Understanding Reproducibility of Bioinformatics Workflows

Sehrish Kanwal, Andrew Lonie, Richard Sinnott
The University of Melbourne, Australia

Thesis and associated publications available online at: https://minerva-access.unimelb.edu.au/handle/11343/198443

ABSTRACT

The expectation for science to be reproducible is considered fundamental but often not tested. Importantly, the main problem itself has still been poorly defined. What it means to achieve reproducibility, what are the limitations and what aspects of a solution should be evaluated. Even, the terminologies such as repeatability, replicability, reproducibility, reusability, around this important concept are not standardized. Despite the critical need for reproducibility, the efforts required to make bioinformatics workflows reproducible are still largely in their infancy. This is due to an incomplete understanding of reproducibility, its requirements and assumptions related to the varied workflow definition approaches.

This is where the PhD project “Understanding reproducibility of bioinformatics workflows” makes major contributions. It establishes rigorous definitions and explicit requirements associated with bioinformatics workflows. This research explores the relationship between workflow definition approaches and how those impact or limit reproducibility. It further establishes important dimensions on which reproducibility can be measured and enhance understanding where implicit or explicit assumptions in different workflow systems are leading to reduced reproducibility. Overall, this research proposes a framework to provide a coherent explanation of concepts associated with reproducibility and evaluate the levels of reproducibility offered by existing workflow approaches, which can guide improvement of these systems.

1. Motivation

- Ability to trust the results of previous research or extend/reuse the analysis procedures that were originally deployed to produce these results.
- Incomplete understanding regarding requirements for assuring reproducibility of workflows.
- Assumptions relating to the many different approaches that can be taken when defining and enacting (executing) workflows.
- Lack of clarity in terminologies and concepts associated with reproducibility.

2. Research Overview

3. Reproducibility Framework

3.1 Reproducibility Range: Higher reproducibility as a result of decreased dependence on a specific runtime environment

4. Portability (portable workflows are now labelled as “the new normal”)

5. Contributions and Conclusions

- Reproducibility of both dimensions i.e. workflow software and results have to go hand-in-hand.
- Categorization of workflow definition and implementation approaches into three broad categories; pre-built pipelines, GUI based integrative frameworks and standardized workflows.
- Characterize and highlight assumptions and omission of factors associated with workflow design and implementation.
- Identify implicit and explicit requirements of reproducibility and categorize these according to proposed dimensions of reproducibility.
- Ultimately, define the term reproducibility for bioinformatics workflows to develop a coherent understanding of the concepts associated with reproducibility.