Creating a good learning and sharing environment for bioinformatics

T. Klingström[1*] & J.I. Ohlsson[1]

[1]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Box 7023, 750 07 Uppsala, Sweden; [*]Tomas.Klingstrom@slu.se

Abstract

Most bioinformatics researchers create their own working environments depending on their training, access to computing resources and what advice they receive from other researchers. This makes the development of training programs difficult as much knowledge and work is locked into each researcher's unique computing environment for data analysis.

Using software container solutions such as Biocontainers and a 'one image per task' coding pattern when developing a workflow, it is easy for an application expert to rewrite workflows for research, training and dissemination of new workflows for analysing data. This paper describes the reasoning behind this approach which is now being tested in a collaboration between livestock researchers in Sweden and South Africa.

Introduction

Advancements in high-throughput technology for genotyping and phenotyping mean that data acquisition, storage, analysis and dissemination is an increasingly large part of the work in agricultural research. In a survey at the Swedish University of Agricultural Sciences, PhD students spend 25 % of their time dealing with challenges caused by a lack of training or tools necessary to perform data analysis (Klingström, 2017). Developing and providing the tools and training necessary to work efficiently with data is therefore an important task for the university to improve research efficiency. In the Life Sciences, specialisation in this area falls under the designation of bioinformatics which is a research discipline joining together biology and computer science. Depending on the context the subject of bioinformatics can be viewed from multiple perspectives:

- As a tool supporting the development of knowledge in other fields of the Life Sciences.
- As an independent research subject pushing the boundaries of human knowledge of its own accord.
- As an interface between biology and computer science providing biological inquiries access to high-powered computing resources.

The different perspectives make it difficult to disseminate the wide array of tools available for bioinformatics. Tools for researchers specialising in bioinformatics often prioritise flexibility and control while researchers who only occasionally use bioinformatics require tools prioritising ease of use and high quality training materials. This creates difficulties in sharing bioinformatics workflows between research groups. In the B3Africa project (Klingström *et al*., 2017), we worked on how to enable interdisciplinary collaborations incorporating all three perspectives to enable researchers to better disseminate bioinformatics tools and apply them to available datasets. One of the key conclusions of the project was that Galaxy (Jalili *et al*., 2020) is a versatile solution which helps us break down the matter of bioinformatics into three distinct components:

- A single user interface which can be used by researchers with little training to perform bioinformatics research using virtually any open-source bioinformatics tool available.

- A data management platform handling data access, metadata and how data can be accessed from local storage, public repositories or cloud storage.
- Providing users with access to a suitable computing environment with the bioinformatics software necessary to perform the desired data analysis.

A weakness with Galaxy is that the graphical user interface does not provide the flexibility and control desired by most bioinformaticians. This means that any Galaxy server requires dedicated staff scientists who provide training and manage the server to install new tools on behalf of the user. Thus, the efficiency gains from using networked Galaxy resources are reduced or lost as tools are configured for every specialized use, creating an environment with barriers to communication and productivity. This can be avoided by using containerized software in a computing environment shared between the Galaxy server and bioinformaticians using their own preferred workflow management software.

Software containers derive their name from the intermodal shipping container, which is nowadays a ubiquitous component of the international shipping system (Figure 1). Instead of requiring a specifically designed vehicle for each kind of goods transported, vehicles of today are often designed to carry a standardized shipping container. Likewise, software containers ensure that any containerized software has the environment and configuration necessary to function the same on different servers. With modern computers and software, pretty much any laboratory with bioinformaticians can host a data management platform and the software containers to analyse small datasets while access to computing power, data transfer capacity and working memory (RAM) limit the size of datasets which can be processed within a reasonable timeframe. Using software containers, it is thereby possible to fully test a workflow on a subset of data before applying for funding or access to computing power in a shared e-infrastructure for computing, and using the same workflow to analyze larger datasets.

