Graph Theory Approaches for Optimizing Biomedical Data Analysis Using Reproducible Workflows

Gaurav Kaushik[†1], Sinisa Ivkovic[1], Janko Simonovic[1], Nebojsa Tijanic[1], Brandi Davis-Dusenbery[1], Deniz Kural[1]

[1]*Seven Bridges Genomics*
1 Main Street, Cambridge, MA 02140, USA

As biomedical data becomes increasingly easy to generate in large quantities, the methods used to analyze it have proliferated rapidly. However, for the insights gained from these analyses to be meaningful, the analysis methods themselves must be transparent and reproducible. To address this issue, numerous groups have developed workflow standards or specifications that provide a common framework with which to describe a given analysis method. The diversity of methods demands that the specification be robust and flexible to accurately describe a specific biomedical data analysis. However, a powerful specification alone is insufficient to drive reproducible analysis – even the most completely described workflow must also be 'runnable' on diverse architectures. Thus the complete reproducible workflow ecosystem includes one or more well defined workflow definitions or specifications as well as the software components needed to implement these specifications. Such implementations allow adaptation to diverse environments, provide optimizations to workflow execution, improve computing efficiency, and support reproducibility through task logging and provenance. To meet these requirements, we have developed the Rabix Executor, an open-source workflow engine which utilizes graph theory approaches to enable computational optimization of complex, dynamic workflows, and supports reproducibility in biomedical data analysis.

1. Introduction

In recent years, the concept of organizing data analysis via computational workflows and accompanying workflow description languages has surged in popularity as a way to support the reproducible analysis of massive genomics datasets.[1,2] Robust and reliable workflow systems share three key properties: flexibility, portability, and reproducibility. Flexibility can be defined as the ability to gracefully handle large volumes of data with multiple formats. Adopting flexibility as a design principle for workflows ensures that multiple versions of a workflow are not required for different datasets and a single workflow or pipeline can be applied in many use cases. Together, these properties reduce the software engineering burden accompanying large-scale data analysis. Portability, or the ability to execute analyses in multiple environments, grants researchers the ability to access additional computational resources with which to analyze their data. For example, workflows highly customized for a particular infrastructure make it challenging to port analyses to other environments and thus scale or collaborate with other researchers. Well-designed workflow systems must also support reproducibility in science. In the context of workflow execution, computational reproducibility can be simply defined as the ability to achieve the same results on the same data regardless of the computing environment or when the analysis is performed. Workflows and the languages that describe them must account for the complexity of

* This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN261201400008C.
† Corresponding author

the information being generated from biological samples and the variation in the computational space in which they are employed. Without flexible, portable, and reproducible workflows, the ability for massive and collaborative genomics projects to arrive at synonymous or agreeable results is limited.[3, 4]

Computational workflows may consist of dozens of tools with hundreds of parameters to handle a variety of use cases and data types. They can be made more flexible by allowing for transformations on inputs during execution or incorporating metadata, such as sample type or reference genome, into the execution. They can allow for handling many use cases, such as dynamically generating the appropriate command based on file type or size, without needing to modify the workflow description to adjust for edge cases. Such design approaches are advantageous as they alleviate the software engineering burden and thus the accompanying entropy or probability of error associated with executing extremely complex workflows on large volumes of data. However, as the complexity of an individual workflow increases to handle a variety of use cases or criteria, it becomes more challenging to optimally compute with it. For example, analyses may incorporate nested workflows, business logic, memorization or the ability to restart failed workflows, or require parsing of metadata -- all of which compound the challenges in optimizing workflow execution.

As a result of the increasing volume of biomedical data, analytical complexity, and the scale of collaborative initiatives focused on data analysis, reliable and reproducible analysis of biomedical data has become a significant concern. Workflow descriptions and the engines that interpret and execute them must be able to support a plethora of computational environments and ensure reproducibility and efficiency while operating across them. It is for this reason that we have developed the Rabix Executor[a], an open-source workflow engine that can interpret standard workflow description languages and apply methods based in graph theory to enable unprecedented computational optimization of even the most complex workflows while supporting reproducibility in science.

