A Literature Review of Bioinformatic Workflow Management Tools and Languages

Tra-Vaughn M.C. James
Earlham College
Richmond, Indiana
tjames19@earlham.edu

KEYWORDS

Bioinformatics, workflow, workflow manager, workflow management, workflow management tool, OpenWDL

ACM Reference Format:
Tra-Vaughn M.C. James. 2022. A Literature Review of Bioinformatic Workflow Management Tools and Languages. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Bioinformatics is an interdisciplinary field between biology and software, in which through various methods and techniques software is able to augment, and analyze biological data. In order to convert this data into useful information requires the use of a large number of tools, parameters, and dynamically changing reference data. As a result workflow managers such as Shake and OpenWDL were created to make these workflows scalable, repeatable and shareable. However, many of these workflow managers are bespoke in nature with many being only useful to specific sects of research. As bioinformatics moves along in development, computing resources become more powerful and readily available, and workflows become more complex, new workflow management tools attempt to address this problem through the application of various computing techniques with many achieving great success.

2 WORKFLOWS MANAGERS WITH A MACHINE LEARNING APPROACH

Other than OpenWDL various other workflow software exists, in which offer unique and sometime tailored approaches to their users.However, even more prominently within new workflow managers is the use of machine Learning to augment pipeline capabilities and streamline existing processes.BETSY presents the use of expert systems to address the growing complexity of bioinformatic workflows [2].Expert is composed of a "knowledge base of rules an inference engine that operates on the rules". Every rule dictated how a set of inputs converted to outputs, the engine then contemplates with rules to create a workflow that produces in the intended outcome.The Benefit of BETSY is its ability to create and alter workflows in a less hands on manner. A test was conducted in to see if the system can reproduce the published analysis of a landmark classification result. BETSY generated a "network that will produce classification predictions using the same algorithm (weighted voting") The resulting scores generated were found to be identical to the published scores.Two other manuscripts were tested, with BETSY producing an identical outcome to the published version as well.Unfortunately, frequent updating of the knowledge base cu ration is know needed potentially having an effect on existing workflows. Similarly SciPipe a workflow developed on flow-based programming (FBP) utilizes machine learning to improve the creation of their workflows [5].However SciPipe, Bioinformatic workflows, cheminformatics and many other. FBP allows for dynamic scheduling in which a program can run processes simultaneously and separately, thus schedule new tasks continually during the workflow run. Thus a process can be "created that obtains a value and passes it on to a downstream process as a parameter this allows for the Dynamic scheduling of tasks workflow.Through this and many other nuances SciPipe has achieved great success in the developing pipelines for various fields. Such as creating a pipeline for "complex dynamic workflows in machine learning for drug discovery applications", using the LIBLINEAR software molecules represented by the signature descriptor to train predictive models .The pipeline was constructed by creating components which are defined in separate files and are reusable in future workflows allowing for the creation of "audit logs for full workflow execution for the final output files and to display automatically plotted workflow graphs.This in comparison to SciLuigi a different workflow manager in which training had to be separated into different workflow files, causing the logging to become scattered into multiple log files.The use of Machine Learning algorithms assists in making workflow managers easily adaptable and scalable in nature, however, many require the use of a database and constant updates to them in order to keep there services primed.

3 TOOLS USED IN ADVANCING EXISTING WORKFLOW MANAGEMENT TECHNOLOGIES

While there exists a multitude of workflow management technologies there are many others used to advance and simplify existing ones. These tools provide necessary advancements to one seeking to better there workflow software without the need to embarking on a completely new product. One such toll is aCLImatise in which is designed to streamline the creation of new portable workflows by providing autmatically generated toll definitions for any tool

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

with a conventional command-line Interface [6].aCLImatise initially executes the command "of interest by trying a variety of help flags, storing the standard output from each". The results are then extracted using PEG (Parsing Expression Grammar). Finally the data model is outputted as a YAML data structure, or translated into CWL or WDL workflow formats. As tool definitions basically describe a piece of software and thus has no affect on the creation of a workflow.

