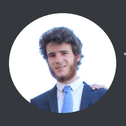
Visualizing metadata change in networks and / or clusters

Tanguy Lallemand, Sylvain Gaillard, Sandra Pelletier, Claudine Landès, Sébastien Aubourg, Julie Bourbeillon

CoRGI Co-Regulated Gene Investigator

Visualizing metadata change in networks and / or clusters



Tanguy Lallemand

Tanguy Lallemand[1], Sylvain Gaillard[1], Sandra Pelletier[1], Claudine Landès[1], Sébastien Aubourg[1] and Julie Bourbeillon[1]

1 IRHS, Agrocampus-Ouest, INRA, Université d'Angers, SFR 4207 QuaSaV, 49071, Beaucouzé, France

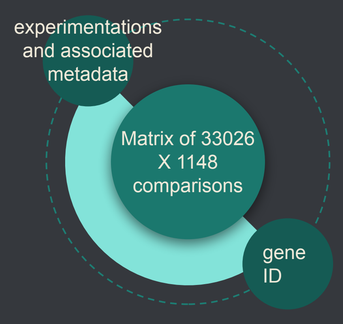
BIDefI

Introduction

- Wider spread of high-throughput experimental techniques and multiplication cross-analysed studies increased the need for **integration of large and heterogeneous data sets from "omics" activities and phenotyping**.
- Classical approaches imply **collating various data sets into a large matrix** and mine the matrix. Results can be visualized using network graphs or heatmaps.
- These are difficult to interpret in biological terms without including comprehensive descriptions of data points to the visualizations. This is challenging when dealing with **annotations extracted from large knowledge representations** such as the Gene Ontology [1]. In addition, iterative approaches in mining techniques and the cinetic notion in biological data sets require representations taking into account a chronology through the data mining steps or the course of the biological process.
- In this context we are developing a web-based tool to **represent ontology annotations associated with biological network graphs or cluster heatmaps through time** using:
  - series of snapshots corresponding to successive steps
  - representation of differences between two steps.
- Our approach builds on methods such as GO enrichment analysis and visualization of changes in networks.

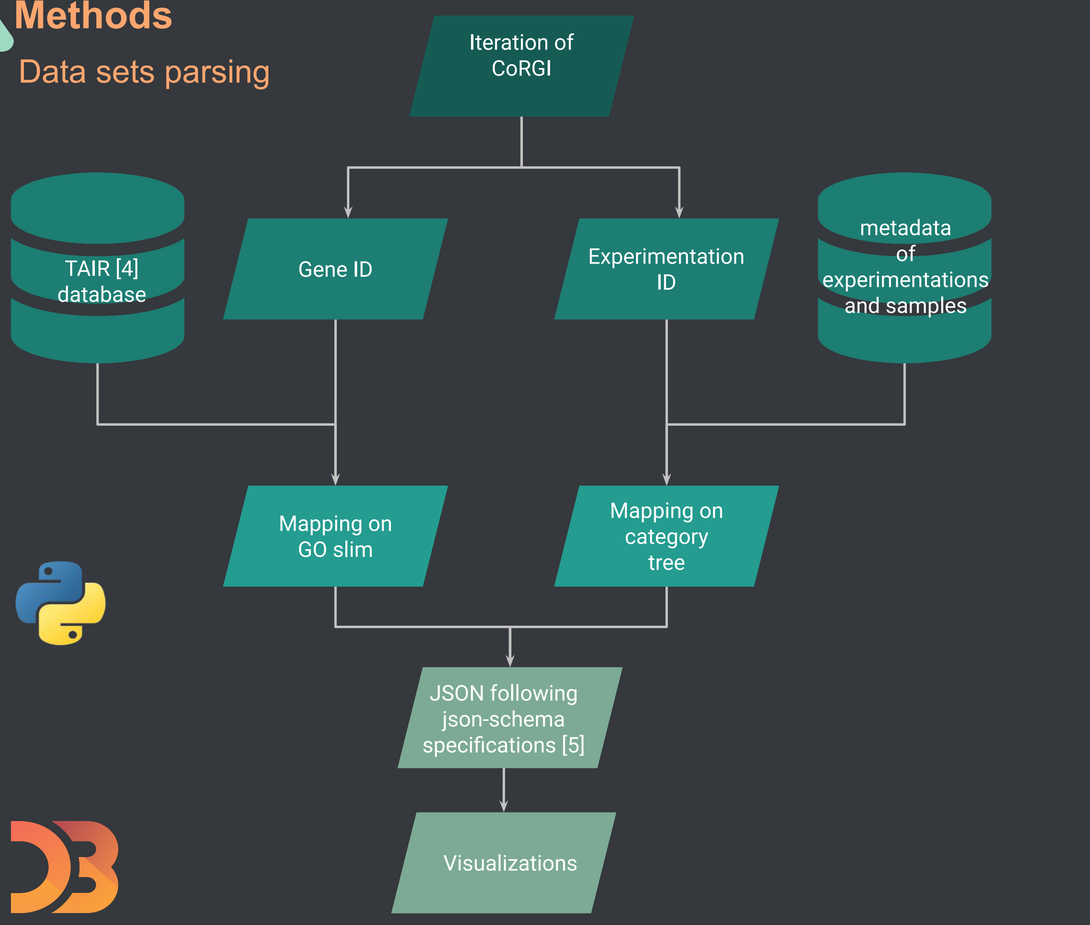
Dataset

- *Arabidopsis thaliana* transcriptomic experiments from the CATdb [2] (1043 differential analysis) and from Gene Expression Omnibus [3] (105 differential analysis) databases.
- 33026 genes
- Thanks to the convergence of Normal law toward Binomial law , CoRGI provides a subgroup of genes jointly regulated in a subset of experimental conditions for each iteration.