2. Rabix uses a modified composite pattern model for workflow execution

The Rabix Executor allows users to execute applications described by a workflow description language. First, the workflow description is submitted through a frontend or interface. Then, the Rabix engine interprets the workflow description and translates it into discrete computational processes or "jobs." Finally, the jobs are queued to a backend or computational infrastructure, such as a local machine, cluster, or cloud instances, for scheduling and execution. Each component of the executor (frontend bindings, engine, queue, backend bindings) is abstracted from each other to enable complete modularity. Developers are able to design custom frontends (e.g. command line or graphical user interface), bindings for the engine to parse different workflow languages, use the queuing protocol of their choice, and submit computational jobs to different

[a]The Rabix Executor is available on GitHub: http://github.com/rabix/bunny

backends. This flexible software model means that Rabix can be modified to perform data analysis on many different infrastructures as desired by the user or developer and achieve identical results.

3. Graph representation of data analysis workflows

Abstractly, computational workflows may be understood as a directed acyclic graph (DAG)[2,5,6], a kind of finite graph which contains no cycles and which must be traversed in a specific direction. In this representation, each node is an individual executable command. The edges in the DAG represent execution variables (data elements such as files or parameters) which pass from upstream nodes to downstream ones.

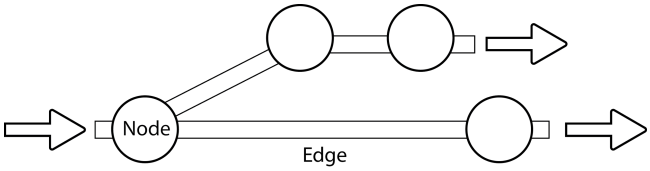


Figure 1. Illustration of a directed acyclic graph (DAG). The DAG may be traversed from left-to-right, moving from node-to-node along the edges that connect them.

In practice, workflows are described with machine-readable serialized data objects in either a general-purpose programming language (GPL), domain-specific language (DSL), or serialized object models for workflow description.[2,7] For example, an object model-based approach may describe the steps in a workflow in JSON format with a custom syntax. This workflow description can then be parsed by an engine or executor to create the DAG representation of the workflow. The executor may then translate the directions for workflow execution to actionable jobs in which data is analyzed on a computational infrastructure or backend.

3.1. *General structure of a workflow execution*

There are three general steps in workflow execution: interpretation of a machine-readable workflow description, generation of the workflow DAG, and finally decomposition into individual jobs that can be scheduled for execution. At the beginning of execution, a workflow engine or interpreter is provided with the workflow description and the required inputs for execution of the workflow, such as parameters and file paths (Fig. 2a). The workflow description object is then parsed and a DAG is created (Fig. 2b), which contains the most minimal set of nodes and edges required for computation.

In addition to representing the steps in the workflow as a DAG (Fig. 2c), many workflow ontologies model computational jobs as a composite (tree) pattern in which there are “parent nodes” (workflows), which can contain multiple executables or “leaf nodes” (executables) (Fig. 2d).[8,9,10] The Rabix engine extends this model by generalizing “parent” nodes to include transformations of the DAG, such as when parallelizations are possible at that node. The engine handles the "execution" or parsing of these parent jobs, while leaves are queued for scheduling and execution on a backend. This model allows for more efficient resolution of DAG features such as

nodes in which steps can be parallelized or are nested. Thus, the burden of flattening a workflow is moved from the workflow developer to the engine itself.