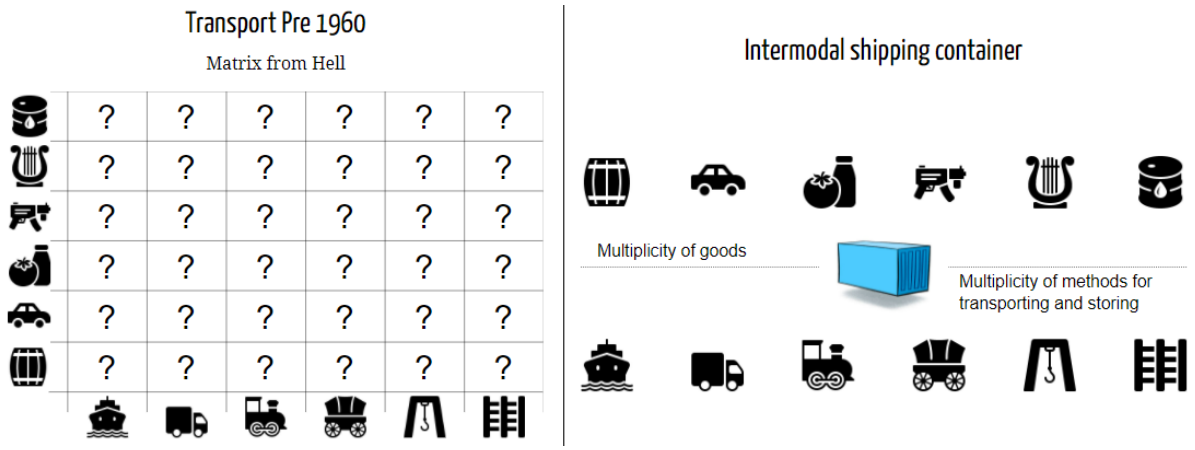

Figure 1. How the intermodal shipping container solved the matching matrix from hell in international shipping. An adaption from "Docker and Galaxy" by Bérénice Batut and Björn Grüning, licensed under CC BY-ND 4.0.

There are a limited number of researchers available for training and the large number of different specialisations within bioinformatics makes it hard for a single university to cover the width of applications for bioinformatics in current-day biology research. The containerized software model makes it easier to document and update your computing environment, and to accommodate workflows and training material developed by other labs.

The Galaxy Training Network is a network of researchers who all provide training material which can be used on many different Galaxy servers (Batut *et al*., 2018). By using software containers on the local server it is possible to keep local computing resources up to date and ensure that researchers can execute the workflows they've learned from the Galaxy Training Network.

The Galaxy platform is an excellent learning tool as it enables researchers to quickly start working with different tools and understand the different steps of data processing in a bioinformatics workflow. It is however not popular as an everyday tool for bioinformaticians developing workflows as it is inherently less flexible than scripting and command line tools. There is a plethora of options for bioinformaticians setting up their working environment but workflow managers such as Nextflow (Di Tommaso *et al*., 2017) are gaining popularity as they provide the flexibility of scripting while improving transparency and reusability of code.

Results

To accommodate these different requirements, we are testing a simple coding pattern to ensure that workflows developed in Nextflow can easily be transformed to run as workflows on Galaxy systems. By using the large repository of Biocontainers (da Veiga Leprevost *et al*., 2017) which contain container images for a vast number of commonly used bioinformatics tools and a 'one image per task' pattern, it is possible to ensure that workflows developed in Nextflow can also be run on a Galaxy server as both workflow managers can be configured to run Biocontainers on any computing environment that supports Docker or Apptainer (Singularity) containers. The Nextflow scripts can then easily be converted into Galaxy workflows and configured to submit jobs to the same scheduler on a computing cluster, meaning that actual computations will be run using the same software and compute nodes but jobs are defined and initiated by two different workflow management systems with different user requirements. This approach means that developed bioinformatics workflows can be quickly disseminated to other researchers as Galaxy workflows and training material be written as necessary to provide researchers with the knowledge and tools to perform their research.

Developing workflows with Nextflow we use four simple rules for coding (figure 2).