Methods

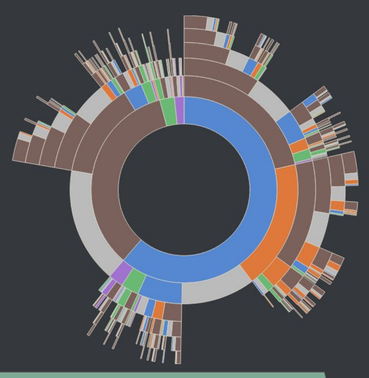
Data sets parsing



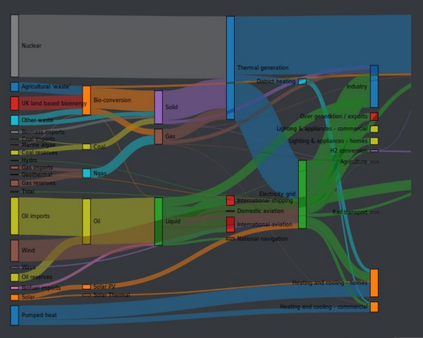
Summary of data sets parsing

Visualizations

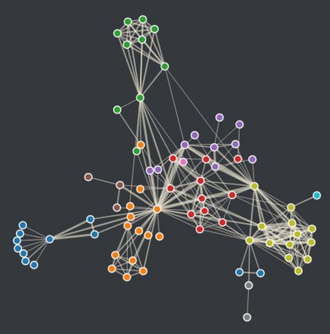
- To represent high volume of sequential metadata in two dimensions few approaches are possible [6] including:
  - **Networks diagram** can deal with sequential data but not with an efficiency and readability (iterations using colors or shapes codes).
  - **Sequence sunburst** is more compact but also less visual. In fact there is no links displayed between nodes.
  - **Alluvial diagrams** are sometime used to visualize tree paths. Inspired by this approach, this type of visualization has been adopted to show changes of matrix at each iteration.
- Final choices:
  - Alluvial diagrams to display changes in networks.
  - Force directed graph deployed to explore a given iteration
- Alluvial diagrams built using D3.js [7] version 5 and in particular the sankey module version 0.12.1. Some functions were rewritten to fit with project requirements and in particular variable width of links.



Example of sequence sunburst



Example of sankey diagram to illustrate tree path



Example of force directed graph

Statistical status of links

- In order to display only interpretative informations, each links is statistically tested using one GO enrichment approach. For each GO Slim term or experiment term, a contingency table is constructed and a two tailed Fisher's exact test are carried out.
- Links are colored following statistical significance or not.

Results

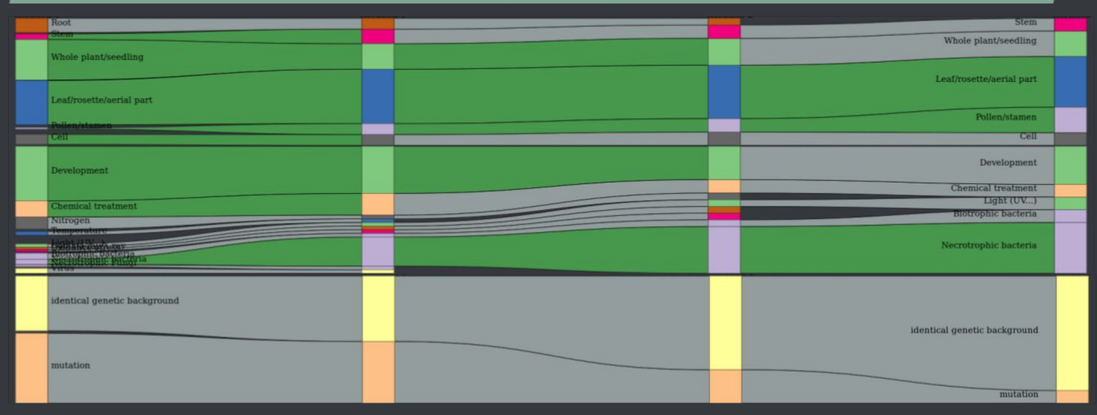
Different visualizations produced:

- Two groups of alluvial diagrams representing experiment metadata and gene annotation changes. Representation using:
  - Nodes as rectangle. Their height represents the percentage of representation of this term in relation to all items mapped on this part of GO for each iteration.
  - Links between nodes. Their colors are based on statistical status.
- One force directed graph deployed on user request, to display Gene Ontology graph with a given GO Slim term as root and all genes with this GO Slim reference mapped on their GO term. Number of genes mapped on each term is represented by node size.

