ReproVirtuFlow (Action MaDICS) Reproductibilité des expériences d'analyse de données scientifiques : bilan et perspectives

Sarah Cohen-Boulakia

Université Paris-Sud, Université Paris-Saclay, LRI CNRS INS2I UMR 8623

Christophe Blanchet

Institut Français de Bioinformatique (IFB-Core) CNRS INSB UMS 3601

Contexte, enjeux

Reproductibilité *computationnelle*

Nombre croissant de résultats scientifiques non reproductibles

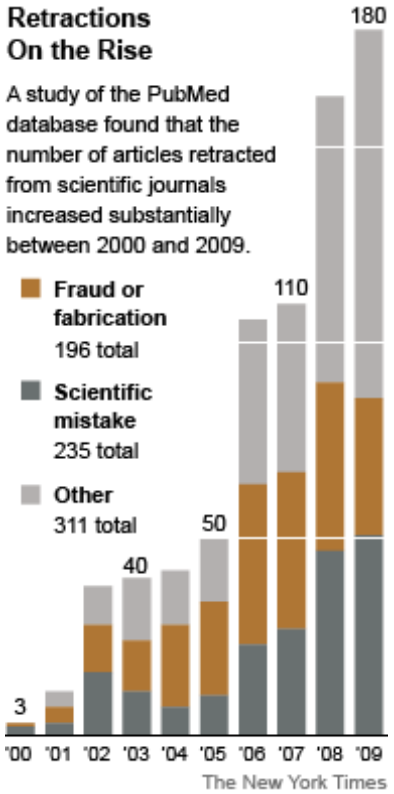
Nombreux domaines concernés

Enjeux économique majeur

- Non reproductibilité des études pré-cliniques évalué à >$10 milliards annuel pour les USA

Devient une obligation contractuelle

- Projets NSF, certains editeurs

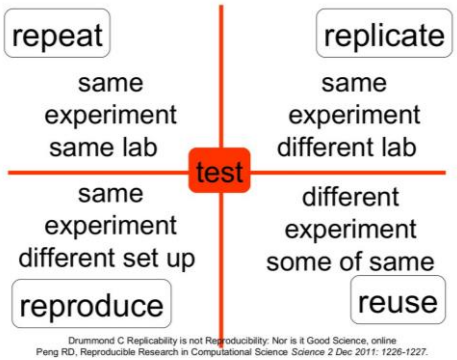


Objectifs de l'action

Sensibiliser la communauté

Etat des lieux et niveaux de reproductibilité

- Solutions ? Niveaux à considérer ? Couverture ?



Focus sur trois types d'approches

- Definition de l'analyse
  - Workflows scientifiques : outils - ordre
- Données et paramètres d'entrée
  - Provenance : Trace des executions
- Environement d'exécution
  - Virtualisation, Packaging : Trace de l'environement (OS, librairies…)

Communautés :
Base de données, représentation des connaissances, algorithmique (graphes), grilles, systèmes, compilation, langages…

Membres

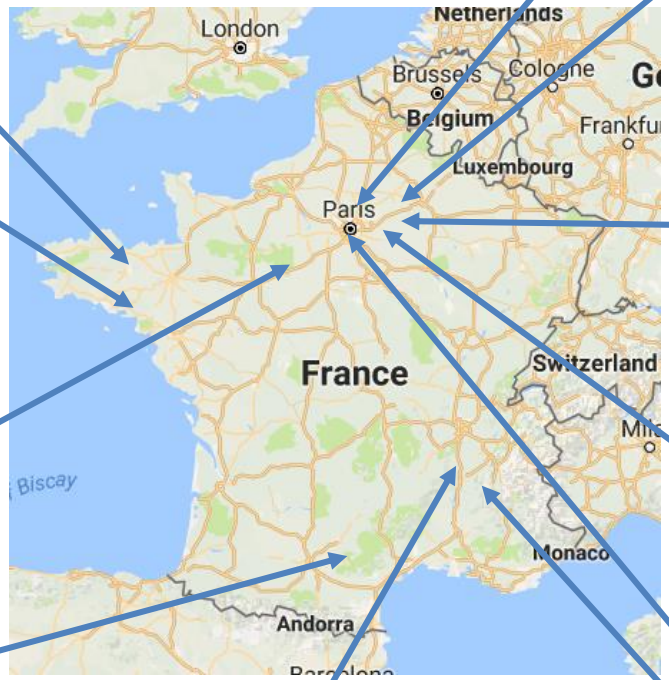
UMR et UMS du CNRS

IRISA Univ. Rennes

Univ. CHU Nantes

Centre de Biophysique Moléculaire, CNRS Orléans

IRD, CIRAD, INRA, Inria, Univ. Montpellier



Univ. Lyon 1 LIRIS

LRI Univ. Paris Sud

CDS, Center for Data Science Saclay

Institut Francais Bioinformatique Gif s/Yvette

Institut Pasteur, Paris

Lamsade Univ. Paris Dauphine

LIG (Grenoble)

