Reconstruction and visualisation of gene regulatory networks from different provenances : application at *Arabidopsis thaliana*

Estelle CHAIX*, Robert BOSSY, Claire NEDELLEC
MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France
(* attending member. e-mail : estelle.chaix@inra.fr)

In biology, genetic regulatory networks can be reconstructed and visualized as networks, where nodes are biological entities (*e.g.* gene, RNA, protein) and links denote biological processes (binding, interaction, regulations …). This networked visualization is very important for biologists who can infer cascading behaviours following certain processes (for example, the mutation of a gene upstream of a process can involve a regulatory cascade that will extinguish other genes etc...).

Different data provenances can have an impact on the representation of the network. Indeed, they can come by different means such as automatic or manual annotations, automatic or manual network reconstructions, etc.... We hypothesize that depending on the source of the data, we can rely more or less on the parts resulting from the recreated network. We aim to characterize the impact of different data sources on a more global regulatory network, seeking to integrate data from different sources.

The goal of the proposed work is the capacity to build a general knowledge base from different sets of annotations. The aim is to evaluate on a complete network which ratio of information comes from the different sets of data, as well as the reliability that can be brought to this global representation. The visualization of the final regulation network will have to take into account the different origins, which will allow the biological expert to get an idea of the confidence to place in the different data (Figure 1).

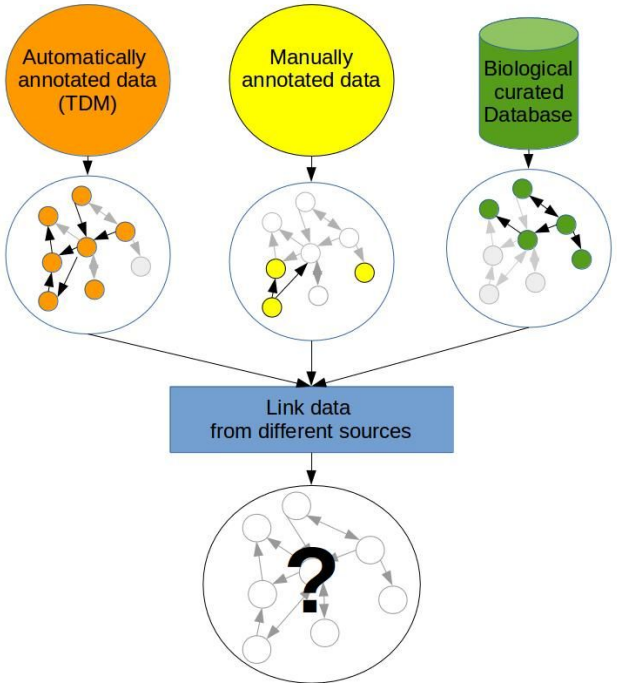

Figure 1: Proposal for the task of rebuilding a regulatory network from different data sources.

Biological context :
A comprehensive understanding of the molecular network underlying the regulation of plant development is a major scientific challenge with high potential impact on fundamental research,

agriculture and industry. For example, plant development requires the coordinated growth of different tissues that involves complex genetics and environmental regulation. The application example is the reconstruction of the genetic regulation network in the plant, and more particularly in the *Arabidopsis thaliana*, a model of study in biology.

Most of this knowledge is spread in thousands of articles, in different forms. Regulatory network can be provided by text, represented in figures more or less formalized, or from common resources such as IntAct[1] or BioGRID[2]. Automatic annotations from the literature can be used to recreate networks, or to supplement existing ones.

Annotation data sets :

- We provide manual annotation data from 21 full-papers (BioNLP-ST SeeDev[3]), automatic annotation from relevant PubMed abstract, locations of relevant external resources (database).
- We provide a standardized semantics (knowledge representation schemata) between these different resources.
- The automatic annotation set will be integrated in PubAnnotation (*e.g.* gene, RNA, protein of *Arabidopsis thaliana*).

Proposed Work:

We propose the following work for the hackathon:

- Develop code that generates network representations for an automatic annotated text from a PubMed corpora, a manual annotated corpora (SeeDev BioNLP-st) and curated database data.
    - For each resource, generate a knowledge base network representations for all kind of entities (gene, RNA, protein) and links between them (binds, interaction, regulation)
    - Merge these different base network representations from different provenances, to provide a general knowledge base, taking account of the origin and data confidence.
- Evaluate the utility of this general knowledge network with representations by:
    - Creating visualizations using open source visualization tools (such as Cytoscape).
    - Providing descriptive network statistics (*e.g.* number of nodes/edges, average degree, percentage of each resources used to build the main network).
    - Assess the level of confidence in this network by taking into account the confidence of the different sources of annotations.

We provide expertise in this biological domain, semantic representation, as well as automatic data extraction from text.

Current identified challenges :

- What are the most relevant ways or tools to visualize these kind of data?
- What to do in case of conflict? How to identify them?
- How to visualize the different data sources in a global representation?

Perspective works and collaboration :

- PubAnnotation annotations in a Cytoscape view (RDF to CSV format) or any other relevant visualization tool
- Improvement of methods for extracting data from subject text

---

[1] All IntAct data and software is freely available to all users, academic or commercial, under the terms of the Apache License, Version 2.0, from https://www.ebi.ac.uk/intact/downloads

[2] All data provided are 100% freely available to both academic and commercial users under the MIT License.

[3] Chaix, Estelle, et al. "Overview of the regulatory network of plant seed development (seedev) task at the bionlp shared task 2016." Proceedings of the 4th BioNLP shared task workshop. Berlin: Association for Computational Linguistic. 2016.