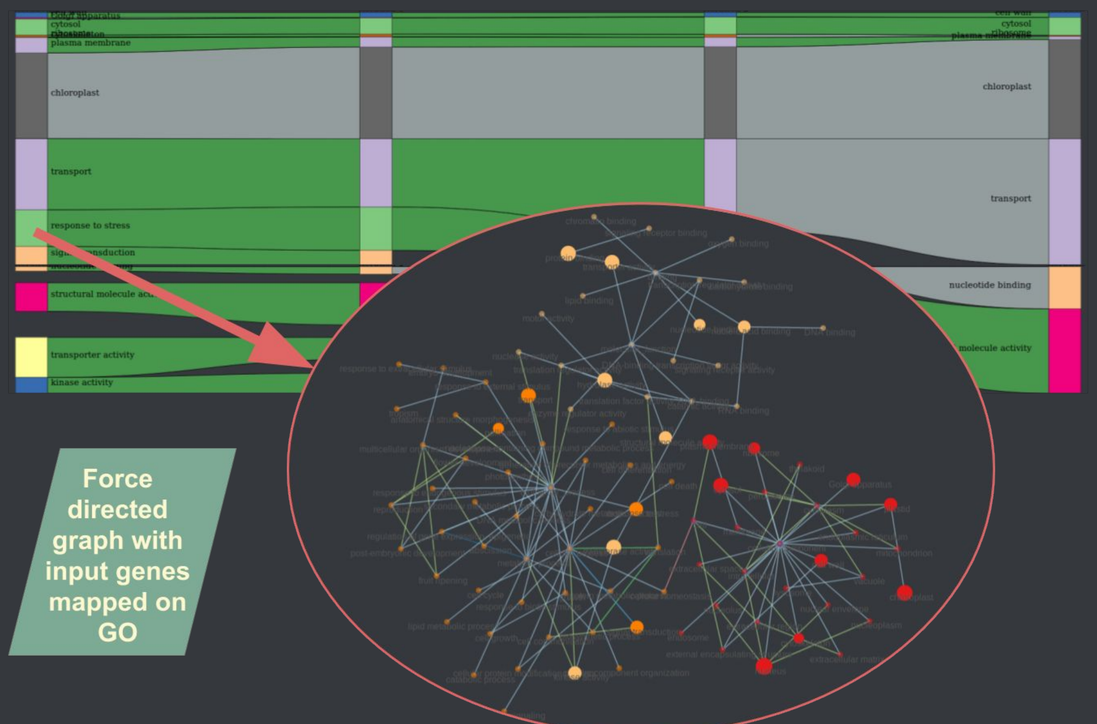
Initial visualization of down regulated genes in clusters

Initial State | Iteration 1 | Iteration 2 | Iteration 3

Alluvial diagrams used to visualize changes in experimentations or genes metadatas



Experiment metadatas



GO Slim

Force directed graph with input genes mapped on GO

Conclusions

- Client side application, need to precompute as possible. Need to minify all JSON datas to limit size of transferred files.
- Using adapted pre-existent visualizations, this project allows to visualize at same place and with efficiency, results of an heterogeneous matrix bi-clustering using associated metadata from different reference ontologies.
- Use of alluvial diagrams to represent changes in networks allowing evolution display of a matrix composed by genetic and experimental metadata in an interpretative way with statistical information.
- This project allows to cross-analyze different types of information in the same place with a visual and efficient presentation.

References

[1] Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–29.

[2] Gagnot, S., Tamby, J.-P., Martin-Magniette, M.-L., Bitton, F., Taconnat, L., Balzergue, S., Aubourg, S., Renou, J.-P., Lecharny, A., Brunaud, V. (2008). CATdb: a public access to Arabidopsis transcriptome data from the URGV-CATMA platform. Nucleic Acids Research, 36, D986-D990. , DOI : 10.1093/nar/gkm757

[3] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository Nucleic Acids Res. 2002 Jan 1;30(1):207-10

[4] Berardini, TZ, Mundodi, S, Reiser, R, Huala, E, Garcia-Hernandez, M,Zhang, P, Mueller, LM, Yoon, J, Doyle, A, Lander,G, Moseyko, N, Yoo, D, Xu, I, Zoeckler, B, Montoya, M, Miller, N, Weems, D, and Rhee, SY, (2004) Functional annotation of the Arabidopsis genome using controlled vocabularies. Plant Physiol. 135(2):1-11.

[5] A. Wright and H. Andrews, "JSON Schema: A Media Type for Describing JSON Documents," Internet-Draft draft-handrews-json-schema-01, Internet Engineering Task Force, Mar. 2018.

[6] Dudáš, M., Lohmann, S., Svátek, V., & Pavlov, D. (2018). Ontology visualization methods and tools: A survey of the state of the art. The Knowledge Engineering Review, 33. doi:10.1017/s0269888918000073

[7] Michael Bostock, Vadim Ogievetsky et Jeffrey Heer, D3: Data-Driven Documents, IEEE Press, coll. « IEEE Transactions on Visualization and Computer Graphics », octobre 2011