Integrated Development of the Software with Literate Programming: An Example in Plant Breeding

X. Mi , H.F. Utz, A.E. Melchinger*

Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, Germany

* Contact author: melchinger@pz.uni-hohenheim.de

Keywords: Multi-stage Selection, QTL, Resources Optimization, Portfolio Analysis, Literate Programming

The resource is always a limiting factor in a plant breeding program. The breeders try to optimize the allocation of limited resources, such as time, money, locations and so on, in order to enhance genetic gain. Furthermore, there is a continuous information explosion, in the present era: (1) generation of huge amount of information and data, such as molecular markers, biometric data of various genotypes in different environments, cross validation and meta analysis outputs and (2) new algorithms, new generations of computers, software and data bases developed for solving our problems and storing the data and information. To meet the challenges, large multi-inter-discipline scientific teams have to operate together to generate huge amount of data sets, manage the data sets and undertake the required analysis.

In this situation, an efficient software management technology is required: (1) to handle large data sets quickly, (2) should be well documented, (3) should be easily implementable and amenable to the incorporation with new features, such as C-port and object oriented, (4) should be popular with plenty users and (5) should be easily interpretable by the partners. We choose the management software Sweave and StatWeave with literate programming technology and this technology generates programs like writing a scientific paper, to build our package *PlabPortfolio*. This package is applied in plant/animal breeding for maximizing the gain of a multi-stage selection procedure under certain restrictions (e.g. for a given annual budget or certain risk limits at each stage). By implementing the R-package *mvtnorm*, which is one of the core packages of R and calculates the multi-variate normal distribution, the number of independent variables in the multi-normal regression model is increased from 3 to 1000. This makes it possible to use huge amounts of marker and QTL information. A variance analysis function, which achieves an efficient distribution of the budget between the QTL tests and field trials, will be implemented into the project.

References

Tallis, G. M. (1961). Moment generating function of truncated multi-normal distribution. *Journal Of The Royal Statistical Society Series B-Statistical Methodology*, 23(1):223

Utz, F. (1969). Mehrstufenselecktion in der Panzenzuechtung. *PhD thesis*, University Hohenheim.

Knuth, Donald E. (1992). Literate Programming. California: Stanford University Center for the Study of Language and Information. ISBN 978-0937073803.

Hothorn, T., Bretz, F., and Genz, A. (2001). On multivariate t and Gauss probabilities in R. *R News*, 1(2):27-29.

Mi, X., Miwa, T. and Hothorn, T. (2009). Implement of Miwa's analytical algorithm of multi-normal distribution *R News*, accepted.

Mi, X. (2008). Model Selection Procedure with Familywise Error Rate Control for Binomial Order-Restricted Problems. *PhD thesis*, University Hannover.