1. Data is stored on a file storage system. Nextflow is configured so that containers are mounted on the file storage area and can access, modify and store the files used and created by the workflow.
2. Containers are loaded from a registry (usually Biocontainers). Each container contains the software necessary to execute *one task* within the workflow (e.g. quality control, trimming, alignment, etc.).
3. Configuration scripts are set to let the workflow engine take the workflow code, pass the job to the right system and execute the workflow with output being stored on the storage area.
4. All written code is to be commented and committed to the SLU Global Bioinformatics Centre Github. This is where everything is documented so that workflows upon completion of a project can be implemented in Galaxy.

By adhering to these principles bioinformatics research can quickly be turned into reusable workflows which can be adapted and used within the many different projects carried out on different species and breeds of relevance to agricultural researchers.

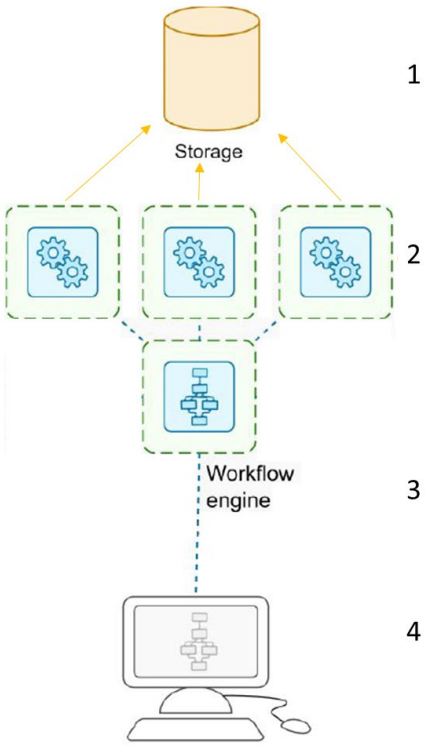

Figure 2. The Nextflow working environment. A workflow consists of an orchestrated and repeatable pattern of operations through which a piece of work is processed from initiation to completion. With the 'one image per task' pattern, a workflow engine like Galaxy or Nextflow mount a succession of software containers on the file storage which each execute the defined set of steps necessary to process the data (Spjuth *et al*., 2021).

Acknowledgments. The authors acknowledge the support of the Galaxy community, the Freiburg Galaxy Team and Björn Grüning, Bioinformatics, University of Freiburg (Germany) funded by the Collaborative Research Centre 992 Medical Epigenetics (DFG grant SFB 992/1 2012) and the German Federal Ministry of Education and Research BMBF grant 031 A538A de.NBI-RBC. The exchange project with the Agricultural Research Council in South Africa has been funded by a STINT Initiation grant for Joint South Africa-Sweden Research Collaboration (SA2018-7728)

References

Batut B., Hiltemann S., Bagnacani A., Baker D., Bhardwaj V., *et al*. (2018). Cell Systems 6, 752-758.e1. https://doi.org/10.1016/j.cels.2018.05.012

da Veiga Leprevost F., Grüning B.A., Alves Aflitos S., Röst H.L. (2017). Bioinformatics 33, 2580–2582. https://doi.org/10.1093/bioinformatics/btx192

Di Tommaso P., Chatzou M., Floden E.W., Barja P.P., Palumbo E., Notredame C. (2017). Nat Biotechnol 35, 316–319. https://doi.org/10.1038/nbt.3820

Jalili V., Afgan E., Gu Q., Clements D., Blankenberg D., Goecks J., et. al (2020). Nucleic Acids Res 48, W395–W402. https://doi.org/10.1093/nar/gkaa434

Klingström, T., (2017). Acta Univ. agric. Suec., Silv. 1652-6880 ; 2017:99 Sveriges lantbruksuniversitet., Uppsala.

Spjuth O., Capuccini M., Carone M., Larsson A., Schaal W., *et al*. F1000Research 2021, 10:513 https://doi.org/10.12688/f1000research.53698.1