COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Ali Rahnavard (George Washington University), October 2022

Title: *A novel platform for data integration and deep learning on COVID-19*

Ali Rahnavard CIC Database Profile

NSF Award #: 2028280

Youtube Recording with Slides

October 2020 CIC Webinar Information

Transcript Editor: Saanya Subasinghe

Transcript

Ali Rahnavard:

Slide 1

Thank you so much for the opportunity. So this is a collaboration work at the Computational Biology Institute, George Washington University.

Slide 2

In brief, you know, the focus of what we do in my lab is - with the high throughput technologies, we can measure many different omics. And our goal is to see if we can get a better picture using integrating these different omics data, including genomics, metabolomics, proteomics, and viral genomes as well.

Slide 3

So indeed, the objective or NSF proposal was to develop a computational platform that includes two sets of - two different sets of analytical approaches to investing - analytical approaches and softwares to investigate COVID-19 related omics data.

Slide 4

So as you seen earlier, in the different, you know, presentation, the main source of this pandemic is the virus, and we are focusing on the genome of virus as a resource to to investigate how the virus behaves,

and if there are specific proteins or regions in the genome of the virus that we need to target for vaccine development.

Slide 5

So the way we look at the data is using the sequencing data, we get the genome of the virus across all the infected, you know, individuals in the pandemic. And we focus on the genome variation also, on the specific region of the genome, including the proteins and other regions like the non-structural proteins, the way we do that - we get this genome sequences for all the individuals in our population, we try to calculate the variation that explained in each of these samples compared to the other samples, and we try to evaluate and see what those variation we see mean.

Slide 6

So here, I'm showing a roadmap of how we can calculate - how we can look at the genome variations in specific regions. In the first row, what you see is the genome variation, correlation to other regions of the genome. And we see four regions, five here, that they are highly correlated to the variation that they have the genome that explains across the populations. And two of those regions are really interesting. One of them is the spike protein region, which, that number you see, 32.5, it says that the variation that we see in the genome of the virus across the population is correlated between the S and the whole genome.

And also we see for the NSP three regions. So the other three regions that you look at the yellow, there are big regions, and they have subsets. That's why we didn't focus too much on those, because we expected when you take a big region of the genome, and you try to look at how that carries the information that the whole genome takes, it should be highly correlated. So we see here that there are specific places in the genome of the virus, like this protein spike protein, we already know that it facilitates the virus, you know, entrance to the human cells. This is highly correlated to non-structural protein 3, and those two all together, co-relate with the whole genome variation that we see in the population.

So as a part of our NSF grant, we developed an approach named omeClust. What it does, you give it a set of points here, the points are the individuals or the strains of the virus, and the three - when we run it using the information from the three different regions, first, the whole genome of the virus. Second, the spike protein and third, the nonstructural protein. We see that those three, they give us the same communities. And that's consistent across all of them. That suggests these two distributions are important to be targeted and investigated further.

Slide 7

Also, we look at this variation that you calculated across the population to see how those are correlated with the epidemiological data. So, we didn't have that much information except like sex and age are a good example here, that the virus genome doesn't really correlated well with the sex and age as expected. The outcome could be unrelated, but not the genome of the variation of the genome.

Slide 8

So, also in our work, what we are doing, part of it, is to integrate the omics data that we measure from infected individuals, like proteins or metabolics. We want to see how those information when we collect them together tell us information that they lead us to hypothesis from the data that we can target. And for example, what you are looking at here, there is an approach we are developing named ‘btest’. What it does, you get given the metabolites information from the patients and the proteins from the patient and body tell us - it gives us a block of relationships, how, what are the metabolites that they are related to the proteins in or infected individuals. Also, we are developing a deep learning approach, which you give it the sequencing data and it tells you - individually, in the person that gets infected, how is the severity going to look based on the genome of the virus that we get from the infected person. So, as you see, we are developing a different, you know, approaches, methods, deep learning approaches, to investigate the omics data, the first focus was on the genome, the viral genome, and now we are moving to the metabolized proteins and try to integrate all this information using the deep learning approach.

Slide 9

So here there are several students in my lab and also co-PIs from the computational biology, Computational Biology Institute to Washington, they collaborate in this work. And it's a team effort, so you can learn more about our approaches and COVID-19 results. We post them on a web page in my lab website. And also the softwares that we are developing that you can use for investigating your data. With that, I want to thank you for your attention.