Distributed Management of Scientific Workflows for High-Throughput Plant Phenotyping

Christophe Pradal, Sarah Cohen-Boulakia, Gaëtan Heidsieck, Esther Pacitti, Francois Tardieu, Patrick Valduriez

Distributed Management of Scientific Workflows for High-Throughput Plant Phenotyping

by Christophe Pradal (CIRAD), Sarah Cohen-Boulakia (Univ. Paris-Saclay), Gaetan Heidsieck (Inria), Esther Pacitti (Univ. Montpellier), François Tardieu (INRA) and Patrick Valduriez (Inria)

High-throughput phenotyping platforms allow acquisition of quantitative data on thousands of plants required for genetic analyses in well-controlled environmental conditions. However, analysing these massive datasets and reproducing computational experiments require the use of new computational infrastructure and algorithms to scale.

Plant species will need particular characteristics to survive in the world's changing climate. To this end, plant scientists are analysing traits of interest to identify both the natural genetic variations in plants and the genetic control of their responses to environmental cues.

In the last decade, high-throughput phenotyping platforms have allowed acquisition of quantitative data on thousands of plants required for genetic analyses in well-controlled environmental conditions. These platforms in controlled conditions are essential to understand the variability of yield in the field depending on climatic scenarios [3]. Recently, national infrastructures such as Phenome have used high-throughput platforms to observe the dynamic growth of a large number of plants under in field and platform conditions. These instrumented platforms produce huge datasets (thousands of images, environmental conditions and sensor outputs) with complex in-silico data analyses that result in the definition of new, complex variables [2]. The seven facilities of Phenome produce 200 terabytes of data annually, which are heterogeneous (images, time courses), multiscale (from the organ to the field) and originate from different sites. This infrastructure paves the way for the collection of data at an even larger scale. Indeed, farmers and breeders can increasingly capture huge amounts of diverse data, which at this stage is mostly environmental, but which can involve detailed maps of yield in fields (precision agriculture), and images originating from remote sensing and drone imaging. Hence, the major problem becomes the automatic analysis of these massive datasets and the ability to reproduce large and complex in-silico experiments.

Figure 1 The Phenoarch phenotyping platform is one of the Phenome node in Montpellier. It has a capacity of 1,500 plants per experiment with a controlled environment (e.g., temperature, humidity, irrigation) and automatic imaging through time (13 images per plant per day).

Such experiments can actually be represented and executed in an efficient and reproducible way by means of scientific workflows in which computational tasks can be chained (e.g., upload input files, preprocess the data, run various analyses and simulations, aggregate the results). OpenAlea is a scientific workflow system that provides methods and software for plant modelling at different scales [L1]. We have used it in the context of Phenome, by developing the OpenAlea Phenomenal software package dedicated to the analysis of 3-D plant architecture, whose outputs can be used for ecophysiological modelling. Phenomenal provides fully automatic workflows dedicated to the 3D reconstruction, segmentation and tracking of plant organs. It has been tested on maize, cotton, sorghum and apple trees. OpenAlea radiative models are used to estimate the light use efficiency and the in-silico crop performance in a large range of contexts.

Executing workflows on large datasets is time-consuming, and thus often incompatible with the scientist's way of working, where trial and error is an essential component. We have designed the infrastructure InfraPhenoGrid [1] to distribute the computation of workflows using the EGI/France Grilles computing facilities. EGI provides access to a grid with multiple sites, each with one or more clusters. This environment is now well suited for data-intensive computing, with different research groups collaborating at different sites. In this context, the goal is to address two critical issues in the management of plant phenotyping experiments: 1) scheduling distributed computation and

