Future Generation Computer Systems 75 (2017) 284–298

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities

Sarah Cohen-Boulakia [a,b,c,*], Khalid Belhajjame [d], Olivier Collin [e], Jérôme Chopard [f], Christine Froidevaux [a], Alban Gaignard [g], Konrad Hinsen [h], Pierre Larmande [i,c], Yvan Le Bras [j], Frédéric Lemoine [k], Fabien Mareuil [l,m], Hervé Ménager [l,m], Christophe Pradal [n,b], Christophe Blanchet [o]

[a] Laboratoire de Recherche en Informatique, Université Paris-Sud, CNRS UMR 8623, Université Paris-Saclay, Orsay, France
[b] Inria, VirtualPlants, Montpellier, France
[c] Inria, Zenith, Montpellier, France
[d] University Paris-Dauphine, PSL Research University, CNRS, UMR 7243, Centre Lamsade, 75016 Paris, France
[e] IRISA, Rennes, France
[f] INRA, UMR729, MISTEA, F-34060 Montpellier, France
[g] Nantes Academic Hospital, CHU de Nantes, France
[h] Centre de Biophysique Moléculaire (CNRS UPR4301, Orléans), France
[i] IRD, DIADE, F-34394 Montpellier, France
[j] EnginesOn / INRIA, Rennes, France
[k] Institut Pasteur, Unité Bioinformatique Evolutive, Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI, USR 3756 IP CNRS), Paris, France
[l] Institut Pasteur, Hub Bioinformatique et Biostatistique, Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI, USR 3756 IP CNRS), Paris, France
[m] Institut Pasteur, Centre d'Informatique pour la Biologie, Direction des Systèmes d'Information, Paris, France
[n] CIRAD, UMR AGAP, Montpellier, France
[o] CNRS, UMS 3601; Institut Français de Bioinformatique, IFB-core, Avenue de la Terrasse, F-91190 Gif-sur-Yvette, France

HIGHLIGHTS

- Use cases from the Life Sciences highlighting reproducibility and reuse needs.
- Terminology to describe reproducibility levels in scientific workflows.
- Criteria to define reproducible-friendly workflow systems and evaluation of systems.
- Challenges and opportunities in scientific workflows reproducibility.

ARTICLE INFO

Article history:
Received 13 May 2016
Received in revised form 14 October 2016
Accepted 7 January 2017
Available online 16 January 2017

Keywords:
Reproducibility
Scientific workflows
Provenance
Packaging environments

ABSTRACT

With the development of new experimental technologies, biologists are faced with an avalanche of data to be computationally analyzed for scientific advancements and discoveries to emerge. Faced with the complexity of analysis pipelines, the large number of computational tools, and the enormous amount of data to manage, there is compelling evidence that many if not most scientific discoveries will not stand the test of time: increasing the reproducibility of computed results is of paramount importance.

The objective we set out in this paper is to place scientific workflows in the context of reproducibility. To do so, we define several kinds of reproducibility that can be reached when scientific workflows are used to perform experiments. We characterize and define the criteria that need to be catered for by *reproducibility-friendly* scientific workflow systems, and use such criteria to place several representative and widely used workflow systems and companion tools within such a framework. We also discuss the remaining challenges posed by reproducible scientific workflows in the life sciences. Our study was guided by three use cases from the life science domain involving in silico experiments.

* Corresponding author at: Laboratoire de Recherche en Informatique, Université Paris-Sud, CNRS UMR 8623, Université Paris-Saclay, Orsay, France.

E-mail address: cohen@lri.fr (S. Cohen-Boulakia).

http://dx.doi.org/10.1016/j.future.2017.01.012
0167-739X/

1. Introduction

Novel technologies in several scientific areas have led to the generation of very large volumes of data at an unprecedented rate. This is particularly true for the life sciences, where, for instance, innovations in Next Generation Sequencing (NGS) have led to a revolution in genome sequencing. Current instruments can sequence 200 human genomes in one week whereas 12 years have been necessary for the first human genome [1]. Many laboratories have thus acquired NGS machines, resulting in an avalanche of data which has to be further analyzed using a series of tools and programs for new scientific knowledge and discoveries to emanate.

The same kind of situation occurs in completely different domains, such as plant phenotyping which aims at understanding the complexity of interactions between plants and environments in order to accelerate the discovery of new genes and traits thus optimize the use of genetic diversity under different environments. Here, thousands of plants are grown in controlled environments, capturing a lot of information and generating huge amounts of raw data to be stored and then analyzed by very complex computational analysis pipelines for scientific advancements and discoveries to emerge.

