Biomedical research in the Cloud: considerations for researchers and organizations moving to (or adding) cloud computing resources

Michelle Holko

Google, Google Public Sector
Washington, DC 20001, USA
Email: michelleholko@google.com

Nick Weber and Chris Lunt

National Institutes of Health
Bethesda, MD 20892, USA
Email: wspc@wspc.com

Steven E. Brenner

University of California, Berkeley
Berkeley, California 94720, USA
Email: brenner@compbio.berkeley.edu

As biomedical research data grow, researchers need reliable and scalable solutions for storage and compute. There is also a need to build systems that encourage and support collaboration and data sharing, to result in greater reproducibility. This has led many researchers and organizations to use cloud computing [1]. The cloud not only enables scalable, on-demand resources for storage and compute, but also collaboration and continuity during virtual work, and can provide superior security and compliance features. Moving to or adding cloud resources, however, is not trivial or without cost, and may not be the best choice in every scenario. The goal of this workshop is to explore the benefits of using the cloud in biomedical and computational research, and considerations (pros and cons) for a range of scenarios including individual researchers, collaborative research teams, consortia research programs, and large biomedical research agencies / organizations.

Keywords: cloud computing; data; bioinformatics; compute research infrastructure.

1. Background

1.1. *Growing use of the cloud in biomedical research*

For at least 30 years, biomedical research data have been growing exponentially, largely since Wally Gilbert first quantified the size of genomics data in 1990 and projected exponential growth until 2040 with a genome for everyone. NHGRI notes that "estimates predict that genomics research will generate between 2 and 40 exabytes [2] of data within the next decade [3]." Making sense from data often requires large and extensible storage and compute capacity, not only because of the sheer size of the data but also because of the complex nature of biology and systems. Additionally, data become more valuable over time, as they grow and also as we learn more about the context surrounding the data. Thus, models that encourage data stewardship and longevity have a greater chance of unlocking discovery.

Many large research organizations are moving to the cloud to handle computational biology research, including the National Institutes of Health (NIH), the National Science Foundation (NSF), the Department of Energy (DoE), the National Aeronautics and Space Administration (NASA), and many academic research institutions. NIH's Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) program is a model for enabling NIH-funded researchers to use cloud resources [4]. It provides choice to researchers by partnering with Google Cloud, Amazon Web Services (AWS), and Microsoft Azure. Through STRIDES, cloud adoption can be done at the organizational (e.g., university) and individual researcher/research lab level. NSF has also been a leader in developing tools like CloudBank for researchers to make it easier to use and track cloud computing in their research grants [5].

Biomedical research increasingly makes use of Machine Learning/Artificial Intelligence (ML/AI) research, as funding opportunities and a focus on developing public policies for ML/AI research grow [6]. These types of research efforts often require large compute and/or supercomputing, beyond what is available to many researchers, from students to principal investigators, on their own laptops. For researchers at institutions who do not have access to large on premise computation and/or supercomputers, the cloud can be a good option to enable research on larger scales. The ability to use tools for ML/AI, such as TensorFlow, can enable researchers to get the most out of their data.

1.2. *Benefits of cloud computing*

Cloud computing can also be used to increase access to compute and storage for researchers at institutions with less infrastructure or IT support. Cloud deployments are almost always more environmentally-friendly, due to both efficient use of computing resources and engineering, and site engineering that minimizes environmental impacts. Data silos are often a problem with on-premise environments, as the data on one's laptop aren't discoverable or easily shareable with

collaborators. This can be overcome with cloud computing, but only if the systems are engineered to improve collaboration and data sharing. Best practices in cloud implementations and engineering are critical to avoid the need for data duplication, re-deploying systems in multiple places, and data leaks. These challenges are not inherent to cloud computing, but are often a result of the technology not being used efficiently. They are also likely signs of an evolving technology and the relevant organizations figuring out how to meaningfully incorporate cloud computing into their funding model to enable researchers.

In addition to filling an immediate need, broader adoption of the cloud into a researcher or organization's infrastructure requires a thoughtful approach and deep understanding of the technology, often in partnership with private sector colleagues. Incorporating cloud computing in an IT infrastructure means the involvement of many different teams, likely including financial, administrative, central IT, research IT, and the researchers themselves. The decision making process often happens at the level of the organization, while the needs of the individual researcher and research groups need to be accounted for in this process.

1.3. *Organizational deployments of cloud computing for research*

Beyond individual research labs, research groups, and organizations adopting cloud, there are many examples of large research consortia building databases and communities in the cloud. The *All of Us* Research Program is a good example [7]. It has developed a custom implementation of Terra, a secure, scalable, open-source, cloud-based platform for biomedical researchers to access data, run analysis tools, and collaborate [8]. The UK Biobank initially used a data download approach and has now moved to a cloud-based platform built by DNAnexus to prevent download and promote centralized data access [9]. The National Cancer Institute's (NCI's) Imaging Data Commons is also cloud-based and provides cancer images and other related data to the research community [10]. NHGRI's AnVIL platform, another implementation of Terra, for genomics provides cloud-based resources for researchers to compute directly on the platform but also allows for data download [11]. When possible, many researchers still tend to download data and compute locally versus leverage cloud computing centrally. This stifles not only collaboration, but also the potential for data reproducibility that centralized platforms with data, tools, and researcher community can offer. Another challenge is that some researchers get accustomed to one system or one cloud platform, and portability can be an issue if a system or cloud platform changes. There are tools to help with this, and many cloud providers are developing multi-cloud solutions to enable portability between and among systems, but this is another thing for researchers to consider in their cloud consideration journey. At the organizational level, the *All of Us* Research Program is committed to expanding to multi-cloud to give researchers the freedom of choice in terms of platforms and tools.