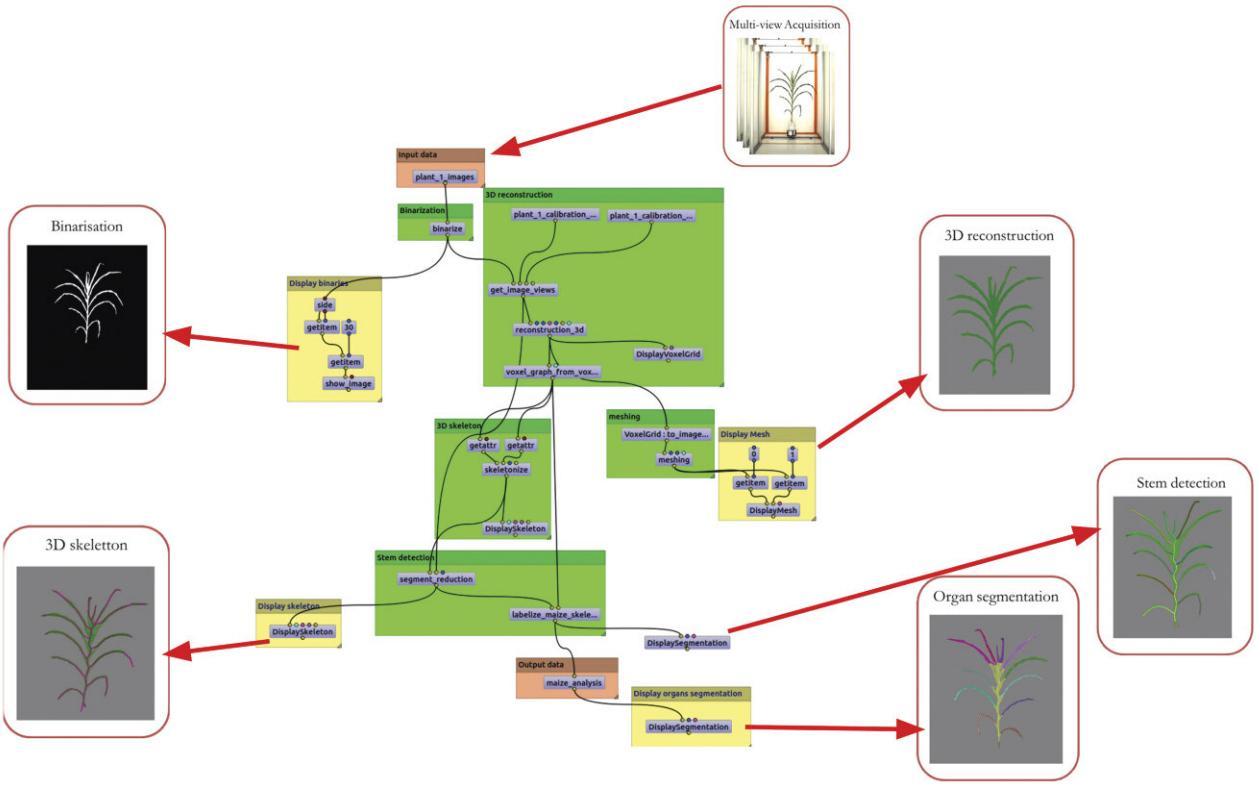

Figure 2: The 3D reconstruction workflow Phenomenal in the OpenAlea visual programming environment is applied to one plant at a given time. The same workflow is run on thousands of plants through time, consuming several terabytes of data, and distributed transparently on the European Grid infrastructure (EGI).

2) allowing reuse and reproducibility of experiments:

1.Scheduling distributed computation

We have adopted an algebraic approach suited to the optimisation and parallelisation of data-intensive scientific workflows. The scheduling problem resembles scientific workflow execution in a multisite cloud [2]. In the context of the #DigitAg project [L2], our objective is to propose new scalable and elastic heterogeneous scheduling algorithms that will transparently distribute the computation of these very large computational experiments on local servers, where the data are stored but with limited resources, on multisite clouds, and on the european grid, which provide computing power but with lower availability and longer delays to access to the data.

2.Allowing reuse and reproducibility of experiments

The second challenge we addressed was to help scientists to foster discovery of novel traits and mechanisms based on the processed datasets, while providing a robust methodology with support for reproducibility and reuse. Modern scientific workflow systems are now equipped with features that offer this support. The provenance information (parameter settings and data sets consumed and produced) can be systematically recorded. Moreover, InfraPhenoGrid and OpenAlea workflows enhance reproducibility and reuse by providing users with the means to interact with provenance information through Jupyter electronic Notebooks [L3] [1].

The authors acknowledge the support of France-Grilles for providing computing resources on the French National Grid Infrastructure.

Links:

[L1] https://openalea.gforge.inria.fr
[L2] http://www.hdigitag.fr
[L3] https://jupyter.org

References:

[1] C. Pradal, S. Artzet, J. Chopard, et al.: "InfraPhenoGrid: A scientific workflow infrastructure for plant phenomics on the Grid", Future Generation Computer Systems 67: 341-353, 2017.

[2] J. Liu, E. Pacitti, P. Valduriez, et al.: "Multi-objective scheduling of Scientific Workflows in multisite clouds", Future Generation Computer Systems 63: 76-95, 2016.

[3] F. Tardieu, et al.: "Plant Phenomics, From Sensors to Knowledge", Current Biology 27(15):R770-R783, 2017.

Please contact:

Christophe Pradal, Inria, France
+33 4 67 61 97 94
christophe.pradal@cirad.fr

Patrick Valduriez, Inria, France
+33 4 67 14 97 26
patrick.valduriez@inria.fr