Western Science

Western Science Speaks Podcast
Season 5, Episode 9

EPISODE TITLE

Classifying New Viruses Within Minutes

PODCAST SUMMARY

What happens when you combine computer scientists, evolutionary biologists, and a global pandemic? A game-changing classification tool. A collaborative team, co-led by Dr. Kathleen Hill and have determined the genomic signature of Covid-19 utilizing a new machine learning program that will allow researchers to easily classify any newfound deadly virus.

Interview

Henry Standage 0:00
Hey, welcome to the Western Science Speaks podcast. Today we have a very exciting team of computer scientists and biologists from Western professors Kathleen Hill, Lila curry, and graduate students, Gurjit Randhawa and Max Soltysiak cracked the genomic signature of COVID-19, officially confirming that the disease originates from bats. To do this, they utilized machine learning to create a new data discovery tool that will allow researchers to easily classify a deadly virus such as COVID-19 in a mere matter of minutes. They came on to the podcast to talk about their ground-breaking research and the importance it could have in the future. Here we go.

All right, why don't we start by having everyone introduce themselves and explain what their role was within this research. Kathleen, why don't you start.

Kathleen Hill 1:24
So, in this particular group, I think one of the major roles is to be a mentor, to be a teacher, to be available and to be listening. So, it's to contribute on knowledge about genomes and their sequence information. And I'm very interested in how genomes change and how they're related to one another.

Gurjit Randhawa 1:50
So, I'm computer scientist. So, my goal is, my role here is to look for interesting problems and then apply my programming skills to develop these tools, software development, and then draw conclusions, which Max can verify later with his biological knowledge.

Henry Standage 2:13
Lila.

Lila Kari 2:14
I am a computer scientist and I shoulder to shoulder with Kathleen, have been for a long time now on the quest of the ultimate universal method that can attach your numerical quantity to DNA and use that in order to exactly classify it and compare it. So, in this particular study, Gurjit, my PhD student and I, were the computational side of the team, and we together worked on the methodology design and the software design, in order to tackle the problem of COVID-19 virus classification.

Max Soltysiak 2:47
Right.

Henry Standage 2:48
And you're at Waterloo correct?

Lila Kari 2:50
I was at Western for 20 plus years. So, my students are still at Western, I'm still a joint professor at Western, but currently, I'm a professor at the University of Waterloo.

Henry Standage 3:01
Okay, I was going to say you're the first non-Western affiliated guest on the podcast.

Lila Kari 3:06
No, I am with Western and I'm faithful to Western.

Henry Standage 3:11
Okay, good. We cleared that up. Max?

Max Soltysiak 3:14
So, I'm a resident biologist kind of with this. And I helped take a lot of the previous work and literature and put our results into a greater biological context and how it relates to this previous work and greater evolutionary context.

Kathleen Hill 3:29
So, Henry, you'll have to keep in mind it's a genome and then it's RNA as opposed to DNA. It's very easily, you know, downloaded as DNA and treated as DNA. But biologically, when we start out, it's an RNA sequence.

Henry Standage 3:47
So, explain to our listeners what RNA means.

Kathleen Hill 3:52
So, they might think about DNA and they might think of a four-letter alphabet such as A, T, A, C, and a G. But in this case, viruses can be RNA sequences. So instead of a T, you would think of a U, we would think of eurocell. And we would have RNA sequences, and they code for proteins. And there are viruses that don't need, you know, to be DNA to be viruses. So those sequences are four letter alphabets. And computationally, those four letters can become other representations, we can make them into the typical ones and zeros that computers like, what we do is make them into images. The idea is that just keeping in mind that the virus being RNA as opposed to DNA, but being the same, you know, in terms of the number of letters and an alphabet, we can take it and easily treat it like DNA and combine it with things that are DNA when we're doing the analysis. So I think a lot of it first of all is observation and asking a lot of questions.

Henry Standage 5:02
And right now, obviously, there is something out there that we don't know a lot about, but we do need to know a lot about it, so with regards to the recent findings, Gurjit, can you tell the story of how you got the wheels spinning on this research?