GDR Bioinfo, Groupes de travail IFB

Centres *Data Sciences internationaux*

Analyses Bioinformatiques

Des Données aux connaissances

- Données

  Distribuées, **Heterogènes**
- Outils

  Different types, nb parametères
- Pipelines d'analyses (*workflows*)

  Complexes

Cas d'utilisation cibles

- NGS (cancer), Phenotypage Plantes

  Données massives
- ELIXIR Europ. Research Infra.

  21 Pays, 180 partenaires

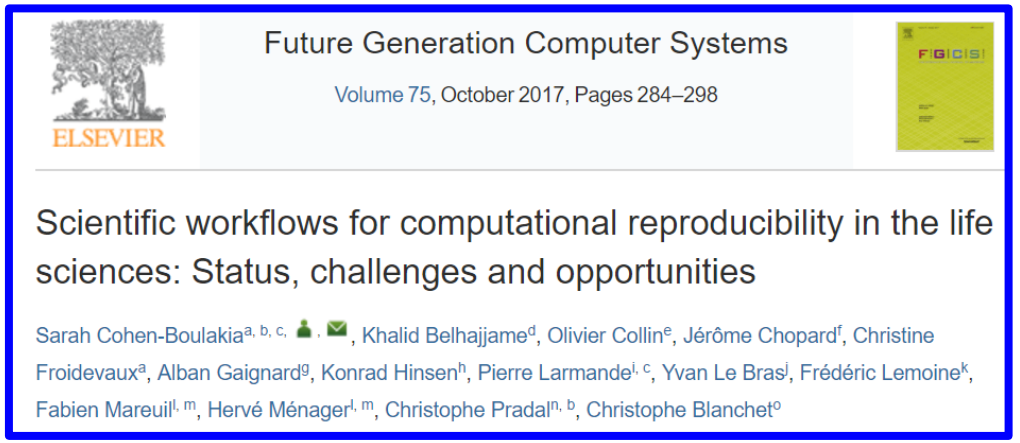
  ➔ Analyses avec workflows scientifiques

Résultats

(1) Etat des lieux sur les solutions existantes à base de workflows + verrous

(2) Webinar en ligne tutoriel et 2 démonstrations

(3) Conception et organisation du premier ReproHackathon



ReproHackathon 1 : Reproduction d'un pipeline d'analyse de données RNA-Seq publié (mélanome Uvéal)

1-2 Juin 2017
Gif s/Yvette
25 participants

IGRoussy, Curie, Pasteur, Saclay, Paris, Nantes, Lyon...

Perspectives

2015-2017: Reproductibilité Computationnelle …

- Traitements de données à l'aide de **systèmes de workflows**
- Etapes d'analyses **déterministes**
- Données de **biologie moléculaire**

Généralisation, nouveaux problèmes ? Solutions ?

- Workflows scientifiques → **scripts**
- Analyses **non déterministes**, apprentissage (statistique)
- Nouveaux **types de données** à analyser (images, graphes…) et nouveaux **domaines** (robotique, reproductibilité et éthique)

Animations futures (2018)

Journées de rencontre et montage d'au moins **deux nouveaux ReproHackathon**

- Images, machine learning, traitement du signal…
- Collaboration MaDICS : EADM, ATLAS, GRAMINEE…
- **Numéro spécial** revue IEEE CiSE (Computing in Science Engineering)

Journée **éthique et reproductibilité**

- CERNA - Commission de réflexion sur l'Éthique de la Recherche en sciences et technologies du Numérique d'Allistene
- Groupe travail « Recherche et innovation responsables : Éclairages interdisciplinaires » (Maison des sciences de l'Homme Paris-Saclay)








ReproVirtuFlow @ cnrs MaDICS

https://www.madics.fr/actions/actions-en-cours/reprovirtuflow/