Using Data Warehouse Technology in Crop Plant Bioinformatics

Christian Kuenne, Ivo Grosse, Inge Matthies, Uwe Scholz, Tatjana Sretenovic-Rajicic, Nils Stein, Andreas Stephanik, Burkhard Steuernagel and Stephan Weise*

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Corrensstr. 3, 06466 Gatersleben, Germany

Summary

Plant-specific data is managed in heterogeneous formats and is dispersed geographically. Based on this data, efficient analyses require a materialised integration, often realised with data warehouse technology today. We describe the requirements, problems and solution strategies for domain-crossing integration as the fundament for analysing plant biological data based on three current case studies. First, we introduce a system for retrieval of markers and mapping positions based on clustering of ESTs. The second case study illustrates the steps for diversity studies after genotyping a collection of about 3,000 ryegrass accessions (*Lolium* spp.), whereas in the third example data of approximately 250 barley cultivars (*Hordeum vulgare*) were used for associating haplotype- and SNP-patterns with malting parameters. For all case studies, we integrate data from different domains - sequence and marker data as well as IPK Genebank data including passport and phenotypic information. Specific problems associated with plant biological data and possible solution strategies are shown.

1 Introduction

Today, data amount is increasing enormously in plant biology. The use of modern high-throughput methods is more and more ruling out the traditional way of researching [1]. The scientific focus is moving away from the single-data-domain and problem-oriented approach towards work crossing the borders of data domains. Bioinformatic tools help to analyse data at a large scale. Often, extensive data sets gained from biological experiments cannot be handled individually, especially in the area of genomic data.

While the main focus of the scientific community is on human and several animals, such as fruit fly and mouse, plants are often under-represented. Several projects are working on a certain plant species and maintain relatively independent data sources, e. g. for barley or pea, designed for special application areas by using different technologies [2].
Very often, data about plant genotypes is quite rare and needs to be supplemented by data of related genotypes. We think, for integration and analysis of plant data, the issues of the heterogeneous and sparse data of plants should be addressed.

In this study, we use data from different domains (phenotypic data, marker data, sequence data and taxonomic data), which can be linked using abstractions of the plants as central objects. Therefore, so-called passport data is used. Passport data serves as identificator of genotypes. It comprises parameters such as accession number, habitat or scientific name. A possible linkage

*To whom correspondence should be addressed. Email: weise@ipk-gatersleben.de