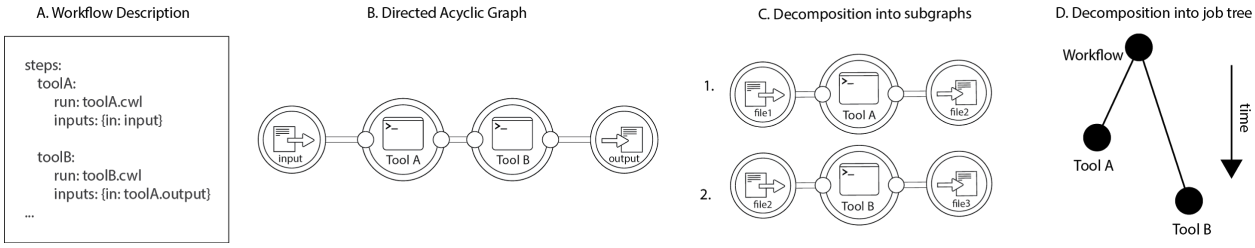


Figure 2. The process of parsing a workflow description. **A.** The machine-readable document is interpreted, from which **B.** a DAG is produced. From the DAG, **C.** subgraphs representing computational jobs that can be sent to backends for scheduling/execution and **D.** a job tree is resolved, which identifies "parent" and "leaf" nodes. Each leaf represents an individual job.

3.2. *Optimization of workflow executions through graph transformations*

After the minimal DAG is evaluated from a workflow description, it may be possible to perform transformations of the DAG to enable further decomposition, thereby allowing additional opportunities for computational optimization. A subset of DAGs may be completely flattened at runtime. In such cases, all possible parameters and inputs are known at runtime and there are no possible modifications to the DAG. Such workflows are synonymous with explicitly declarative workflow descriptions, which do not allow modifications to data structures in the workflow during runtime.[11] However, modern workflow description languages often allow for external transformations of the data elements in the DAG to enable more flexibility in the workflow description. Due to these external transformations, additional edges and nodes could be created in the DAG during execution, which in turn could be decomposed into additional jobs for optimization. However, it is difficult to do this generally. If an engine is capable of performing decomposition of jobs even when the workflow has external transformations, further optimizations can be made to reduce cost or time of an analysis.

4. Algorithm for execution optimization via graph transformations

In the following sections, we discuss workflows specifically described with the Common Workflow Language (CWL), an emerging standard for describing tools and workflows that has found a significant audience in the bioinformatics and biomedical imaging communities.[1]

The Common Workflow Language is used to describe individual "applications," which can be either a single tool or an entire workflow. Workflows are described as a series of "steps," each of which is a single tool or another, previously-described workflow. Each step in the workflow has a set of "ports" which represent data elements that are either inputs or outputs of the tool. A single port represents a specific data element that is required for execution of the tool or is the result of its execution. For data elements which are passed between applications, there must be an output port from the upstream tool and a complementary input port on the downstream application.

When a CWL workflow is represented as a DAG, applications become nodes and edges indicate the flow of data elements between ports of linked tools. In the case of a simple workflow, there are no possible transformations of the DAG; each node represents a single command line execution and all data elements are simply passed from tool-to-tool as-is (Fig. 3).

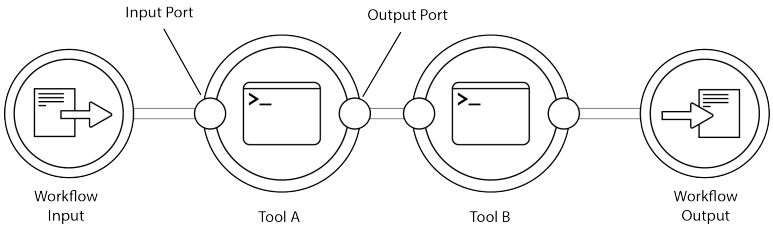


Figure 3. A DAG created from a workflow described by the Common Workflow Language which contains two tools (**A, B**). Tools have input and output ports, which define discrete data elements that are passed downstream along the edges of the DAG.

