ASaiM

Lessons learned from developing a framework for biologists

Bérénice Batut — October 16th, 2015


BIOSCIENCES

PhD thesis in bioinformatics and computational biology

- Contribution to aevol project
- Development of simple Python scripts

Post-doc in bioinformatics

- Development of ASaiM project

ASaiM project

Objectives

Development of a bioinformatics environment to analyze data from gut microbiota

Gut microbiota

Community of microorganism species that live in the digestive tracts

"Forgotten" organ

Metagenomic: study of microbiota

DNA, RNA

Bioinformatician work

Who's there?
What are they doing?
How are they doing?

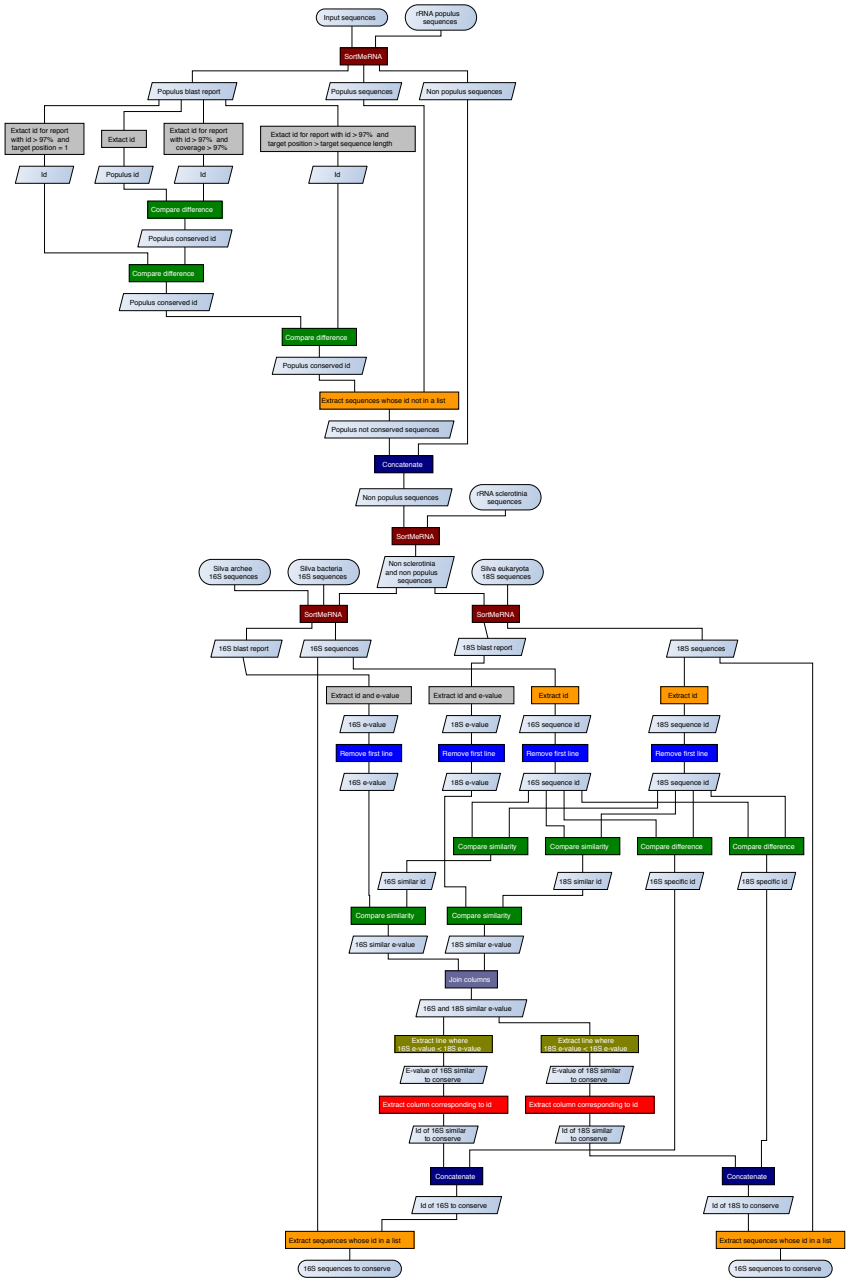
ATGCATTATTTGGCGGCAATA
ATGCATTATTTGGCGGCAATA
ATGCATTATTTGGCGGCAATA
ATGCATTATTTGGCGGCAATA
ATGCATTATTTGGCGGCAATA ATGCATTATTTGGCGGCAATA
ATGCATTATTTGGCGGCAATA
ATGCATTATTTGGCGGCAATA
ATGCATTATTTGGCGGCAATA
ATGCATTATTTGGCGGCAATA

Sequences

Complexity

- Short sequences
- Sequence variability
- Uncomplete reference databases
- ...

