DIVIS

Project title: Biological Data Integration and Visualisation Acronym: DIVIS

Project duration: 36 months – Start date: 01/01/2018 End date: 31/12/2020

Key-words: heterogeneous datasets, data integration, ontology, data visualisation

Coordinator: Julie Bourbeillon, IRHS / Bioinfo, julie.bourbeillon@agrocampus-ouest.fr

Financial support from « Objectif Végétal »: 22.2k€ (Région Pays de la Loire)

Summary:

In biology high-throughput techniques and the multiplication of databanks have caused the scale at which data are acquired and shared to increase exponentially in recent years. The field is entering the so called “Big data” era. However, the datasets accumulated by biologists during the course of their experiments are often heterogeneous. Moreover, even if they can be stored and retrieved together through databanks they are usually not bound together because there are discrepancies between the experimental settings. This poses problems to biologists to analyse jointly these disparate datasets in order to acquire new insights into their biological interpretations.

However existing software does not provide complete user-friendly solutions to gather these datasets together, render their size manageable for instance through data summaries and allow scientists to interpret them easily through visual displays. The goal of the project is therefore to perform a preliminary study and explore new bioinformatic approaches to solve this issue in a generic way, that is to say not tailored for a specific organism or a limited range of experimental data.

During the course of the project we will develop a prototype software using apple and seed datasets. The tool will manipulate a large matrix containing all datasets extracted from the research unit database and: (I) normalise datasets so that they can be treated jointly; (ii) group similar samples analysed in similar experimental settings. The grouping process will be performed based on previous knowledge stored in specifically designed ontologies; (iii) represent each group by a representative archetype individual; (iv) summarise data for the archetype individuals; (v) build a visual display of the data summaries which the biologist will be able to navigate to acquire an understanding of the underlying datasets.

Project description

1. SCIENTIFIC PROJECT

Objectives (one half-page maximum)**:**

Integration of heterogeneous omics and phenotypic datasets at a large scale is a mandatory task for current scientific studies. However, methods automating this approach are still at a research stage and to our knowledge no operational and user-friendly software yet exists. Experiments are performed independently and resulting data are manually a-posteriori cross-analysed by scientists. For instance, the different IRHS biology teams have been accumulating datasets of different natures (transcriptomic, biochemistry, physical measures, sensory analysis, etc.) regarding perennial, annual and biannual plants. These datasets are described using reference ontologies enriched with in-house knowledge and stored in a Laboratory Information Management System (LIMS) which is developed and distributed by the IRHS Bioinformatics team[1].

The main objective of the present project (DIVIS) is therefore to develop a directly usable prototype of such a data analysis tool, by combining the most promising integration and visualisation approaches in the data science domain to exploit data stored in our LIMS. At a first stage, the tool will gather from the LIMS database and normalise experimental datasets regarding samples of a close nature, across levels from the molecule to the organism, across nature of experiments and experimental designs, which is seldom performed by existing software, in particular in plant biology. The originality of the integration approach regards the analysis of the resulting matrix: (I) reduce the number of individuals by regrouping similar samples using a similarity score, (ii) calculate this score based on similarity between their metadata variables stored in a specifically designed ontology and (iii) represent each group by an archetype sample. The visualisation approach will allow to present data regarding these archetype samples in a multi-layer display separating various subsets of coherent data and to navigate through the results.

In order to validate the methodology, the tool will be developed based on two test datasets acquired as part of matching experiments (including several studies performed on the same samples) where there are no missing data. An apple fruit dataset including transcriptomic, biochemical, physical, sensory data and a seeds dataset containing phenotypic data on germination, longevity, biochemical composition, physical attributes and genotyping data (SNPs) will be considered to assess how the approach can be adapted to different experimental contexts with an equivalent level of complexity.

Context and Scientific / Socio-economic issues (one page maximum)**:**

In the general biological research community, recent years have seen two major shifts in the way biological research is conducted. On the first hand, high-throughput techniques have permitted to increase the scale at which experiments are conducted and are slowly spreading from the molecular level to the phenotype and even population level. On the second hand, the resulting datasets are more and more shared through public repositories which are beginning to host large amounts of data (for instance Genbank is closing on the 200 millions sequences as of October 2016[2]). Because of this new context, biologists are faced with the "5Vs" challenges of big data: Volume, Velocity, Variety, Value, Veracity [1].

Beyond the volume of available data these new trends in biology are indeed causing new problems. A first difficulty lies with the **data source heterogeneity**. There is a myriad of available databanks, some generalist and some focusing on a specific topic, some publicly available and some in-house for a given research laboratory. These databanks host data with heterogeneous formats and interfaces and coexist with data files stored on

[1] https://sourcesup.renater.fr/projects/elvis/

[2] Source: GenBank statistics page at https://www.ncbi.nlm.nih.gov/genbank/statistics/