Additionally, CWL workflows can be designed such that data elements and the execution itself can be transformed during runtime. Developers are given several options for describing workflows which can enhance their utility and flexibility in handling biomedical data analysis:

1. The ability to generate "dynamic expressions" or transformations on data elements, inputs, outputs, and other command line arguments.
2. The ability to perform "scatter/gather I/O (input/output)", also known as vectored I/O, in which execution of the input data can be parallelized based on specific criteria. A common genomics use case for this is performing an analysis per chromosome, in which the set of chromosomes is delivered to a node as an array (e.g. [1, 2, 3, X]).
3. The ability to nest workflows within workflows, which allows for rapid composition of complex workflows and the ability to quickly reuse existing code.

Though CWL allows such descriptions, the exact way in which these features are implemented is left entirely to the execution engine that is interpreting it. Therefore, the Rabix engine has been designed to handle CWL descriptions with two additional advances:

1. The ability to enable graph transformations at runtime to attempt to fully flatten the DAG.
2. The ability to further flatten the DAG during execution by inspecting pending executions, referred to as the "look-ahead" method.

These functionalities enable the Rabix engine to create additional edges and nodes as needed, in order to more fully decompose the workflow DAG. In this way, more choices are available for computational optimization of dynamic or transformable workflows.

4.1 *Algorithm and data model for workflow execution*

The engine is capable of flattening a workflow DAG to identify all possible jobs and enable additional optimization. Concretely, the workflow DAG is stored in three tables, Variables, Jobs,

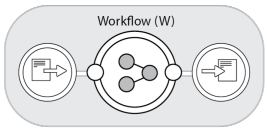
and Links, which are evaluated at various stages during traversal of the DAG during execution. The Variables table contains the ports and their explicit values. The Jobs table stores each node of the workflow and a counter for the inputs and outputs that have been evaluated at that node. The Links table stores the edges in the DAG that is traversed.

Suppose for example, Rabix is executing the workflow in Figure 4. The engine will first parse the workflow description as a workflow DAG with two variables (W.I, W.O; Fig. 4a), which are yet to be evaluated. Additionally, there are two ports (#In, #Out), an input and an output. Next, the engine inspects the contents of the workflow (Fig. 4b) and is able to see the following steps: Tool A, Tool B, each of their ports, and the link between each step within scope.

After this, any known values are curried downstream through their links. The input for the workflow (W.I) is curried to Tool A through the link that has been identified between the two (W.I → W.I.A). The input job counter (#In) for Tool A is decremented to 0, thereby triggering an input event where a job (execution of Tool A with value1) is distributed to a backend for computation. The engine now waits for an event in which the output of Tool A (W.A.O) is reported as value.

Once the output for the job is evaluated and reported to the engine (value2), an output event is triggered. The output port for W.A is decremented to 0, the link from W.A.O to W.B.I is traversed, and W.B.I is evaluated as value2. This reduces the #In counter for W.B to 0 in the Jobs table and triggers a job, the execution of Tool B with its input (Fig. 4c). The execution finally concludes until the input port counter for W reaches 0 and W.O is evaluated (Fig. 4d).

A

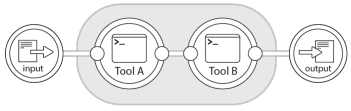


VARIABLES	Values
W.I	value1
W.O	

JOBS	# In	# Out
W	1	1

LINKS

B

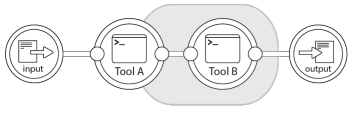


VARIABLES	Values
W.I	value1
W.O	
W.A.I	value1
W.A.O	
W.B.I	
W.B.O	

JOBS	# In	# Out
W	0	1
W.A	0	1
W.B	1	1

LINKS
W.I → W.A.I
W.A.O → W.B.I
W.B.O → W.O

C

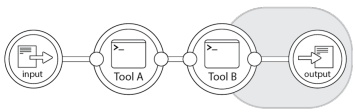


VARIABLES	Values
W.I	value1
W.O	
W.A.I	value1
W.A.O	value2
W.B.I	value2
W.B.O	

JOBS	# In	# Out
W	0	1
W.A	0	0
W.B	0	1

LINKS
W.I → W.A.I
W.A.O → W.B.I
W.B.O → W.O

D



VARIABLES	Values
W.I	value1
W.O	value3
W.A.I	value1
W.A.O	value2
W.B.I	value2
W.B.O	value3

