original Id numbers are not supported and individuals should be sequentially numbered. If needed, this recoding step from original Id to sequential number could be done by the *ped_util.f* program.

Outputs are either in English or in French, according to a compilation option.

AVAILABILITY

Pedig (standard Fortran 77 sources, user's guide in French and English, examples) can be freely downloaded from http://www-sgqa.jouy.inra.fr/diffusions.htm.

REFERENCES

Boichard D., Maignel L., Verrier E. (1997) *Genet. Sel. Evol* **29** : 5-23.
Meuwissen T.H.E., Luo Z. (1992) *Genet. Sel. Evol*. **24** : 305-313.
VanRaden P.M. (1992) *J. Dairy Sci*. **75** : 3136-3144.