Optimized Bioinformatics Workflows from Requirement Engineering of Solution Specifications

Jennifer Horkoff (jenho@chalmers.se), *Francisco de Oliveira Neto* (gomesf@chalmers.se) Assistant professors, SE Division. *Alexander Schliep* (alexander.schliep@cse.gu.se) Associate professor, Data Science Division. *Marcela Davila* (marcela.davila@gu.se), Head of the Bioinformatics Core Facility.

Problem/Motivation: Bioinformatics solutions are most often implemented as workflows, combining several different existing tools, which interoperate based on standard file formats, either using workflow systems (e.g., Galaxy[1]) or using scripting languages [1]. In other words, the workflows are composed by a sequence of steps, each with their own set of tools that produce outputs used by its following step. An example of a workflow to perform DNA Sequencing is presented in Figure 1.

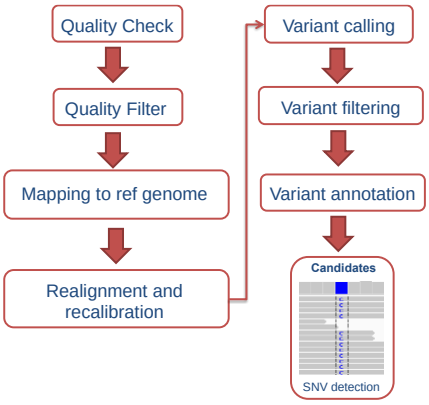

Figure 1- Example of a workflow for DNA Sequencing.

The distinct tool options used in a particular workflow step comprise different parameters, functions and performance. As an example, TrimGalore[2], PRINSEQ[3] and FastX[4] are different tools for DNA filtering. However, the choice between tools is mostly situational and convenience-based (e.g., a tool that "feels" more familiar), such that bioinformaticians do not exploit the benefits of the different tools in the workflow. In fact, one of our findings is that the choices of tools across the workflow is mostly arbitrary and can differ across people, labs, etc. The application of a workflow, as well as its connected artifacts, is supported by a lot of implicit, tacit information. As a consequence, most of the information is not clear for bioinformaticians. For instance, newcomers to a lab have to question colleagues in order to capture this implicit knowledge and apply to their own activities. Important elements such as the intent of the tasks, the requirements of the workflow and the explicit criteria for decision making are missing. Therefore, the workflows become sub-optimal, error-prone and easy to misinterpret. The challenge is capturing the variability and details of bioinformatics workflows in such a way as to facilitate updating, re-use, comparison, adaptation, and optimization.

By making intent and requirements explicit, through easier communication with non-bioinformaticians, is also likely to improve concordance between computational analysis and biological experiment. We overcome those challenges by bridging bioinformatics and software engineering (SE) while investigating the following research questions:

- **RQ1:** What are the defining and unique characteristics of bioinformatics workflows?
- **RQ2:** How can approaches for modeling and reasoning over intentions and variable workflows from Software Engineering be adapted to effectively capture bioinformatics workflows?
- **RQ3:** How can we use the captured workflows to optimize against developed benchmarks?

Methodology and Results: The initial plan was to perform five distinct tasks: 1) Adapt requirements engineering (RE) solutions, 2) Collect data from bioinformatics workflows, 3) Sketch bioinformatics workflow language. 4) Instrument testing frameworks, 5) Develop benchmarking tools. We collected data mainly through interviews with bioinformaticians during several iterations. However, the complexity and amount of information regarding the workflows, as well as the difficulties in capturing such knowledge using modelling notations, hindered our progress such that we could not complete activity 5. Additionally, activities 3 and 4 led to limited conclusions since the data collection revealed, respectively, significant limitations on i) current modelling notations to properly capture the workflow process, and ii) the technicalities of

[1] https://galaxyproject.org/
[2] https://github.com/FelixKrueger/TrimGalore
[3] http://prinseq.sourceforge.net/
[4] http://hannonlab.cshl.edu/fastx_toolkit/

instrumenting the different tools (e.g., different reporting templates) hindered their integration. In fact, the bioinformaticians asked us to prioritize the elicitation process over the tool instrumentation. On that note, those limitations yield research opportunities for the next steps of our research, where extension to current modelling notations and tool platforms can contribute to the fields of bioinformatics and software engineering.