JOBS	# In	# Out
W	0	0
W.A	0	0
W.B	0	0

LINKS
W.I → W.A.I
W.A.O → W.B.I
W.B.O → W.O

Figure 4. The algorithm which enables flattening of a workflow DAG as it is traversed. **A.** The engine interprets the top-level of the workflow description and **B.** inspects the contents of the workflow node and determines the DAG structure and links between each step (edges). The currying of value1 from the workflow input to the input of Tool A triggers an input event, where a job (analysis of Tool A with its inputs) is sent to a backend node. **C.** The execution continues and the engine traverses the DAG. **D.** The workflow is completed when the output of the final tool (W.B.O., value3) is curried to the overall workflow output (W.O). The port counters allow the engine to track when nodes are ready to be executed.

In the case where the engine is traversing a portion of the workflow that maps to a parent node beneath the root parent node, each output update event will generate an additional output update event. This strategy allows the engine to "look-ahead" towards future executions and apply optimizations to dynamic portions of the DAG, as outlined in the following sections.

4.2. *Graph transformations: enabling parallelization with scatter/gather*

By evaluating workflows through this counter and trigger system, Rabix is capable of flattening parallelizable nodes in the DAG. Suppose we have a workflow where a data file and an array are inputs for a single tool, which then produces an output file (Fig. 5a). In this case, the tool is capable of being scattered over an array of variables (e.g. [1, 2, 3]). Normally, these executions will be performed sequentially on a single core, or on multiple threads if the tool allows it. However, on a workflow level, Rabix can enable additional parallelization by scattering the data over three separate executions of the tool based on the values in the array (Fig. 5b), thus allowing the jobs to be distributed to separate computational instances as needed.

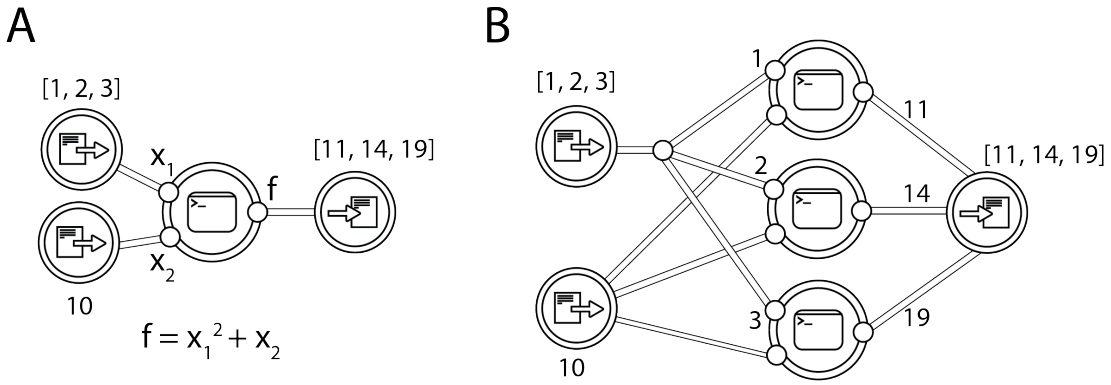


Figure 5. Graph transformations when performing parallelization. In this workflow, a function is performed on two inputs, an *int* and an *array of ints*. **B.** The flattened DAG created by the engine. Each value of the array is scattered as a single process to reduce computation time.

The advantages of this approach is further demonstrated by another use case, in which there are two sequential, parallelizable jobs (Fig. 6a). Rabix employs a "look-ahead" strategy (Fig. 6b) which can mark downstream jobs as ready even though not all sub-jobs (leaves) are done from the upstream parent job.