Gurjit Randhawa 5:17
Yeah, so it was a second week of January, when I was reading the news. And it was like Chinese scientists, they successfully sequenced this genome of COVID-19 virus, and they made it publicly available. So, I was curious enough to see how we can contribute, can we provide something there? Can we just explore it and look for something? So, from my curiosity, what I did is I ran some tests overnight, and I got some amazing preliminary results that I was so excited to share. I knew we had something to contribute because we are correctly classifying it and with so much accuracy, and with so much speed. So, I shared the results early in the morning with the rest of my team members, my colleagues, that this is definitely something we should look into. And we should expand it further. And that is how it all started.

Henry Standage 6:23
Right?

Lila, how does one input something such as the Coronavirus into machine learning?

Lila Kari 6:29
So, what we need is a sequence genome. So, we were very much dependent on this very nice aspect of the COVID-19 virus research, that researchers made the data about the sequence of the virus publicly available. And once we have that, then all we need is an input, we can use the DNA sequence of the genome as an input for the alignment for classification to determine the genomic signature of the virus and then use that genomic signature in order to train the machine learning algorithm and then predict it and train it on existing database of viruses and then predict the label or the taxonomic classification of the COVID-19 virus.

Henry Standage 7:13
So, what does that sample of the virus look like?

Lila Kari 7:17
Well, what we are basing this data on was the publicly available sequence of the virus that was obtained by biologists or medical doctors in their labs. So, they did this process, but they publish their sequences and sequences, which is just like a 30 plus 1000, letter long DNA sequence, ACGT, GGG, CCDBG, and so on. And then that's all we need for our computational experiments. We use real data, not simulated data.

Henry Standage 7:47
What might surprise us about the makeup of COVID-19?

Kathleen.

Kathleen Hill 7:52
So, if we back out for a bit, the virus would have an RNA sequence. And I think people are very fascinated, looking at the components of it. In our work, we don't actually look at the things that it's coding for. I mean, there'll be other people that will figure out how that sequence is coding for various components. The parts that we're most fascinated in is the sequentially, the composition of the letters of its particular genome. And the things that were fascinating to the students and to us is their nature, how they're put together. It's not a random assortment of these letters, but they will have patterns that we can actually illustrate. We can show you pictures of and the manuscript ID and then we can compare it to how the sequences are arranged in close relatives and distant related relatives of that sequence.

Henry Standage 8:56
Max, how much of an eye are you keeping on other group's literature and data while you're in the midst of trying to confirm that Coronavirus originates from bats?

Max Soltysiak 9:07
So, it was pretty interesting when it first came out when people had ideas of where it came from, specifically regarding its origin and obviously a lot of news coverage on for example, the white market and whatnot. But when we're first classifying it based on sequences. Coronavirus is a broad family and it has many different types of coronaviruses. Some, infect, birds some infect, mammals, a lot of them tend to infect bats. When we first did our analyses, other groups were saying as well that it looks like this could be for bats. So when we were doing our first analyses and we were seeing what Coronavirus is and what it was related to, it was pretty interesting and it was great that it was aligning with what other people were saying as well. And from the literature originally when Gurjit first messaged me and said hey, do you want to help with this. I was like, yeah, I want to help with it, I had to read every single preprint on COVID-19 that was out at the time. And obviously that was expanding exponentially, like every single day that it passed on. So, I think by the time we finally started writing, the amount of articles just blew out of proportion. But from those articles, we found that different groups independently found that three specific strains of Coronavirus were important and the analysis. Two of them start with a Z and another one is RATG13. And initially, we didn't have those in our analyses. And then we had to try and dig and find where those sequences

were from. So we can include them. And they ended up being extremely influential on the results. And the fact that all these other groups are also independently coming to that conclusion means that these sequences are important. And that's kind of what was really interesting about this analysis, and how closely related these three sequences are and how other groups are independently coming to the same results with different data. It's just, for example, we can do it within five minutes, without any genes or anything like that.

Henry Standage 11:03
Something really unique about this situation is how much news about it there is out there, but how little is actually known. So, when you're digging through articles, how do you choose which ones you're really going to focus on? Or take on board with you? And which ones you kind of disregard?

