

A Modeling Approach for Bioinformatics Workflows: A Design Science Study

Laiz Heckmann Barbalho de Figueroa¹, Rema Salman¹, Jennifer Horko^{1,2}, Soni Chauhan¹, Marcela Davila³, Francisco Gomes de Oliveira Neto^{1,2}, and Alexander Schliep^{1,2}

¹ University of Gothenburg, Gothenburg, Sweden, ² Chalmers University of Technology, Gothenburg, Sweden, ³ Bioinformatics Core Facilities, Sahlgrenska Academy, University of Gothenburg, Sweden

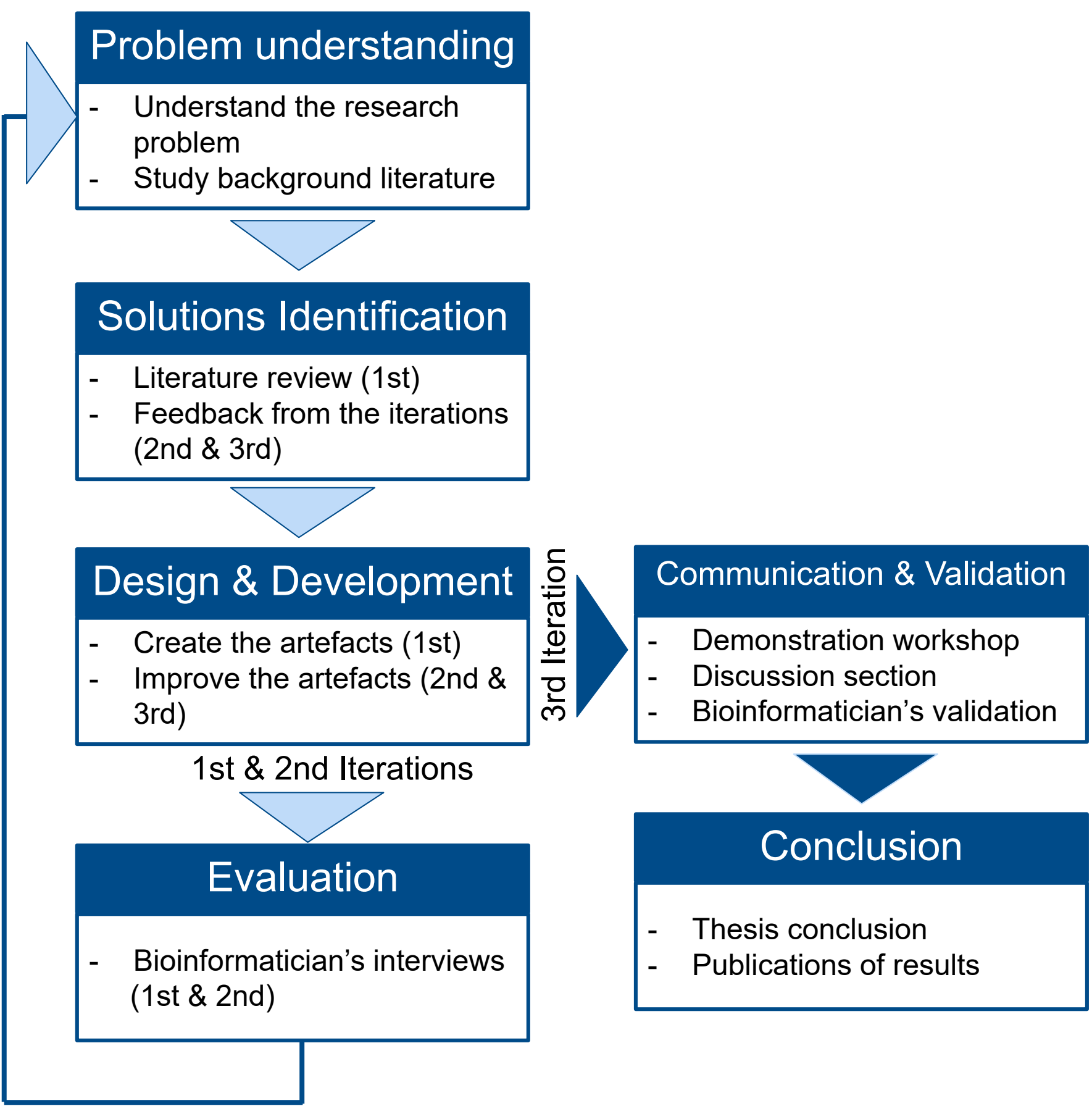
Modelling bioinformatics workflows

Bioinformaticians execute daily, complex, manual and scripted workflows to process data. There are many tools^{1,2} to manage and conduct these workflows, but there is no domain-specific way to textually and diagrammatically document them. This limit bioinformaticians and researchers to visualize, share, replicate and improve workflows.

