The POPREP Web Pipeline: Monitoring Animal Genetic Resources

E. Groeneveld[*] and H. Lichtenberg[*]

Background

Conservation of breeds under the threat of extinction has long been an issue of importance. In the wake of the Rio Convention on Biological Diversity (1992), the focus has expanded to within breed diversity of small but also large populations. As they are responsible for much of the food production their deterioration would have a wide spread negative effect. Therefore, the National Action Program on Animal Genetic Resources in Germany stipulates continuous monitoring of active breeding populations.

The objective of this contribution is to develop a software system that produces a generic report based on a minimum set of individual animal information. The intended use is three-fold: firstly, it can serve as a documentation of parameters relevant to biodiversity issues to be included in the yearly reporting on breeds as proposed in Groeneveld (2003). Secondly, its outputs can be used as an early warning system in the management of big and small populations allowing to counteract negative trends as soon as they become apparent (Groeneveld (2010)). Finally, intermediate output will be made available for further research.

Implementation

POPREP is a web application (http://poprep.tzv.fal.de) operating on the principle of uploading data to the web server where all computations are done and returning the results via email. The software is available under the GNU General Public License (Free Software Foundation, Inc (1991)), but the availability of the POPREP web server makes installation on the user's side superfluous and provides a very low entry threshold.

Data inputs

Minimum requirements for the input data set are: unique identification of all animals, for each animal the sire, dam, birthdate and sex needs to be given. Original animal IDs can be used which may contain alpha numeric characters. The data must be in ASCII format with five pipe '|' delimited columns.

In the web interface, the email address needs to be specified to which the results will be returned, as well as information about the coding of sex and the date format. Name of breed, user name and affiliation are optional and mainly help to beautify the final reports. Finally, the pedigree file is uploaded and the user informed that the job has entered the execution queue.

[*]Institute of Farm Animal Genetics (FLI), Höltystr. 10, 31535 Neustadt/Mariensee, Germany

The resulting email contains two reports plus an optional archive with intermediate data used in the report for the user's post processing.

Web server pipeline

The POPREP web service is an automatic operation without any human intervention. The series of computational steps are executed by different subsystems. The nine major steps of the pipeline are:

STEP 1: input data upload After filling the POPREP web page with sufficient information, the datafile is uploaded and the cycle of steps 2 to 9 is triggered.

STEP 2: input data verification Input verification is done as a first step before a dataset is processed any further. After checking the input file format (ASCII, five pipe delimited data columns, date format, sex encoding), the consistency of the pedigree is checked. These are: parent older than offspring, animal can show up only as sire or dam depending on its own sex and pedigree loops. The latter are particularly difficult to find. If a loop is found, a graphical presentation of the animals in that loop is generated and provided as a pdf document to the user. If major errors occur, the pipeline is stopped and the user informed accordingly via email. After correction the input data can be uploaded to start the process again.

STEP 3: database setup and loading of data For each run initiated by a data upload a new APIIS (Groeneveld (2004)) project is created. All computations are done on these pedigrees being loaded into a PostgreSQL database.

STEP 4: SQL runs and creation of auxiliary tables All data used for the Population Structure Report and Inbreeding Report are generated through SQL commands and stored in auxiliary database tables which are at a later stage accessed for the report generation.

STEP 5: parallelization of compute-intensive tasks The basic layout of the reports considers birth years of animals as a group, computing statistics for each of them, thereby allowing the user to observe the development of the breed over time. Some of the calculations are compute intensive, mainly the pedigree completeness index (PCI) and the additive genetic relationship (AGR). But as the birth year statistics are independent from each other, these processes are executed in parallel, starting as many jobs as CPUs are available on the web server. Currently, our web server has 2 quadcores cutting the computational time into 1/8 for the PCI and AGR computations.

STEP 6: assembly and typesetting of reports The input part of the reports is made up from 2 components: the numerical data computed on the basis of the submitted pedigree and text blocks that are part of the POPREP package. Both are merged and typeset with LaTeX to yield high quality camera ready reports. They are self contained documents, presenting a fully fledged status of the population dynamics of the breeds under investigation with texts explaining the tables and graphs generated from the data. Typically, they are more than 20 pages each with a total of 14 tables and 13 figures.