For the sake of space, the deliverables of our project (models and templates) are presented in an online Appendix[5]. We began activities 1 and 2 with elicitation to model the. We tried several notations, such as Data-flow Diagrams, Sequence Diagrams, BPNM [6]and ultimately Activity Diagrams. As an answer to RQ1, we identified that the workflows are based on i) complex and repeated tasks, that can be decomposed in several smaller tasks; ii) many quality checks, using threshold values. Also, we identified the following bottlenecks: i) constant splitting of tasks between people and tools; ii) data emphasis where files are exchanged back and forth; iii) unclear motivation behind tasks.

Regarding RQ2, we identified that one modelling notations is unable to capture the complexity of the workflows. We should use different ones, or extend some of their elements. For instance, we propose extending activity diagrams in order to include motivations/goals and distinguish between sources, files, thresholds, etc. Another contribution is a draft template for bioinformatics workflow elicitation.

Answers to RQ3 are still on a higher level, since the priority in elucidating the workflow process, hindered actual integration of the toolchains to run the benchmark. These toolchains can be integrated into a workflow management system, which allows creation of multi-step computational analyses [2, 3]. Mainly, the workflows can be compared in terms of their data-flow, control-flow, service deployment and the different operations it supports (i.e., specific activities in the workflow) [2]. The service deployment is relevant since the use of Graphical User Interfaces (GUI), allows non-programmer researchers to use the workflow and operate tools. The control-flow can be correlated to the flow on the Activity diagrams, such that different tool dependencies can be mapped from the diagram. For instance, we identified that a FastQC program is always run, whereas a MultiQC program should only execute if there are more than one sample.

The models were insightful to the bioinformaticians. Their holistic view exposed bottlenecks in some of their decisions, such as which activities should be performed based on different threshold for quality checks on the dataset obtained from labs. Additionally, knowledge transfer between colleagues is now easier. From an SE perspective, we found that modelling notations are limited to express bioinformatic workflows from the data-flow (threshold, files, tools) and control-flow (activities, steps, decisions) perspectives. We believe that goal modelling can fill the gap from Activity Diagrams to represent motivations to the process.

Table 1- Preliminary financial summary.

Type of Costs	Values (SEK)
Salary	85183.00
Operation	2502.00
Local	6672.00
Indirect	30923.00
Total	**125,280.00**

Financial Outcome: The financial outcome of the project is shown in Table 1. These numbers were capture before the final costs are processed, since the project runs until the end of December. Particularly, the salary costs in Table 1 cover the researchers (Horkoff, de Oliveira Neto and Schliep). Costs due for December will also include a MSc. Student that worked 96 hours in this project.

Future Plans and Disseminations: Most of our future work is towards validating the proposed templates and new modelling languages for workflows. Ultimately, our aim is to develop a recommender's system based on rules and criteria to select steps and tools in a bioinformatic workflow. Additionally, we aim to integrate the tools and build a prototype where the different tools are regarding their unique features. The results from this project will be submitted to ER'19[7], while the ongoing investigation regarding the toolchain can be submitted to conferences such as ICST Tool Track[8]. Future funding submissions include calls from Vinnova, VR and Swedish Foundation for Strategic Research (SSF).

[5] https://goo.gl/adcu2z
[6] Business Process Model and Notation
[7] 38th International Conference on Conceptual Modeling: http://www.inf.ufrgs.br/er2019
[8] 12th International Conference on Software Testing http://icst2019.xjtu.edu.cn/Calls/Testing_Tool_Track.htm

References:

[1] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, and P. Li. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics, 20(17):3045–3054, 2004.

[2] V. Curcin and M. Ghanem, "Scientific workflow systems - can one size fit all?," 2008 Cairo International Biomedical Engineering Conference, Cairo, 2008, pp. 1-9. doi: 10.1109/CIBEC.2008.4786077

[3] M. Abouelhoda, S. Alaa, and M. Ghanem. Meta-workflows: pattern-based interoperability between Galaxy and Taverna. In Proceedings of the 1st International Workshop on Workflow Approaches to New Data-centric Science (Wands '10). ACM, New York, NY, USA, Article 2, 8 pages. DOI=http://dx.doi.org/10.1145/1833398.1833400