COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Praveen Rao (University of Missouri-Columbia), October 16, 2020

Title: *Democratizing Genome Sequence Analysis for COVID-19 Using CloudLab*

Praveen Rao CIC Database Profile

NSF Award #: 2034247

YouTube Recording with Slides

October 2020 CIC Webinar Information

Transcript Editor: Cora Cole

Transcript

Praveen Rao:

Slide 1

Good afternoon to many of you, it's still 11:50 [a.m.] here, and I'm going to talk about how I'll work on democratizing genome sequence analysis for COVID-19, using CloudLab, which is an NSF-funded experimental testbed for cloud computing. I'm at the University of Missouri-Columbia.

Slide 2

And so the motivation for our work is pretty straightforward: there's a growing interest in understanding how an individual's genome actually impacts the symptoms that a person sees due to COVID-19, as well as the severity of the disease, as well as the eventual outcome - whether they survive the disease or not. So by doing genomic analysis of the genome of COVID-19 patients, we can improve our understanding of the disease, and this can enable us to have new treatment strategies and faster drug discovery. There are a number of publications that are coming up, and one of them is in the New England Journal of Medicine that did a genome-wide association study - and this was about using about 1,900 patients and studying their genetic variants in the genomes. Another effort that has been around is the COVID Human Genetic Effort, and it's an international consortium and the goal is to basically identify how an individual's genome impacts their response to COVID-19. So our genomes may hold the answers to fight COVID-19, and this is an important area to focus on.

Slide 3

So the goals of our project are basically, you know, twofold: the first is to enable researchers to perform variant analysis at scale on human genome sequences, and the goal is to give them the resources at no charge. So variant analysis essentially detects variations in the individual's genome - for example single nucleotide polymorphisms or small inserts and deletes, as well as, we can even think of structural variants such as copy number variations. Now, the other part of the research is going to focus on developing an efficient *de novo* assembly of the human genomes so that we can do deeper analysis of the variants of the genomes of individuals, you know, one belonging to a group that were not impacted by the disease, and the other belonging to the group that were impacted by the disease - and in this particular context, we are looking at COVID-19.

Slide 4

So to achieve our goals, what we're going to do is to develop a software infrastructure using CloudLab. And CloudLab has been around for several years, it was originally designed for computer systems research and it was not really planned for data intensive workloads, but in this particular effort we are going to show how we can leverage CloudLab and have workarounds around some of the limitations that it has to build an infrastructure that can support large-scale genomic analysis using cluster computing technologies, as well as open source tools, so we're going to be looking at the best practices that are out there for genomic pipelines. One of them is GATK, we're also going to look at the BD Genomics project, we're going to use Apache Spark to achieve parallelism, and we're also going to use some of the open source tools that are widely used in the genomics community. And the second part would be to develop an efficient algorithm that is going to help us perform what we call an 'exhaustive variant analysis' using *de novo* assembly.

Slide 5

So essentially we are talking about two groups of patients here, and using bipartite graph modeling, we will be looking at the pairwise comparison between these individuals and have deeper analysis of the variants in their genomes that will help us better understand the disease. And on the right side what you see is essentially the whole ecosystem that we are putting together - leveraging whatever is available in terms of open source software, and building our own components (such as the exhaustive variant analysis engine). And the goal is, at the end of the day, researchers should not worry about having to pay high costs for cloud computing resources either through commercial vendors or other, you know, venues. So CloudLab is a free academic platform and we would like to leverage that to basically empower researchers with the ability to do large-scale genome analysis in an effort to find a cure for COVID-19. We also would like to understand: how does the genomic workloads impact the computer and network

systems, you know? How can we build future systems that are better geared towards processing genomic workloads at scale?

Slide 6

Now here is a project site - we have it actively being hosted on Github - and we are able to allow users to sign up on CloudLab and do variant analysis on a single load, as well as on a cluster. They can also do *de novo* assembly on sequences, we have access to two publicly available resources: one is the Thousand Genomes Project, which is of course not related to COVID-19 but it gives us a lot of data to test our software, and then we have access to the COVID-19 data portal where some of the some of the projects list sequences that are available for us to, you know work with, we also report some performance evaluation on our initial efforts - how fast can we actually do the variant analysis on these sequences in a cluster mode - and as well as *de novo* analysis.

Slide 7

So please follow this link [https://github.com/MU-Data-Science/EVA] if you're interested in doing large-scale genome analysis at no cost, and here is a simple UI that we are currently building so that you can provide access, or provide the URLs for your files, and then you can just say execute and give your email ID and then we will send you back the variant analysis file once the process is completed.

Slide 8

Here is our team, it includes a diverse set of researchers ranging from pathology to genomics to bioinformatics to epidemiology, and my Ph.D. student Arun Zachariah is actively involved in building the software along with me, so feel free to reach out to us if you have any questions or interest in using our platform, and thank you so much for your attention.