When evaluating the possibility of using cloud for research, researchers and organizational IT professionals often consider the cost, size, and age of on-premise infrastructure, familiarity with and ability to implement cloud-based systems, as well as the research-specific factors like size and

persistence of data sets, frequency of use, types of analysis workflows, and bioinformatics tools and languages. The choice of which cloud(s) to use often also involves cost comparison and an evaluation of which tools are available on the various cloud platforms. Peculiarities of the academic research environment, including especially funding models, complicate the decision about whether to migrate to cloud computing. There is also an ability for organizations to create multi-cloud and hybrid solutions so that the cloud can be used to extend on-premise environments, act as a bridge to cloud computing, and/or enable choice among researchers as to which cloud platform to use. This flexibility means that there are a wide variety of options available, which can also make the decision more confusing and the path forward less clear.

2. Relevance to biocomputing

The size of data, types of data, and types of ML/AI analytic workflows that are used in biocomputing research are relevant for cloud computing, particularly as data grow and are more voluminous. As this trend towards the cloud continues, it is important to share considerations and discuss challenges together as a community. The topic is timely since not only is there a growing use of the cloud, but also growth in data and an emphasis on ML/AI research - all of which require flexible compute and the storage that the cloud can provide. NIH has addressed this topic recently in a Virtual Workshop in September 2021 on Broadening Cloud Computing Usage in Biomedical Research, MSIs, HBCUs, TCUs, etc [12].

The text string “cloud computing” search on PubMed has been growing, with 63 publications in 2021 (Figure 1). Other biomedical conferences that have covered cloud computing include the American Medical Informatics Association (AMIA) and the American Society of Human Genetics (ASHG).

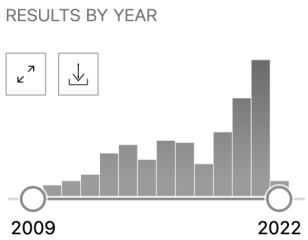

Fig. 1. Number of publications with “cloud computing” in PubMed from 2009-2021

The new policy on data sharing that will go into effect in January 2023 also means that cloud computing will be even more useful for researchers whose data don’t fit neatly into one of the existing NIH primary data archives [13].

3. Workshop overview

This workshop, first and foremost, will be a balanced discussion about the pros and cons of moving to the cloud in a variety of situations, while considering different-sized labs and

organizations, and for a wide range of research applications. This balanced perspective is a key feature to ensure that the discussion is an opportunity for learning and information exchange. The focus will include a range of compute options, including various public cloud providers, on-premise, hybrid and multi-cloud options.

Specific research use cases for biocomputational research in the cloud will be shared, and considerations for researchers and organizations who are evaluating the possibility of moving to the cloud, along with the range of possibilities including hybrid and multi-cloud. A discussion of the evolving technology and the relevant organizations is critical to figuring out how to meaningfully incorporate cloud computing into funding models to enable researchers.

The workshop is organized into talks and a panel discussion. The talks set the stage for the panel discussion, and cover considerations of moving to the cloud and how this went/is going. Talks include both researchers who are using the cloud, and those who are not using the cloud but have evaluated the possibility and decided against it. Session organizers also participate in talks and the panel discussion. The session includes diverse viewpoints, both from the cloud adoption perspective and the organizational type, size, and considerations perspective.

For the panel discussion, private sector researchers were invited to participate, to include the industry perspective along with larger organizations, including NIH. The panel is meant to spark discussion amongst the workshop participants. For both the talks and the panel, diversity and inclusion were goals incorporated into the final workshop organization.

References

1. Y. A. M. Qasem, R. Abdullah, Y. Y. Jusoh, R. Atan and S. Asadi, "Cloud Computing Adoption in Higher Education Institutions: A Systematic Review," in IEEE Access, vol. 7, pp. 63722-63744, 2019, doi: 10.1109/ACCESS.2019.2916234.
2. https://www.backblaze.com/blog/what-is-an-exabyte/
3. NHGRI (2021) Genomic Data Science. Accessed February 8, 2022.
4. https://datascience.nih.gov/strides
5. https://www.cloudbank.org/
6. https://www.whitehouse.gov/ostp/news-updates/2021/06/10/the-biden-administration-launches-the-national-artificial-intelligence-research-resource-task-force/
7. https://allofus.nih.gov
8. https://app.terra.bio/
9. https://www.ukbiobank.ac.uk/enable-your-research/research-analysis-platform
10. https://datacommons.cancer.gov/repository/imaging-data-commons
11. https://anvilproject.org/
12. https://datascience.nih.gov/data-ecosystem/nih-workshop-on-broadening-cloud-computing-usage-in-biomedical-research
13. https://sharing.nih.gov/data-management-and-sharing-policy/about-data-management-and-sharing-policy/data-management-and-sharing-policy-overview#after