Our aim is to extend the Unified Modeling Language (UML) Activity Diagram to the bioinformatics domain. This will ultimately aid in establishing a shared understanding and consistency between the activities (data processing) between bioinformaticians and researchers.

The process

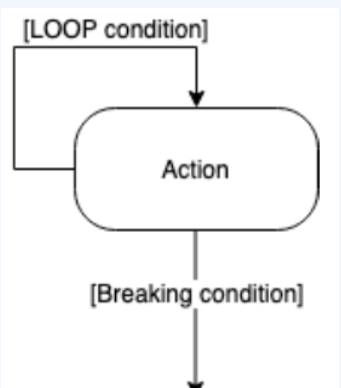
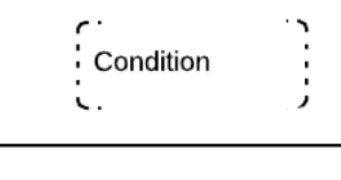
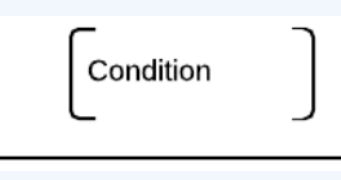
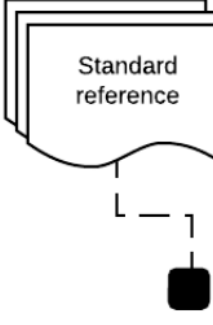
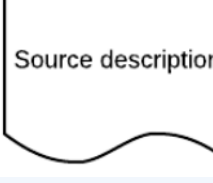


A Design Science Research Methodology (DSRM) approach was used to investigate the unique characteristics of bioinformatics workflows:



Seven bioinformaticians from three different Facilities^{3,4,5} participated in this study.

UML Activity Diagram

Unique characteristics of bioinformatics workflows were identified (*tool*, *tools settings and parameters*, *diagramSeparators* and *stantardReference*) and implemented in the UML AD extension model⁶. The final version of the concrete syntax contains 14 standards and 9 extended notations, for a total of 23 shapes (*see some examples below*). Despite of the high graphical complexity, the average understandability was 4.3 (out of 5).

Name	Base Class	Description	Notation
Loop	ActivityEdge	An iterative set of activities and actions until reaching the defined condition.	
Soft Condition	ActivityEdge	A condition with a limited soft-condition value, which is used for test outcomes. The condition is predefined within dashed guards on the outgoing edges.	
Hard Condition	ActivityEdge	A condition with a limited hard-condition value, which is used for test outcomes. The condition is predefined within solid guards on the outgoing edges.	
StandardReference	ObjectNode	Data, usually a standard, that are used for comparisons, such as the human genome.	
Source	ObjectNode	A link, document title, or person's name, which is the source for a specific set of actions.	
Tool	ObjectNode	An automatically operated tool or software used to perform an activity with its description.	
Tool	ObjectNode	A manually operated tool or software used to perform an activity with its description.	

Workflow Description Specification Template (WDST)

Bioinformaticians prefer diagrams over text. The proposed WDST (*see excerpt below*) can aid in knowledge sharing and documentation, however it requires refinement and automation, as its usability was evaluated negatively (1.3 out of 5) by the participants.

Workflow Description Specification			
Workflow ID:	<<the workflow name or identifier>>		
Date of creation:	<<data in which this document was created>>	Number of steps:	<<amount of steps>>
Workflow version:	<<version of this document>>	Modification date:	<<date of modification>>
		Workflow creators:	<<name>>
Workflow			
Workflow goal:	<<what do you want to achieve with this workflow?>>		
Workflow source:	<<is this workflow created locally or it follows a reference – in that case, add link to the reference or name the person>>		
Workflow responsible:	<<person who signs the final output or who uses this workflow>>		
First Step (start point)		Final Step (end point)	
Step ID: <<The name or identifier of the first step>>		Step ID: <<the name or identifier of the last step>>	

Future work

We will i) validate the concepts created with a broader bioinformatics community, ii) assess and improve workflows using the new language, iii) add workflow automation and iv) explore modeling tools and frameworks.

References

- 1) Fernando, T., et. al. Workflowds: Scalable workflow execution with provenance for data analysis applications. 2018
- 2) Roux-Rouquié, M., et. al. Using the unified modelling language (uml) to guide the systemic description of biological processes and systems. Biosystems. 2004
- 3) Bioinformatics Core Facility, Gothenburg, <https://cf.gu.se/english/bioinformatics>
- 4) Genomic Medicine Sweden, Gothenburg, <https://genomicmedicine.se/en/>
- 5) Translations! Genomics Plattform, Gothenburg, <https://wcmtm.gu.se/research-groups/genomics-platform> Guo, B., et al.
- 6) <http://www.cse.chalmers.se/~jenho/BioinformaticsWorkflows/>.

