The use of R as part of a large-scale information management and decision system

Joris DE WOLF, Koen BRUYNSEELS, Rindert PEERBOLTE and Willem BROEKAERT
CropDesign

February 28, 2006

This presentation describes the successful use of R as a part of an information system in a industrial setting.

An information management and decision system has been developed for TraitMill^{TM}, a highly automated plant evaluation platform allowing high-throughput testing of the effect of the introduction of transgenes on agronomically valuable traits in crop plants. The screening is based on plants grown in-greenhouse in specific experimental layouts, imaged at weekly intervals, and harvested on an individual plant basis. About one hundred thousand plants are screened annually. The measurements are automated to a large extent and a vast amount of data is stored directly in a central relational database. This database is used to manage the information flows but also gives input to a decision support system that assists in detecting interesting genes in the test population.

The relational database is built in Microsoft SQL Server and accessed through a Java fat client or a web based front-end, all running on a Linux platform. Additional components were needed to carry out formal statistical analysis and inference as well as for producing graphs. These components had the following requirements: (i) be highly versatile and programmable in-house, (ii) be able to perform complex statistical analysis (linear mixed models, survival analysis, non-linear curve fitting among others) in a reliable and automated manner, and (iv) exchange data and results with the database swiftly and reliably. The graphic component had to (v) produce graphs for integration in websites and for high-quality publication.

R has been chosen to perform both the analytical and graphical tasks. It fulfilled all requirements mentioned above. For the core of the statistical analysis, off-the-shelf R packages and functions proved to be sufficient and performed these tasks swiftly enough. The accessibility of the code and the relative simplicity of the language made the development of scripts for specific goals straightforward. The only low-end adaptation that needed to be developed was the connection between R and MS SQL Server. The latter functionality has been put in the open source domain by CropDesign.

Currently R is used in two ways in this system. First, batches of scripts are run unsupervised in the background, using data from the database and storing results back into the database and graphical output into file systems. Second, R is accessed interactively by the user via the Java interface. For the latter Rserve is used.

A drawback of R is the high turnover of new releases and the problems or suspicion of backward incompatibility that this may bring along.

Despite this downside, R has proven in the last three years of high-througput operation of TraitMill^{TM} to be a valuable ressource for information management.