Demystifying Bioinformatics Pipelines

Enthought
6,108 followers

September 25, 2023

What is a Bioinformatics Pipeline?

Bioinformatics—used extensively in genomics, pathology, and drug discovery—combines mathematical and computational methods to collect, classify, store, and analyze large and complex biological data. The set of biological data analysis operations executed in a predefined order is commonly referred to as a "bioinformatics pipeline". In other words, **a bioinformatics pipeline is an analysis workflow that takes input data files in unprocessed raw form through a series of transformations to produce output data in a human-interpretable form.**

Typically, a bioinformatics pipeline consists of four components: 1) a user interface; 2) a core workflow framework; 3) input and output data; and 4) downstream scientific insights.

The core framework contains a variety of third-party software tools and in-house scripts wrapped into specific workflow steps. The steps are executed in a particular environment via a user interface, taking raw experimental data, reference files, and metadata as inputs. The resulting output data is then used to drive scientific insights through downstream advanced results, visualization, and interpretation.

Bioinformatics Pipelines in R&D

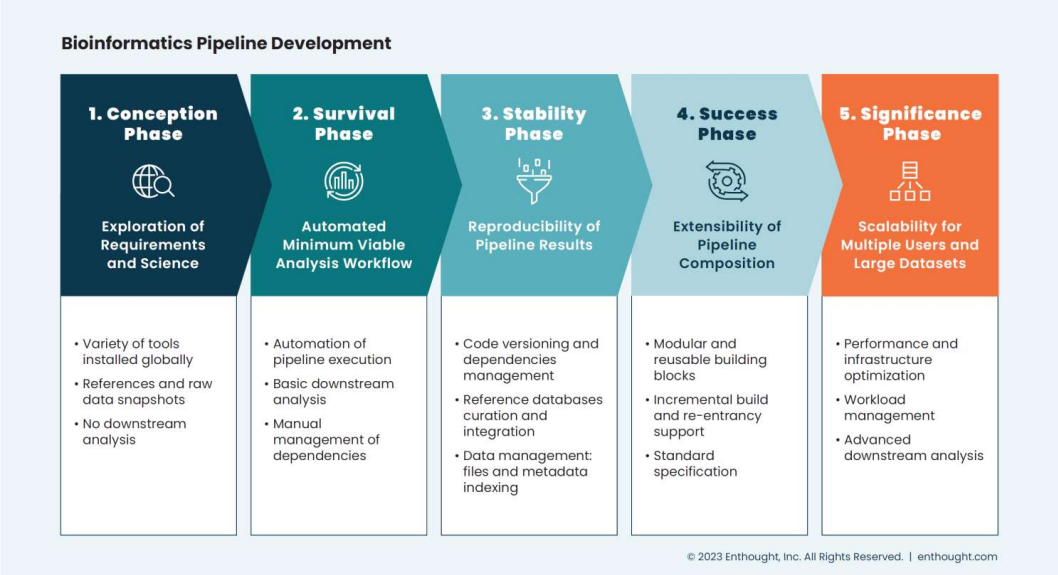
Despite the existence of highly sophisticated data pipelines, there is a frequent need in R&D to create ad-hoc bioinformatics pipelines either for prototyping and proof-of-concept purposes or to integrate newly published tools and methods since existing pipelines are not easily customizable. As the pipelines grow with additional steps, managing and maintaining the necessary tools becomes more difficult. Moreover, the complex and rapidly changing nature of biological databases, experimental techniques, and analysis tools makes reproducing, extending, and scaling pipelines a significant challenge.

Scientists, bioinformaticians, and lab managers are tasked with designing their pipelines and identifying the gaps within their frameworks. The best approach to prioritizing efforts depends highly on the operational need, the scientific scope, and the state of the bioinformatic pipeline.

The very first step, however, is to understand the evolution of a bioinformatics pipeline.

The 5 Phases of a Bioinformatics Pipeline

A bioinformatics pipeline evolves through five phases. Pipeline stakeholders first seek to explore and collect the essential components, including raw data, tools, and references (Conception Phase). Then, they automate the analysis steps and investigate pipeline results (Survival Phase). Once satisfied, they move on to seek reproducibility and robustness (Stability Phase), extensibility (Success Phase), and finally scalability (Significance Phase). Below is a figure of the evolution at-a-glance as well as a description of each phase.