Need for numerous treatments to extract useful information



Example of workflow to sort sequences given their type

ASaiM framework

Bioinformatics framework to generate workflows to analyze data from gut microbiota

Main Requirements

- ☐ Generation of workflow with numerous tools
- ☐ Easy to use
- ☐ Flexibility
- ☐ Heavily and easily documented
- ☐ Easy to maintain

First tested approach

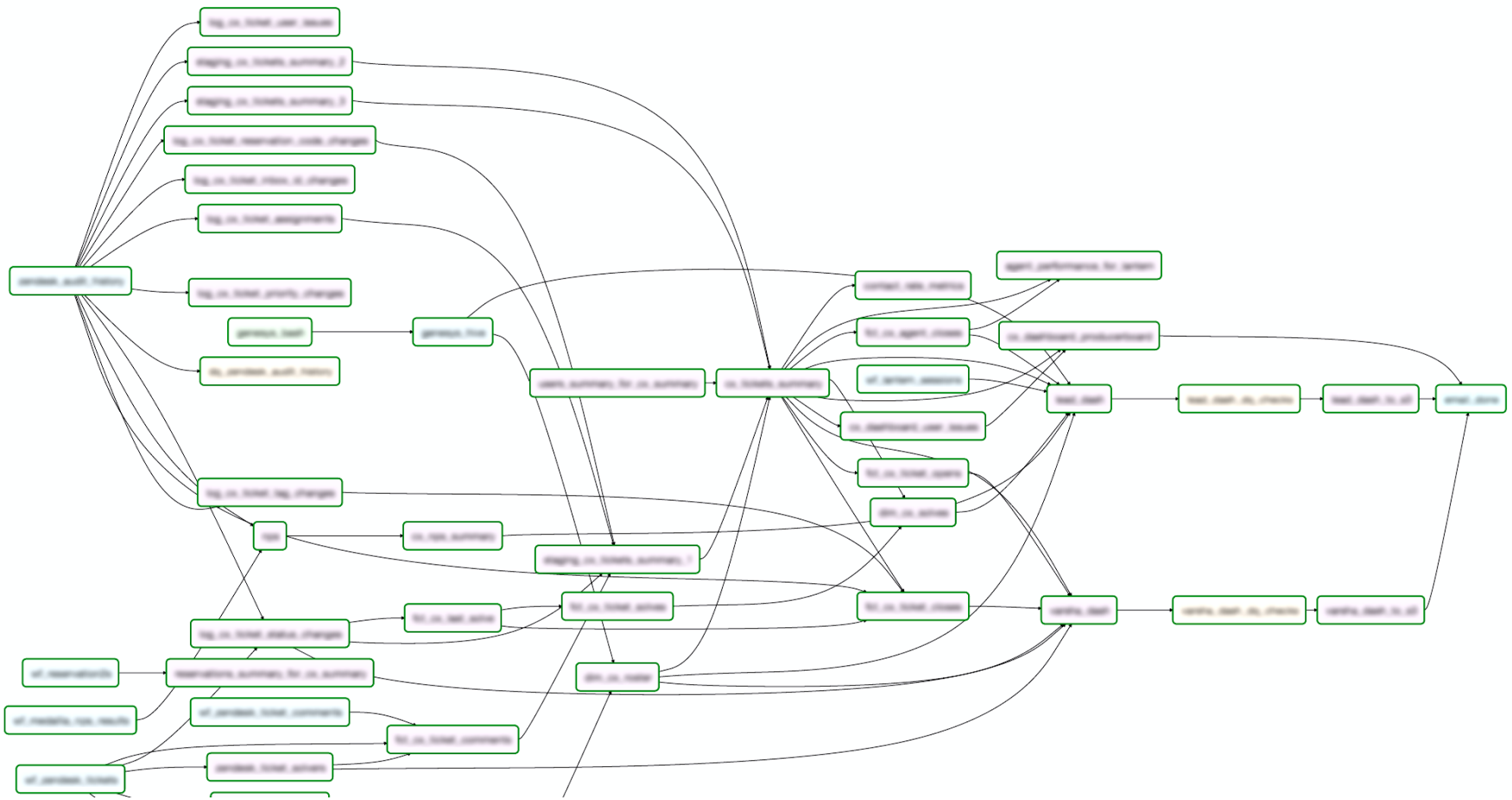
Simple Python scripts

Fit with framework requirements ?

- ☐ Generation of workflow with numerous tools
- ☐ Easy to use
- ☐ Flexibility
- ☐ Heavily and easily documented
- ☐ Easy to maintain

Second tested approach

Worflow manager such as Luigi, Airflow, ...