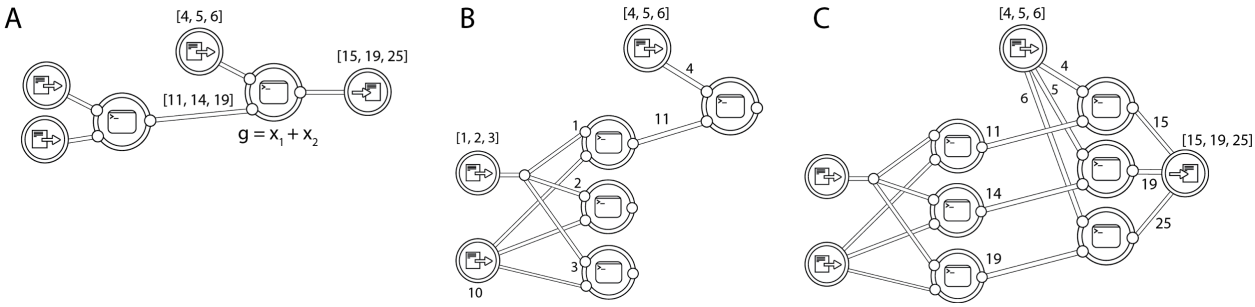


Figure 6. Graph transformations for sequential parallelization. **A.** The workflow from Fig. 5 with an additional downstream function with an input that can be scattered. **B.** During execution, the engine is able to look ahead to the next stage in the workflow. If any input is available (e.g. value of 11 returned by a tool), downstream processes which can proceed are started. **C.** The completed workflow.

Each individual node in the DAG does not need to be scheduled individually. Instead, the engine can create "job groups" from a flattened DAG (Fig. 7). For example, in the case of executions scattered across chromosome number, job groups of multiple chromosomes can be distributed to the same node to optimize cost.

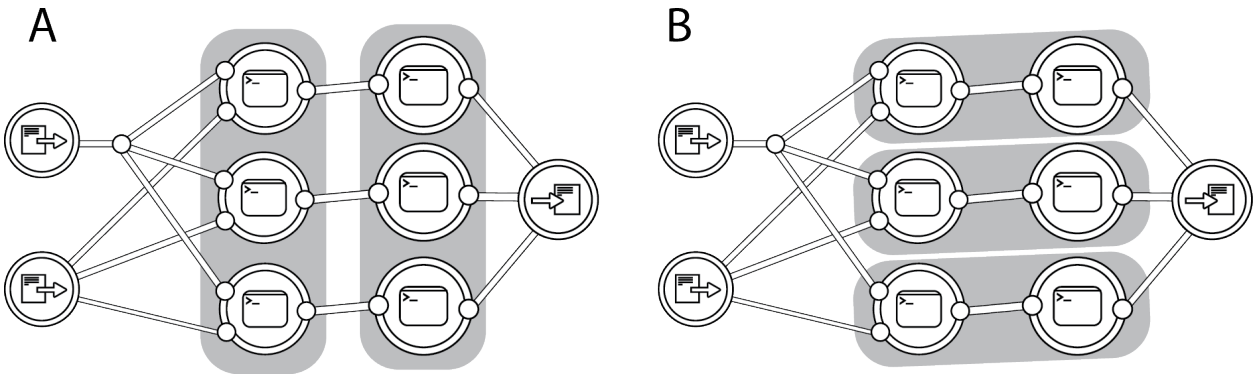


Figure 7. Jobs can be grouped (grey background) for execution on a backend node based on specific criteria set by the workflow or tool author.

Figure 7 demonstrates two possible job group assignments. In the case of Figure 7a, the first tool can be executed simultaneously for each chunk of the data on a single backend node. Once any single job in the first group is finished, the second group of jobs can begin execution on a second node. In the case of Figure 7b, each chunk of data is parallelized across three nodes and the final output is gathered at the end.

The engine is also able to send information to the backends about upcoming jobs. This allows a backend scheduler to pre-allocate resources for upcoming jobs.