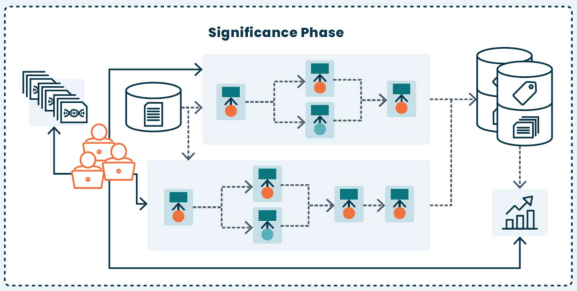
1. **Conception Phase: Exploration of Requirements and Science** | The Conception Phase focuses on the basic

needs of a new bioinformatics pipeline. It consists of exploring all the dependencies (software tools and reference databases), assessing adequate input data and metadata, and running the different steps individually. Building a minimum viable pipeline requires selecting the proper processing steps to run in the correct order and get the appropriate outputs.

2. **Survival Phase: Automated Minimum Viable Analysis Workflow |** The Survival Phase focuses on having an automated minimum viable analysis workflow. The process of analysis automation consists of streamlining the data processing steps to build a consistent analysis pipeline after setting the needed dependencies and reference databases. It ensures that the pipeline can be run as one transaction using a single entry point, which simplifies the analysis execution and the pipeline results investigation.
3. **Stability Phase: Reproducibility of Pipeline Results |** The Stability Phase focuses on the reproducibility of the pipeline results, which is a recognized concern in biological research and is even more critical in a clinical setting. To ensure a full reproducibility of pipeline execution, the software dependencies should be managed automatically and the data should be indexed and searchable. Ideally, a pipeline user should only define the input data, execute the pipeline, and control the downstream analysis without worrying

about the code versioning, the dependencies management, and the data organization.

4. **Success Phase: Extensibility of Pipeline Composition |** The Success Phase focuses on the extensibility of pipeline composition, which is a significant concern for research labs that need to keep up-to-date and adapt to continuously evolving analysis practices. Extensible pipelines can be easily modified and extended to define new custom pipelines including new tools or a custom selection of steps.
5. **Significance Phase: Scalability for Multiple Users and Large Datasets |** The Significance Phase focuses on pipeline scalability. To be scalable, pipelines should be organized and optimized to facilitate multi-user interactions and support running resource-consuming processes and storing large-scale data.



Significance Phase.
Multiple users can run the pipelines. Infrastructure and performance are optimized to enable running resource-consuming processes and storing large-scale data.

More Data, More Complexity

With the improved availability and affordability of high-throughput technologies such as Next-Generation

Sequencing, the challenge in biology and clinical research has shifted from producing data towards developing efficient and robust bioinformatics data analyses. Integrating, processing, and interpreting the datasets resulting from such advanced technologies inevitably involve multiple analysis steps and a variety of tools resulting in complex analysis pipelines.

The evolution of such pipelines raises serious challenges for effective ways of designing and running them:

- Scientific rigor and collaborative projects require consistency and full reproducibility of analysis pipelines.
- The rapid evolution of technologies and bioinformatics require high pipeline extensibility to easily integrate new tools and features as they emerge.
- Handling large volumes of data, more users, and heavier analysis processes requires efficient management of resources, workload, and high scalability of pipeline design and infrastructure.

To address these issues, life science R&D labs need to invest now in designing and developing reproducible, extensible, and scalable bioinformatics pipelines to avoid playing catch-up later.

Enthought has extensive experience in optimizing complex bioinformatics pipelines leveraging machine

learning and AI. Contact us to see how we can help your team.

*--> Learn more about each phase including mini-case studies in the full paper: **Optimized Workflows: Towards Reproducible, Extensible and Scalable Bioinformatics Pipelines***

AAPS PharmSci 360 🏖️

Heading to **American Association of Pharmaceutical Scientists (AAPS) | @aapscomms** #PharmSci360 in Orlando? Let us know in the comments!

And come see Enthought at the Expo at Booth #3210 and meet Dr. **James Corson** at his talk on **Automated Analysis of Organoid Culture Development**!

Report this

Published by

Enthought
6,108 followers
Published • 4mo

+ Follow

In this month's issue of "Digitalizing Scientific R&D," we demystify the 5 phases of bioinformatics pipelines to help scientists, bioinformaticians, and lab managers evaluate where they are today with their pipelines and identify how to overcome the gaps within their frameworks.

#bioinformatics #pharma #biotech #chemistry #scienceandtechnology #machinelearning #newsletter #research #pharmsci360 #pharmaceutical #NBC2024 #aaps

Like Comment Share 26 1 comment

Reactions

1 Comment

Most relevant

Add a comment...

Raja Karmakar • 2nd
Founder & COO, BioSymphony (we're hiring)
3mo •••

Great! Have a look at this amazing bioinformatics opportunity

https://www.linkedin.com/posts/biosymphony_handsabronabrtraining-linux-pythonabrprogramming-activity-7120055778421477379-edxN?utm_source=share&utm_medium=member_android

Like | Reply

Digitalizing Scientific R&D

All things digital and data in scientific research

1,561 subscribers

Subscribe