

Calvin Li

Professor Stacey Suver

CSE300 Research Proposal

13 April 2018

Research Proposal for a Cloud Computing Platform in Cancer Genomics

The following is a proposal for research and development in the area of cloud computing, specifically a platform to support currently existing cancer research methods. Cancer is a leading cause of death in the world, and it continues to trend upwards; 38% of people will develop a form of cancer during their lifetime. As cancer research is a major division of medical sciences, there is a need for dedicated research teams to seek answers regarding the unknown mechanisms surrounding cancer propagation, cause and effect, and more. Such research includes computation of genomics, a notoriously difficult branch of genetics, where the advent of cloud computing offers a new method of investigating cancer genomics. The project will provide a platform with features such as an end-to-end encryption algorithms in order to perform secure data transactions and analytics; in effect, a tool that a wide variety of cancer researchers could incorporate into their needs. Among other issues, security and a lack of computing power into genomic research are bottlenecking these medical scientists, who would greatly benefit from the aforementioned platform. With logging mechanisms and an analytic back-end, the platform would additionally provide field evidence of the efficacy of cloud computing platforms in computational biology, which may influence other fields as well.

There are two main appeals of cloud-based services. First, they are easily scalable- meaning, they can be resized to fit the computational needs of smaller or larger teams. Second, they are cost-effective- internal data systems can be expensive to set up, as there is much overhead in maintaining private data centers. Currently, there is some preliminary data on usage of cloud computing platforms in literature; since cloud computing as a field is a relatively new area, a primary issue is with the lack of widespread adoption. Despite this issue, cloud is promising, and projects such as MelanomaDB and Simplicity (cited below) have shown positive results where it has been incorporated. Further research, as mentioned in these studies, should seek to expand the existing cloud services for the straightforward purpose of providing more data points where the advantages of cloud programs can be weighed.

Put simply, it cannot be reasonably proven that cloud services offer advantages until there is a large enough data set of real-world uses. This is the exact problem that a new, well-designed, and powerful platform would solve. Furthermore, by offering a full-scale service that abstracts the user data set, security issues would effectively disappear. Any database service, such as MelanomaDB, has the natural problem of all security systems; if their database is breached, an attacker would gain access to sensitive data. The platform proposed would not store data, and therefore have higher ease-of-adoption, which is important when considering how it will affect the number of research teams willing to use our platform.

Instead of storing data, we will use complex hashing/salting, and other encryption algorithms to provide government-level security, as well as processing algorithms in order to avoid storing data. Users will be able to input data into a “black box”, and receive output under any variety of selected map-reduce functions. For example, if a research team is studying a

specific gene and how its variations affect breast cancer, they could select a map-reduce function to input a large dataset of genomes, and the platform would internally manage the data until a specific, abstracted output is returned as a result. While this proposed mechanism may seem unwieldy, it is targeted to the day-to-day questions that would otherwise involve intensive analytics and study.

Our intention is for researchers to have quick and relevant answers that can point them to the correct direction of study, rather than providing a platform that will extensively provide analytics and insights on visualization services. As noted by the National Cancer Institute, this behavior is a leading issue in today's cancer research. While an extensive visualization and analytic service is also in demand, we feel that it is too costly of a project to take on while cloud computing services have not yet been widely adopted in this field. Therefore, it is out of scope of this project, with regards to efficient allocation of time, resources, and capital. The platform will be served across the web, where ease of use is again emphasized.

The target audience is two-fold. Primarily, the platform is directed towards cancer researchers as a service. As with previous cloud services in this burgeoning crossover, the expectation is that lower budget, smaller teams will have the greatest need and usage of our platform, but we hope to see larger groups and organizations adopt the platform as well. However, beyond the life of the platform as a service, usage data will be collected and analyzed for academic purposes. Essentially it will serve as a proof-of-concept to further cloud services and platforms, not just in cancer genomics but in wider fields of computational biology.

Our proposal will contribute an easily adopted platform for cancer researchers, alleviating constraints on their daily behavior. Much of the debate regarding whether or not cloud services

have a place in a sensitive field such as cancer genomics, where patient data should be kept anonymous. However, with an intuitive, easy to access platform that does not see user data but still provides simple analytics, we hope to expand the available data on how cloud services can be incorporated into researchers' needs, as well as how we can adjust services to meet privacy and security concerns. Existing work often suffers from either the issue of not enough adoption/usage, or being too difficult to incorporate into existing research projects. Our project design seeks to alleviate these issues as much as reasonably possible. The value of our project should not go underrepresented; the lifetime risk of 38% across both men and women of developing cancer means that it is a serious issue that deserves the highest priority. Our platform will take meaningful steps towards researching cancer and providing life-saving insights.

Works Cited

- Trevarton, Alexander Joseph, et al. "MelanomaDB: a Web Tool for Integrative Analysis of Melanoma Genomic Information to Identify Disease-Associated Molecular Pathways." *Frontiers in Oncology*, Frontiers Media S.A., 1 July 2013, [doaj.org/article/99ec3686bacd40afbfb6e76661631e3a](https://doi.org/10.3389/fonc.2013.00040).
- Walsh, Paul, et al. "Visualizing Next-Generation Sequencing Cancer Data Sets with Cloud Computing." *SpringerLink*, Springer, Cham, 7 June 2016, link.springer.com/chapter/10.1007/978-3-319-50070-6_4.
- "Cancer Genomics Research." *National Cancer Institute*, 13 Apr. 2018, www.cancer.gov/research/areas/genomics.
- "Cancer Stat Facts: Cancer of Any Site." *National Cancer Institute*, 13 Apr. 2018, <https://seer.cancer.gov/statfacts/html/all.html>