Airflow dependency graph (from Airbnb site)

Fit with framework requirements ?

- ☑ Generation of workflow with numerous tools
- ☑ Easy to use
- ☐ Flexibility
- ☐ Heavily and easily documented
- ☑ Easy to maintain

Third tested approach

Homemade approach

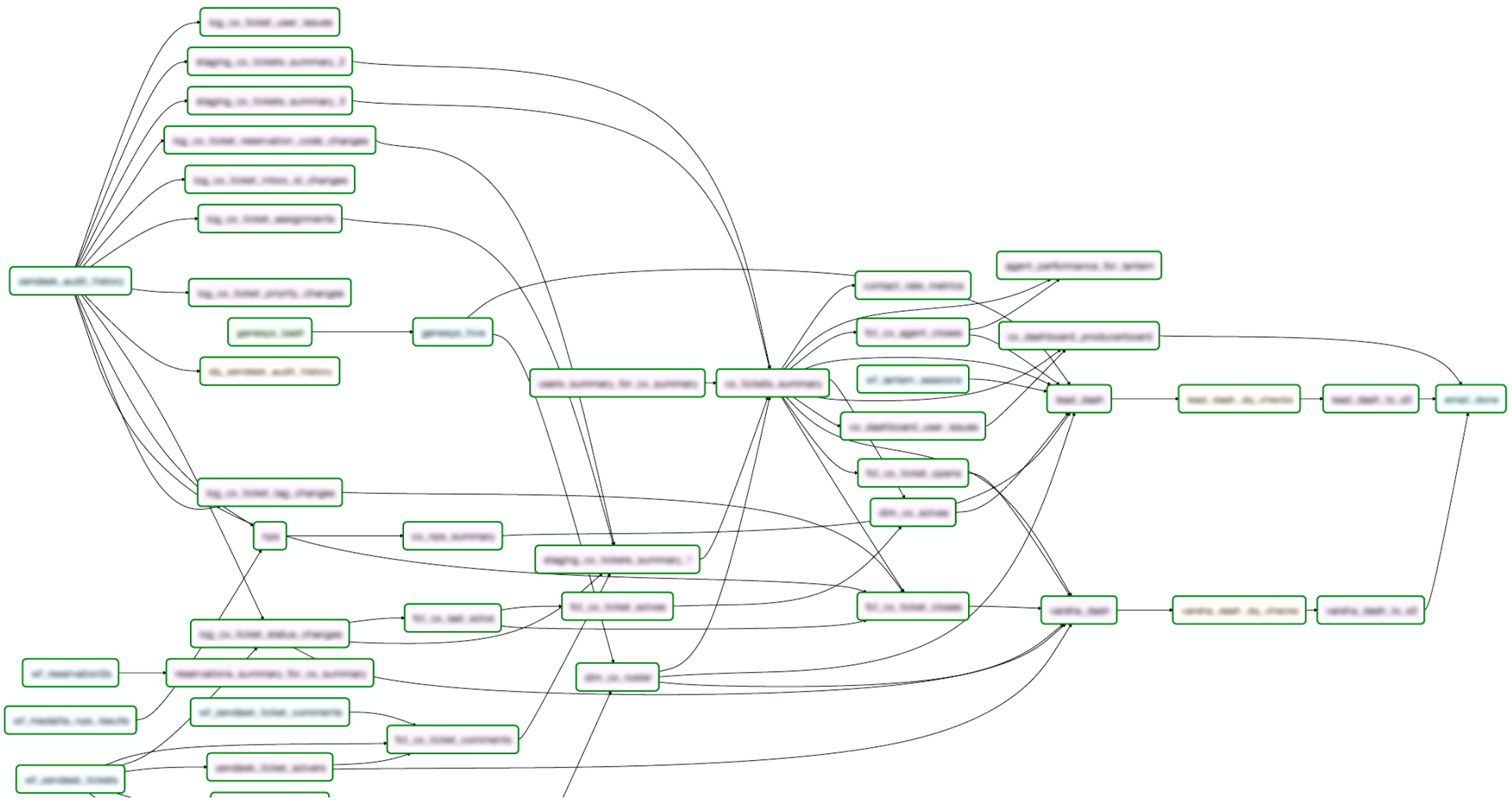
- Configuration file
  - Workflow description
  - Web interface for generation
- Python scripts to execute workflow in configuration file

Fit with framework requirements ?

