Analysis of high-throughput plant image data with the information system IAP

Christian Klukas[1], Jean-Michel Pape[1], Alexander Entzian[1]

[1]Research Group Image Analysis,
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK),
Corrensstraße 3, OT Gatersleben, 06466 Stadt Seeland, Germany

Summary

This work presents a sophisticated information system, the Integrated Analysis Platform (IAP), an approach supporting large-scale image analysis for different species and imaging systems. In its current form, IAP supports the investigation of Maize, Barley and Arabidopsis plants based on images obtained in different spectra.

Several components of the IAP system, which are described in this work, cover the complete end-to-end pipeline, starting with the image transfer from the imaging infrastructure, (grid distributed) image analysis, data management for raw data and analysis results, to the automated generation of experiment reports.

1 Introduction

Recent advancements in the automation of plant imaging make it possible to investigate the phenotype of several hundred plants over time. It now becomes possible to analyze numerous phenotypic properties using non-invasive imaging systems obtained in different spectra. The most commonly used visible light imaging is used to track plant growth properties, near-infrared cameras give insight into the watering status of plants and fluorescence imaging systems make it possible to investigate the photosynthetic activity of plant organs such as leafs, plant stems and flowers. Finally yet importantly infrared cameras are used to track over time even tiny leaf temperature differences between different genotypes, varieties or treatments.

This work presents a sophisticated information system, the integrated analysis platform (IAP), an approach supporting large-scale image analysis for different species and imaging systems. In its current form, IAP supports the investigation of Maize, Barley and Arabidopsis plants using images from visible light, fluorescence and near-infrared cameras. For Arabidopsis also infrared images are utilized for phenotyping purposes.

Several components of the IAP system, which are described in this work, cover the complete end-to-end pipeline, starting with the image transfer from the imaging infrastructure, (grid distributed) image analysis, data management for raw data and analysis results, to the automated generation of experiment reports.

2 Methods

2.1 Data import from imaging systems

At several institutions around the world and also at the IPK-Gatersleben automated imaging solutions, sourced by the LemnaTec GmbH, Wuerselen, Germany, have been installed. Using the vendor-supplied import/export tool it is possible to manually export image data sets from

these systems. But this task is comparably time consuming as annotation data needs to be processed separately and such solution prevents the automation, desired for our approach and needed in an environment where frequent up-to-date analysis is required. For these reasons IAP contains a LemnaTec database access component which has been reverse-engineered from the existing and used PostgreSQL data structures. Relevant for the IAP data access component are the tables which contain measurement labels, plant annotations, image snapshot records with imaging times and configurations and those tables that specify the storage locations of the created image files. These image files are stored at the database server and are accessible using FTP and SFTP.

2.2 In-memory data management

For IAP the VANTED [1] data structures for experiment data have been slightly extended in order to support the storage of exact timings of measurement data. The previous system is limited to storing whole numbers and a time unit, e.g. "day 1". Within IAP in addition to this timing information, which is useful for data display and human interaction with the data, the exact date and time point is handled and processed, if available. In addition, each image and each numeric data point in the data set may be flagged, so that these data sets may be more easily tracked and checked or may be marked as an outlier, which yields to its omission during the image data analysis.

2.3 Database-based data management

In the previous systems DBE and VANTED, which were developed to handle extensive large-scale metabolite-, protein- and gene expression data sets [2] (example use case [3]), an Oracle database based component for shared data management has been developed. As the Oracle database requires special and complicated setup, handling and operation in addition to high licensing fees, for IAP it was desired to switch to an open source alternative. In order to decrease the development effort and to increase the turn-around time during database development, the schema-less so called No-SQL database MongoDB has been selected. Among the advantages of this system are native Java support libraries, the direct support for storing binary (image) files and the easy to implement mapping of Java objects and Java object hierarchies to the corresponding database structures. Other advantages are seamless support for sharding, a technique to distribute the database content onto multiple servers, which increases read-speeds and makes it possible to extend the storage and data management capabilities above the limits of a single database server.

2.4 Building blocks of image analysis pipelines

Traditionally different approaches are used to define image analysis pipelines. On one side graphical editors are used (e.g. within the LemnaGrid software), where pre-composed image analysis blocks are graphically placed on a drawing board and connected via lines, to model and define the input and output of these blocks. On the other hand a single image analysis method may be coded as source code, possibly created using structured procedural programming (e.g. in form of ImageJ macros as in HTPheno [4]). Similar to graphical programming, pre-defined analysis blocks and graphical editing quickly becomes unfeasible, once the complexity of the analysis pipeline reaches a certain point. For IAP multiple camera systems and the processing of images for multiple plant organisms need to be considered. To increase and ease code reuse, object oriented and component-based software development approaches have been used to define the image analysis pipelines.

2.5 Data export and report generation

The interpretation of the numerous calculated numeric properties is eased by the automated report generator, included within IAP. This report generator relies on the availability of an R installation and of an installed LaTex environment. Once the report functionality is selected, the user may specify how the data set should be separated or combined inside the experiment report diagrams. E.g. the user may be interested to get insight into developmental changes for different plant stress treatments, used within his biological experiment. IAP offers the separation and combination depending on multiple defined experiment factors (most importantly species, genotype, variety, and treatment). Depending on the specifics of the particular property either a box-plot, line diagram with displayed variance or a stacked bar plot for histogram data is generated. Finally, the pdflatex command is used to process a pre-defined LaTex file, containing headings, descriptions, and which also incorporates the generated diagrams. Only those properties and descriptions that are specific and available within the data set at hand are included inside the report.

3 Summary

The IAP system is implemented using the Java programming language and incorporates the ImageJ image manipulation library as well as the VANTED system [1] in order to handle and process image data as well as the results of the image analysis procedures. Input and output data sets as well as additional related data such as greenhouse climate data or metabolite measurement data can be loaded into the system, combined, filtered, stored within the MongoDB database and investigated or related to each other. This versatility can for example be used to import and process greenhouse temperature data and to calculate so called growing-degree days (GDD) for different experiments.

It is planned to further extend and improve the capabilities of the system, e.g. to relate different data domains and to incorporate and apply statistical methods to rank the results for different genetic lines or treatments within an experiment.

References

[1] C. Klukas and F. Schreiber: Integration of -omics data and networks for biomedical research with VANTED. *Journal of Integrative Bioinformatics*, 7(2):112, 2010.

[2] L. Borisjuk, M.-R. Hajirezaei, C. Klukas, H. Rolletschek and F. Schreiber: Integrating data from biological experiments into metabolic networks with the DBE information system. *In Silico Biology*, 5:0011, 2004.

[3] T. F. Sharbel, M. L. Voigt, J. M. Corral, G. Galla, J. Kumlehn, C. Klukas, F. Schreiber, H. Vogel, B. Rotter: Apomictic and Sexual Ovules of Boechera Display Heterochronic Global Gene Expression Patterns. *The Plant Cell*, 22(3):655-671, 2010.

[4] A. Hartmann, T. Czauderna, R. Hoffmann, N. Stein and F. Schreiber: HTPheno: An image analysis pipeline for high-throughput plant phenotyping. *BMC Bioinformatics*, 12:148, 2011.