Right. And that's a big golden question that everyone's having to deal with at the time. Preprints are so important, especially with rapidly evolving science and with a pandemic. And there's a need for a lot of this research to be accelerated. A lot of the time, we have to keep in mind that preprints haven't been peer reviewed yet. So just because it's a preprint doesn't mean its good science. And what results that we trust in? Which results can we cite? It’s a little bit of a gray area, and you got to do a little bit of digging into the methodology. And when you start reading some of these papers, a lot of the initial preprints were extremely short. They're like three paragraphs and a single statement saying we should focus on this. And it's like, what does that paper bring to the table, oftentimes it's not much. While a lot of the results could be important, but because they were peer reviewed yet, you don't really know what to trust. A lot of the time, we were taking a lot of their data, and trying to see if we can use their data, less so their results, but more so the sequences and the data that they were saying is important. And we include this in our data, we find the same conclusion. But we can get more out of that preprint and paper, other than just taking it for face value, but actually incorporating their data into our own analyses and figuring out on our own independently, can we trust this? Is this looking good? But it's definitely tricky, especially as it went from three papers to five papers to 20 papers to 60 papers to 120 papers, how many of those are actually important? And how many of those can we kind of trust.

Max Soltysiak 12:51
But prior to this peer review?

Henry Standage 12:53
Yeah, there's such an opportunity to gain notoriety. As a scientist right now. It's got to be tough to dig through all that. Gurjit can you touch on just how expansive and powerful the computational methods you're deploying are.

Gurjit Randhawa 13:08
That's the beauty of our method. It's really, really fast. So, comparing to other methods, if you say, analyze these 1000s of genomes on an alignment based method that will at least take few hours, maybe days. We can do it in under two, three minutes. So, we are very light on computational side because we are just using sequences, we don't need much computational power at our end. And it can be done on any average computer at your home within few minutes, and the software is open source publicly available that so you can download and run on your system and get the same results.

Henry Standage 13:57
I believe your paper notes that there are 25 different DNA sequences currently associated with Coronavirus. How different is one sequence from the other?

At time of our study, we used like all 29 complete genomes available till that date. But as of now there are more than 30,000 sequences available. So, it's expanding. So every country is now sequencing and they are reading 1000s of sequences every week. So, it's still ongoing. Among those 29. They were very, very similar. More than 99% similarity within those 29 sequences.

Kathleen, you look like you're itching to hop in here for a second.

Kathleen Hill 14:42
The 30,000 number is up to date as of today. But also, they're averaging, they're tracking one mutation every two weeks, according to a webinar earlier this week with CanCOVID. So, they're tracking how much it actually changed. That's sort of an average, that'll have variation. And we're really watching science unfold. It's not an experiment that had a design, is being carried out, you get to see the results. You're watching a very dynamic process. So, when you make interpretations, they're sort of dynamic as well.

Henry Standage 15:21
That's an epidemiologists dream I imagine.

Kathleen Hill 15:24
It's a teacher's dream, people at Western should know that this community will be teaching based on this example, with students learning experiential learning in the fall, they will be given or choose their own cases like this and follow it along. And there were classes last year that actually were using these genome sequences and trying some of these methods. So, it'll be fun to be doing things that are live, you know, actual real world problems and trying to participate as we go.

Henry Standage 15:57
On that subject of teaching. We've never been forced to create a cure for Coronavirus in the past. Will it be easier to solve COVID-19 if we become familiar with prior iterations of the virus?

Kathleen Hill 16:11
I could try that one. But I wondered if Max, I'd been pondering that question if you wanted to try it as well, I have a solution. But I'm curious to hear Max first or somebody else.

Henry Standage 16:20
Max.

Max Soltysiak 16:21
As we saw previously, when I was just a kid barely thinking at all. With the SARS, the original SARS epidemic and how that really influenced the world. There was a lot of work going into vaccine development for SARS. And well when SARS no longer became as big an issue as it previously was thought to be a lot of the funding, for example, towards vaccine development, it was kind of cut off and a lot of the progress that worked towards creating a SARS vaccine kind of tapered off a little bit. But similar to other related viruses, a lot of the work that we did with the original SARS pandemic and SARS vaccine development was used as a head start on the development of a vaccine for SARS or COVID-19. So, while it's not really simpler because the SARS epidemic was dangerous. And it was taken very seriously. A lot of the work that previously was done and because COVID-19 is related to SARS, a lot of that work that was done could already be used as a head start and a lot of the information for example, with the spike proteins, its replication, a lot of its effects on the human body, for example, we had a little bit of a knowledge base we could go off. If this was for example, a novel adenovirus or something else along those lines where we don't necessarily have as much information on it, it would take a lot longer and a lot more work in order to get where we are now. So, a lot of these previous iterations and previous versions that we've had to deal with helped for example, with all the research we're doing and where we can base our new progress, that we work on.