- [x] Generation of workflow with numerous tools
- [x] Easy to use
- [ ] Flexibility
- [ ] Heavily and easily documented
- [ ] Easy to maintain

Main issue with these approaches

Dependency between the tasks

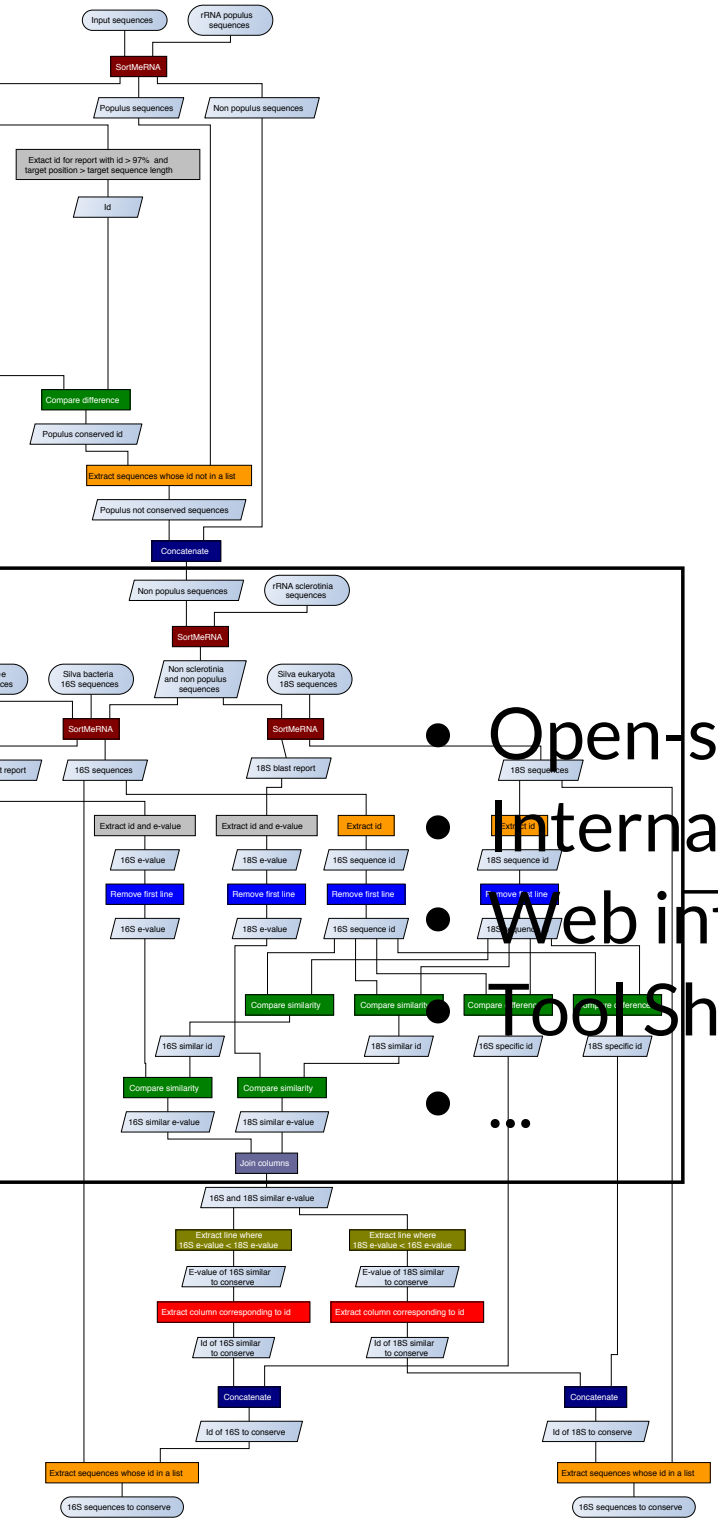


Airflow dependency graph (from Airbnb site)

Final approach

Galaxy

- Open-source project based on Python
- International development community
- Web interface
- Tool Shed
- ...



Fit with framework requirements ?

- ☑ Generation of workflow with numerous tools
- ☑ Easy to use
- ☑ Flexibility
- ⊟ Heavily and easily documented
- ⊟ Easy to maintain

Galaxy dependency graph

Workflow to sort sequences given their type

AsaiM framework

- Configuration of a Galaxy server
- Development of wrappers for tool integration
- Development of scripts to use Galaxy and API

Used tools

- Code
  - Github and submodules
  - Gitlab
- Documentation
  - Sphinx + ReadTheDoc + Github
- Web page
  - Jekyll + Github page
- Management
  - Trello
  - Slack

Learned from this project

- Need to correctly define the conception
  - *No workflow manager with input/output dependency*
- Do no reinvent the wheel
  - *Do not prefer home-made solution*
- Integrate active community
  - *Galaxy*, *Clermont'ech*, *Pycon*, ...
- Need of good tools and good habits in big projects
  - *Roadmap, management and development tools*

Thank You.

Questions?

 bebatut.fr

 github.com/bebatut

 twitter.com/bebatut