4 WORKFLOW MANAGERS BUILT UPON EXISTING FRAMEWORKS

Many Bioinformatic workflows use and build upon existing workflow systems to address specific issues with the tool or to create a new workflow management system for a specific application.These workflow managers allow for the advancement of new workflow technologies while receiving the robust and time tested implementation of a preexisting workflow system.Bioshake is an Haskell Embedded Domain specific language it is built upon Shake a build tool implemented as an EDSL in Haskell [1]. inheriting Shakes "the reporting features, robust dependency tracking, and resumption capabilities. However unlike Shake, BioShake supports forward specification of workflows.One of the most important aspects of Bioshake is its ability to prevent errors before execution in which are caught by there type system. Moreover its interchangeability to use a different back end such as, Toil or Cromwell allowing for "the leverage of the cloud and containerisation facilities" of them both. While more specific to RNA-Seq experiments MIGNON uses WDL as underlying framework [3]. The steps of the workflow are "wrapped into WDL tasks that must executed on an independent unit of containerized software through the use of docker containers.Similar to Bioshake it can also be utilized with cloud based services like cromwell, but also with personal and HPC computers. To test is capabilities MIGNON was tested by 6 different human datasets (total of 42 samples) in which cromwell and docker coupling produced a fast and easy to deploy workflow.

5 CONCLUSION

Given the advancements within the feild and the usage of various new software techniques the workflow management world has definitely became more advanced, streamlined and resourcefull.However, with these new technologies, comes a new learning curve, providing another facet of difficulty to novice and experience Bioinformatic scientists. Using WDL as the basis I seek to develop a work flow management tool coupled with a GUI interface in will take the hassle of learning technologically challenging workflow managers this will allow bioinfomaticians to focus on the analysis and comprehension of their data while having an easy to use tool in which can automate the workflows needed to dissect their data. My implementation will be ambiguous in nature allowing for the development of various Bioinformatics workflows. [4] [7]

REFERENCES

[1] Justin Bedő. 2019. BioShake: a Haskell EDSL for bioinformatics workflows. *PeerJ* 7 (2019), e7223.

[2] Xiaoling Chen and Jeffrey T Chang. 2017. Planning bioinformatics workflows using an expert system. *Bioinformatics* 33, 8 (2017), 1210–1215.

[3] Martín Garrido-Rodriguez, Daniel Lopez-Lopez, Francisco M Ortuno, María Peña-Chilet, Eduardo Muñoz, Marco A Calzado, and Joaquin Dopazo. 2021. A versatile workflow to integrate RNA-seq genomic and transcriptomic data into mechanistic models of signaling pathways. *PLoS computational biology* 17, 2 (2021), e1008748.

[4] Michael Jackson, Edward Wallace, and Kostas Kavoussanakis. 2020. Using rapid prototyping to choose a bioinformatics workflow management system. *bioRxiv* (2020).

[5] Samuel Lampa, Martin Dahlö, Jonathan Alvarsson, and Ola Spjuth. 2019. SciPipe: A workflow library for agile development of complex and dynamic bioinformatics pipelines. *GigaScience* 8, 5 (2019), giz044.

[6] Michael Milton and Natalie Thorne. 2020. aCLImatise: automated generation of tool definitions for bioinformatics workflows. *Bioinformatics* 36, 22-23 (2020), 5556–5557.

[7] Laura Wratten, Andreas Wilm, and Jonathan Göke. 2021. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature methods* 18, 10 (2021), 1161–1168.

A ONLINE RESOURCES

Nam id fermentum dui. Suspendisse sagittis tortor a nulla mollis, in pulvinar ex pretium. Sed interdum orci quis metus euismod, et sagittis enim maximus. Vestibulum gravida massa ut felis suscipit congue. Quisque mattis elit a risus ultrices commodo venenatis eget dui. Etiam sagittis eleifend elementum.

Nam interdum magna at lectus dignissim, ac dignissim lorem rhoncus. Maecenas eu arcu ac neque placerat aliquam. Nunc pulvinar massa et mattis lacinia.