4.3. *Graph transformations: flattening nested workflows*

CWL developers have the ability to reuse existing code and import previously-described workflows into other workflows. This feature means that it is possible to reuse code for additional workflows in lieu of refactoring and potentially introducing errors that break reproducibility. However, the ability to nest workflows presents a challenge to interpretation and optimization by

the engine. If no DAG transformations are applied and nested workflows are only executed recursively, this can lead to unnecessarily prolonged execution time and cost.

Suppose a developer has described a workflow that takes two inputs and produces two outputs from two tools (Fig. 8a). In this workflow, one of the outputs is created by the upstream tool and one from the downstream tool. Later, the developer wishes to reuse this workflow description in another workflow, where the output of the upstream tool is passed to another tool for further analysis (Fig. 8b). As with sequentially scattered tools, the engine is capable of passing values from the nested workflow, once they're produced, to steps downstream using the "look-ahead" strategy. Commonly, the tool outside the nested workflow is blocked from execution until all outputs from the nested workflow are produced, leading to increased computation time and cost.

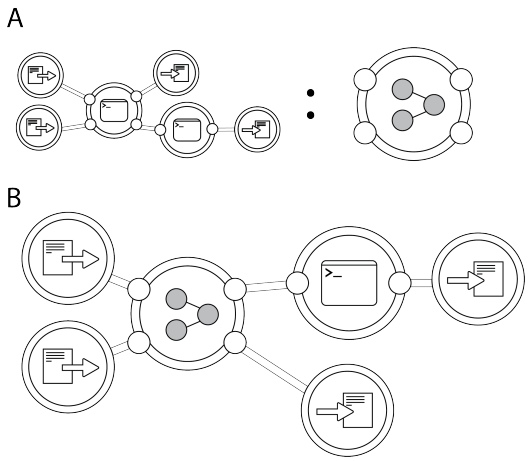


Figure 8. Graph transformations enable flattening of nested workflows to optimize total execution time. **A.** Workflow consisting of two tools. **B.** Workflow in Fig. 8a. extended with third tool. The engine allows the downstream tool to start executing once the necessary inputs are ready, even if the upstream workflow has yet to produce all of its outputs. No code refactoring from the workflow in 8a is required.

4.4. Benefits to Orchestration and Computation

The model used by the Rabix engine allows for improved optimization of data analysis at the workflow level. Further, it provides the ability to implement additional optimizations or features to enhance orchestration of jobs and computation, regardless of whether such features are supported by a workflow description language or specification.

Rabix keeps track of all jobs executed from the workflow and caches results. If the workflow contains a job that has previously been executed and the outputs are still available, the engine can reuse them even if the job was part of a different workflow run. Importantly, even if cached results are not available, the engine will look ahead in the DAG and may encounter cached downstream jobs which do have these files available, and so can resume failed or modified workflow jobs. This makes the caching mechanism comparable to declarative workflow description such as GNU Make.[11]

Additional business logic outside of a workflow specification can also be implemented. For example, CWL does not yet allow for conditional workflows, in which the entirety of DAG is not necessarily traversed but only paths based on checkpoints during the execution. Additionally, though a DAG is acyclic, Rabix can enable loops for a tool or workflow which use iterative operations.

4.5. *Caveats to Graph Transformations and Possible Solutions*

An important caveat for flattening the DAG are external transformations in which the structure of data elements is modified before execution and thus cannot be anticipated by the engine. For example, CWL and other workflow description languages allow for modifications of input types before tool execution. In certain cases, such as for a tool which can be scattered, the data type may change or the length of the array that is being scattered cannot be known ahead of time. If the engine is unable to anticipate the length of an array that must be scattered upon execution, it is impossible for it to flatten the DAG before evaluation. However, such hurdles can be overcome by allowing users to define external transformations, such as array lengths, ahead of time to circumvent this issue. For example, if two sequential tools expect an array of length *n*, the engine can optimize the parallelization of these tools if the user defines the same array as the input for both.