STEP 7: assembly of archive of computed data It was the intention to automatically produce a camera ready typeset report on population structure and inbreeding related issues from a pedigree file. With no human intervention all data submitted in the pedigree file will be used in the computation of the statistics. Often it may be desirable to analyze and present subsets of data given in the reports. This is facilitated by the optional provision of intermediate results as part of the mail returned. For each table and graphic in the reports, the underlying numerical data is available in a CSV file. Furthermore, inbreeding coefficients are provided for each animal submitted. All of these 35 files are assembled into a compressed archive.

STEP 8: mail delivery of results The resulting reports, a basic statistic on the input data and data checks and, if requested, the archive are mailed back to the user. The pdf reports amount to less than 2MB in size while the results archive is largely a function of the input data and therefore may be substantially larger.

STEP 9: removal of data and results Each upload results in a new, encapsulated project created on the web server with its own database and filesystem hierarchy. The last step in the pipeline is the purging of the project with only the contact information given by the user in the home page of POPREP being saved for statistical evaluation.

Results

POPREP has been in automatic operation since September 2009. Although format requirements of the pedigree input data are very simple, the initial data checking had to be tightened because of substantial problems in data preparation. It is not uncommon, that further processing has to be aborted due to errors in data formatting, inconsistencies in birthdate and sex. Furthermore, loops in pedigrees are also a reason for early stops. Depending on the data quality, typically a number of uploads, data checking, data correction and again upload cycles have to be gone through before a complete report can be generated.

Report printing Being delivered as pdf files the main reports can be printed right away on the local printer of the user. For those who would like to modify the typeset report the LaTeX text is supplied as part of the archive. Thus, any changes can be made to suit special needs.

Post processing The archive with the numerical data can be analysed by the user at her own digression. While the tables of the typeset reports contain all years in the data, the user may wish to present only the last 10 years. Then the file corresponding to the report table can be imported into a spreadsheet with a free choice of lines (i.e. years) to be included in the customized presentation.

Research on effective population sizes require a number of decisions to be taken. Groeneveld et al. (2009) have presented six methods to compute the effective population size. Additionally, a number of decisions need to be taken like starting and end year involved, and the degree of pedigree completeness. Estimates for some of the methods can be taken directly from the report while others can be computed on the basis of the spreadsheet data.

Discussion and conclusion

Dissemination of executable programs is straight forward: a program needs to be transferred and executed on the target platform. With the development of complex systems like POPREP the situation is very different: here a number of software components interact, each of which may have to be installed and configured before use. To name but a few, POPREP uses PostgreSQL, LaTeX, gnuplot, Fortran binaries and a number of Perl modules. Even if the source code is available, many users would find it overwhelming to make a complete installation on their own.

There are two options to overcome this issue: the complete system is made available as an appliance or its functionality is offered as a service on the web. An appliance is a large file containing the complete operating system and the application software fully configured and operational. This route is chosen for the deployment of MolabIS (Truong Van Chi and Groeneveld (2010)) and CryoWEB (Duchev et al. (2010)).

The web service alternative chosen here reverts the long standing paradigm: instead of bringing the program to the data the user sends the data to the software. This setup is very efficient for well defined in- and outputs like in pedigree analysis. The largest dataset run on the POPREP web site till now had 2.3 Mio. records. Execution time depends on the pedigree depth and the number of animals ranging from typically minutes to hours and even days.

However, a sufficiently fast Internet connection is required to be able to upload the input data and an email connection to receive the results. That these conditions are met already in many parts of the world, show the use statistics of the POPREP web site for the first 5 months of operation beginning in September 2009. Accesses from 32 countries were recorded with serious use from around 20 with about 350 projects executed. It is interesting to see the globalizing effect of the Internet: the first users came from Malaysia and Columbia.

References

Cong, T. V. C. and Groeneveld, E. (2010). In *Book of abstracts. 9th WCGALP, Leipzig, August 1 to 6, Germany*.

Duchev, Z. I., Gandini, G., Berger, B., et al. (2010). In *Book of abstracts. 9th WCGALP, Leipzig, August 1 to 6, Germany*.

Free Software Foundation, Inc (1991). 59 Temple Place - Suite 330, Boston, MA 02111-1307, USA.

Groeneveld, E. (2003). *Züchtungskunde*, 75(5):317–323.

Groeneveld, E. (2004). *Livestock Production Science*, 87:1–12.

Groeneveld, E. (2010). *Züchtungskunde*, 82(1):29–39.

Groeneveld, E., v.d. Westhuizen, B., Maiwashe, A., et al. (2009). *Genetics and Molecular Research*, 8(3):1158–1178.