Henry Standage 17:56
When we spoke before you all made it very clear that this project is nowhere near the finish line. So Lila, why don't you tell us what's next? And how specific can understanding of the virus become?

Lila Kari 18:11
Oh, I if I were to veer off the question a little bit here, I think that in our method as a mathematician and computer scientist, like me and Gurjit are, I think that the most interesting aspect besides being super fast, which is really very important when time is of essence. But besides that, I think that the most interesting aspect of this research is its universality, if I may use that word with a grain of salt, in the sense that now we have like an off the shelf tool

that we can apply for anything at any moment in time, with no modification, and we get an answer. So rather than you know, have a new thing, you have to find the genes, the proteins, compare it and so on. So, there is lots of biological orbia has to be done. This is SS no modification whatsoever. And you can compare and find out the relatedness literally between anything and anything between real genomes, synthetic genomes, computer generated genomes, whatever you know, your cat walking on the keyboard genomes, anything you want. As long as a sequence of DNA or RNA, you can compare it and place it in sort of the tree of life, so to speak, and find out what it is. So, I think that as a mathematician, I found this sort of deeply satisfying that instead of every case, having a particular solution, here we have a one size fits all solution ready to deploy at any time for the future. So from my point of view, this is sort of the most exciting part.

Henry Standage 19:44
Is it fair to say that ultimately, the legacy of this research is that we'll be more prepared in the future?

Lila Kari 19:51
Absolutely. Absolutely. Absolutely. Because then at least we don't have to wait for three months for the peer review. Now we can do it in the first day and be done with it as opposed to waiting for the validation of the method itself.

Henry Standage 20:05
If I were a researcher looking to create a vaccine now how might I deploy your findings?

Kathleen Hill 20:11
So, step one, I think you want it super early. So, if you know there's a problem and you know it's a pathogen, you're going to want to assemble the circumstantial evidence of biologically what you think that pathogen is. And you're going to want to deploy this as quickly as possible because it's not what Gurjit and Lila and these folks have developed, that's going to take a long time, it's going to be you assembling that and then running it through, because within minutes, you're going to have an answer, but you need to put in the right information. And then you want to do it very early on. And then if you're the vaccine type, the therapy type, the diagnostic type, it's going to be that closest neighbor, you find and doing that properly. That then gives you all the wealth of biology, all of that information that's associated with that closest classification. So, you want to do it early. And you want to right at the scene of knowing you have that pathogen you want all of the circumstantial material of what it is.

Henry Standage 21:24
10 years down the road, where would you like to see this research being utilized?

Kathleen Hill 21:33
I think everybody has something imaginative, I say take it to the classroom, I say put it in the hands of users. I have recruited an uncountable number of students this summer to practice, try, find problems. And then I go into my lab and find crazy amounts of data and ask if we can start applying it. Can it somehow help us classify disease types, something else of interest for me?

Henry Standage 22:05
Well, I just want to quickly express my gratitude for everyone making time for me today. This is a wicked cool podcast for me to do. And I'm excited to see what else comes from this research over this summer and beyond. So thank you all.

Lila Kari 22:19
Wow,

Kathleen Hill 22:20
We have a fascinating meeting. We're going to convene right after we exit your Zoom. I'm going to tell these people I'm going to issue another zoom, because we have an important deadline for the next project.

Henry Standage 22:33
All right, well, enjoy that. Thank you.

Kathleen Hill 22:35
Okay, thank you.

Henry Standage 22:38
That concludes another episode of Western Science Speaks thanks to the team for coming on. If you're interested in reading more about their research, they're getting a lot of press at the moment, and deservedly so. If you enjoyed the show, subscribe to us on Apple, Spotify, PodBean wherever you get your podcasts and make sure you stay up to date with the latest research and conversations from Western Science. Next week we'll be airing a special best-of featuring five really interesting interviews over the years pretending to health sciences. You can catch that exclusively on streaming services, so make sure you subscribe. In the meantime, I'm Henry Standage, signing out. Thanks for listening.