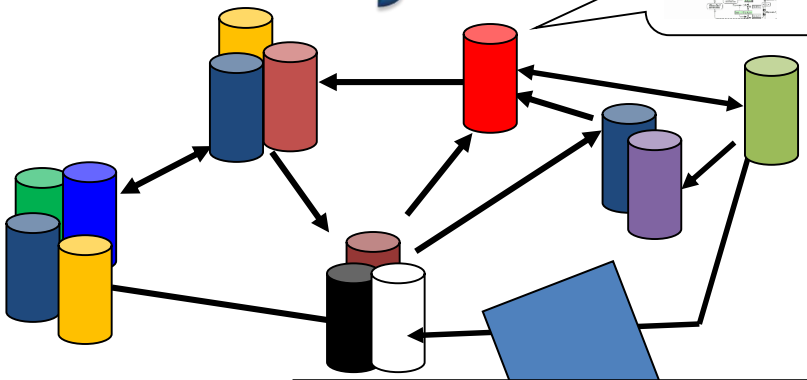
Reproducible Science in Bioinformatics

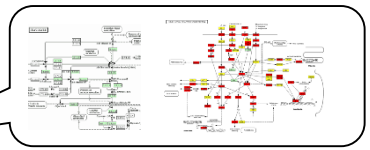
Sarah Cohen-Boulakia
Laboratoire de Recherche en Informatique (LRI)
CNRS UMR 8623
Université Paris-Sud

Biological analysis

Public sources

- Distributed
- Heterogeneous
- Network





CCCTTTCCGTGT
TGCCGTGTGGC
TAAATGTCTGTG
GTCTGTGC...

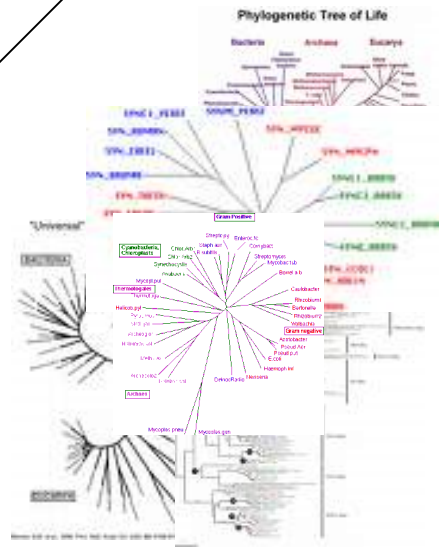
Tools
Scripts
Python

JAVA, Perl
Web services
...

How these data have been generated? With which input data? Which tools? Which parameters?

Tools

- Distributed
- Heterogeneous
- Chained



Workspace

What is the difference between these two experiments?

Take Home Message

Compared to 20 years ago…

- The number and diversity of the sources has increased a lot
  - > 1,500 databases (NAR databases issue)
  - Need for data provenance to determine data quality
- The complexity of the pipelines to be designed has increased a lot
  - Need for process provenance to determine data quality

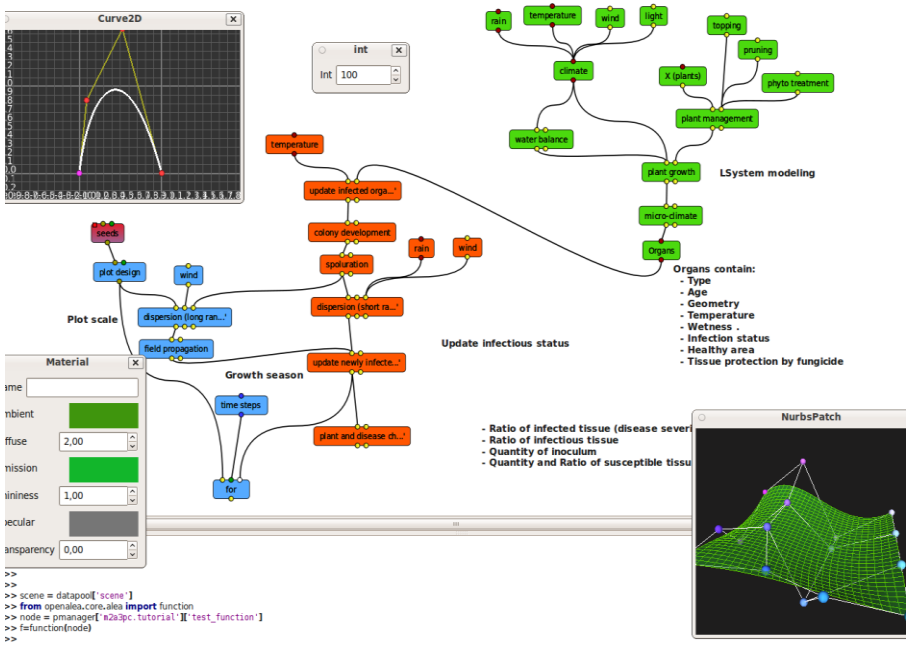
→Increase in the heterogeneity of data
+ Increase in the complexity of analysis pipelines
+ *Increase in the need to publish…*
= increasing difficulties to reproduce experiments!

Analysing & Integrating biological data

- Use scripts (Python, Perl, …)
  - quick to develop
  - hard to maintain
  - almost impossible to share
  - no high level view of the analysis steps…
  - …
- Use Scientific Workflows
  - *Visual programming*: chaining processors (from libraries…)
  - SWF Systems take care of important issues : Scheduling and parallelization, logging, debugging, integration of web services, recovery, provenance

→ Reproducibility

Scientific workflow



- Companion tools
  - Virtualisation & Container techniques (Docker, …)
  - Notebooks
  - …