Faced with the complexity of analysis pipelines designed, the number of computational tools available and the amount of data to manage, there is compelling evidence that the large majority of scientific discoveries will not stand the test of time: increasing reproducibility of results is of paramount importance.

Over the recent years, many authors have drawn attention to the rise of purely computational experiments which are not reproducible [2–5]. Major reproducibility issues have been highlighted in a very large number of cases: while [6] has shown that even when very specific tools were used, textual description of the methodology followed was not sufficient to repeat experiments, [7] has focused on top impact factor papers and shown that insufficient data were made available by the authors to make experiments reproducible, despite the data publication policies recently put in place by most publishers.

Scientific communities in different domains have started to act in an attempt to address this problem. Prestigious conferences (such as two major conferences from the database community, namely, VLDB[1] and SIGMOD[2]) and journals such as PNAS,[3] Biostatistics [8], Nature [9] and Science [10], to name only a few, encourage or require published results to be accompanied by all the information necessary to reproduce them. However, making their results reproducible remains a very difficult and extremely time-consuming task for most authors.

In the meantime, considerable efforts have been put into the development of *scientific workflow management systems*. They aim at supporting scientists in developing, running, and monitoring chains of data analysis programs. A variety of systems (*e.g.*, [11–13]) have reached a level of maturity that allows them to be used by scientists for their bioinformatics experiments, including analysis of NGS or plant phenotyping data.

By capturing the exact methodology followed by scientists (in terms of experimental steps associated with tools used) scientific workflows play a major role in the reproducibility of experiments. However, previous work have either introduced individual workflow systems allowing to design reproducible analyses (*e.g.*, [14,15]) without the aim to draw more general conclusions and discuss the capabilities of the scientific workflow systems to reproduce experiments or it has discussed computational reproducibility challenges in e-science (*e.g.*, [16,17]) without considering the specific case where scientific workflow systems are used to design an experiment. There is thus a need to better understand the core problematic of reproducibility in the specific context of scientific workflow systems, which is the aim of this paper.

In this paper, we place scientific workflows in the context of computational reproducibility in the life sciences to provide answers to the following key points: How can we define the different levels of reproducibility that can be achieved when a workflow is used to implement an *in silico* experiment? Which are the criteria of scientific workflow systems that make them *reproducibility-friendly*? What is concretely offered by the scientific workflow systems in use in the life science community to deal with reproducibility? Which are the open problems to be tackled in computer science (in algorithmics, systems, knowledge representation etc.) which may have huge impact in the problems of reproducing experiments when using scientific workflow systems?

Accordingly, we make the following five contributions: We present three use cases from the life science domain involving in silico experiments, and elicit concrete reproducibility issues that they raise (Section 2). We define several kinds of reproducibility that can be reached when scientific workflows are used to perform experiments (Section 3). We characterize and define the criteria that need to be catered for by *reproducibility-friendly* scientific workflow systems (Section 4). Using the framework of the criteria identified, we place several representative and widely used workflow systems and companion tools within such a framework (Section 5). We go on to discuss the challenges posed by reproducible scientific workflows in the life sciences and describe the remaining opportunities of research in several areas of computer science which may address them in Section 6 before closing the paper in Section 7.

2. Use cases

This paper starts with a set of three use cases, extracted from real projects, where scientific workflow systems are used to manage data analyses.

2.1. Next generation sequencing for diagnosis in oncology

2.1.1. Context

New and powerful next-generation sequencing (NGS) techniques allow to simultaneously and quickly analyze a large number of genes, up to the entire genome, that are assumed to be involved in diseases. As recently highlighted [18], the main challenge in applying NGS to medical diagnosis resides in workflow development fulfilling diagnosis interpretation requirements, such as quality control or variant knowledge annotation.

In this context, the preeminent French health and science agency, National Cancer Institute (INCa), is in charge of cancer control in France. The goal of the INCa is to generalize existing workflows designed for diagnosis in oncology, and deploy them in most French hospital laboratories.

2.1.2. Computational tools used

In INCa, workflows are implemented through both very specific chaining tools using command-lines and workflow systems (Galaxy). As such workflows are used in production (and for diagnosis purpose), a particular attention has been paid in deploying solutions allowing different users to (virtually) work in the same run-time computational environment, ensuring in particular that the exact same version of tools and packages is available.

[1] International conference on Very Large Data Bases.

[2] ACM's Special Interest Group on Management Of Data.

[3] http://www.pnas.org/site/authors/format.xhtml.