4.6. *Furthering reproducibility by extending CWL to execution descriptions*

Workflows described using the Common Workflow Language require two objects for execution: the description of an application and an input object specifying the explicit values of the required inputs. Recording a task that has been previously executed is not, however, within the scope of CWL. However, an analyst may want to reinspect a prior analysis, reuse a workflow with a specific set of parameters on new data, or reanalyze the same data with a different workflow version. It is for these reasons that we have enabled an additional layer of task description and annotation within Rabix, alleviating the burden of logging the workflow execution.

Following the execution of a workflow, additional outputs and logs are produced by Rabix as a matter of course. The explicit command line execution, an object describing the output of the execution, and a description of the workflow execution are all recorded. From these objects, it is directly possible to reproduce a prior analysis or reanalyze additional data with the exact same parameters as previous. Rabix allows for replication of a previous execution or reproduction an exact workflow on new data with a single command line call. In this way, it is possible for an analyst to not only publish a workflow but also the explicit tasks as plain text files. These functionalities can be extended with new modules or plugins to enable a variety of use cases centered on reproducibility.

5. Conclusions

The Rabix executor suite is an open-source project designed to enable scalable and reproducible analysis of portable workflows. Computational reproducibility, the ability to replicate a prior analysis or reuse prior workflows on new data, is required for accurately judging scientific claims or enabling large-scale data analysis initiatives in which synonymous results can be compared.[2,12,13] The Rabix engine additionally aims to optimize workflow executions by intelligently interpreting and handling complex workflows. This is achieved through a composite model in which workflows can be more fully decomposed. Finally, additional logic can be applied to optimize for user-defined variables, such as cost or execution time, regardless of the workflow description language being interpreted.

References

1. Amstutz, Peter; Crusoe, Michael R.; Tijanić, Nebojša; Chapman, Brad; Chilton, John; Heuer, Michael; Kartashov, Andrey; Leehr, Dan; Ménager, Hervé; Nedeljkovich, Maya; Scales, Matt; Soiland-Reyes, Stian; Stojanovic, Luka (2016): Common Workflow Language, v1.0. Figshare. https://dx.doi.org/10.6084/m9.figshare.3115156.v2
2. Leipzig, Jeremy. "A review of bioinformatic pipeline frameworks." *Briefings in bioinformatics* (2016): bbw020.
3. Kanwal, Sehrish et al. "Challenges of Large-Scale Biomedical Workflows on the Cloud--A Case Study on the Need for Reproducibility of Results." *2015 IEEE 28th International Symposium on Computer-Based Medical Systems* 22 Jun. 2015: 220-225.
4. Alioto, Tyler S et al. "A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing." *Nature communications* 6 (2015).
5. Deelman, Ewa et al. "Pegasus, a workflow management system for science automation." *Future Generation Computer Systems* 46 (2015): 17-35.
6. Guo, Fengyu et al. "A workflow task scheduling algorithm based on the resources' fuzzy clustering in cloud computing environment." *International Journal of Communication Systems* 28.6 (2015): 1053-1067.
7. Köster, Johannes, and Sven Rahmann. "Snakemake—a scalable bioinformatics workflow engine." *Bioinformatics* 28.19 (2012): 2520-2522.
8. Belhajjame, Khalid et al. "Using a suite of ontologies for preserving workflow-centric research objects." *Web Semantics: Science, Services and Agents on the World Wide Web* 32 (2015): 16-42.
9. Terstyanszky, Gabor et al. "Enabling scientific workflow sharing through coarse-grained interoperability." *Future Generation Computer Systems* 37 (2014): 46-59.
10. Gamma, Erich. *Design patterns: elements of reusable object-oriented software*. Pearson Education India, 1995.
11. Stallman, Richard M, Roland McGrath, and Paul D Smith. *GNU Make: A program for directing recompilation, for version 3.81*. Free Software Foundation, 2004.

12. Peng, Roger D. "Reproducible research in computational science." *Science* 334.6060 (2011): 1226-1227.
13. Sandve, Geir Kjetil et al. "Ten simple rules for reproducible computational research." *PLoS Comput Biol* 9.10 (2013): e1003285.