COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Ho-Joon Lee (Yale University), February 2022

Title: A landscape of virus-host protein-protein interactions in SARS-CoV-2 infection in humans by machine learning

Funded by NSF Office of Advanced Cyberinfrastructure, Directorate for Computer & Information Science & Engineering (OAC/CISE) through the Northeast Big Data Innovation Hub Seed Fund Program.

Youtube Recording with Slides

February 2022 CIC Webinar Information

Transcript Editor: Saanya Subasinghe

Transcript

Ho-Joon Lee:

Slide 1

Hello, everyone my name is Ho-Joon Lee from Yale School of Medicine and I am going to talk about *An interactome landscape of SARS-CoV-2 virus-human protein-protein interactions by machine learning*.

Slide 2

There are two objectives. The first is to develop the protein sequence based multi-class machine learning or deep learning classifiers for evidence or confidence level prediction using the Viruses.STRING database. The second is to - using those classifiers we want to create a draft interactive landscape of cytoskeletal virus human protein-protein interactions.

Slide 3

So here is an overview of our machine learning and deep learning workflow. So we use the Viruses.STRING database, which did not include the SARS-CoV-2 at the time of the analysis. This is the network of PPI virus-human PPIs which contain more than 80,000 interactions between about 1,200 virus proteins from 102 virus species and about 8,500 human proteins. And each interaction has a combined score ranging from zero to one thousand which we convert into five evidence classes pieces. And this is the distribution of the number PPIs for evidence classes. And we are going to focus on the experimental PPIs which belong to evidence class 3 or 2 based on the zero index here. And based on the data, we first extract node features, another protein features which are fractional compositions of 20 amino acids. And at this point we are developing two different models - one is more canonical [inaudible] models like Random Forests and XGBoost, in this case. And another one is based on deep learning. We specifically use graph neural networks like GraphSAGE or datalized version of HinSAGE. For connected machine learning, we also extract Edge features which are 72 distance or similarity measures between amino acid composition profiles between virus proteins and human proteins. And based on the features, we developed the Random Forests and XGBoost. For Random Forests, we optimize 36 models by research with temp request regulation and 432 models for executive space which has the same contemporary transplantation. And, in short, we obtain up to 67% accuracy and 37% accuracy for Random Forests cases and 74% accuracy and 67% accuracy for XGBoost cases. And this work, this part, has been published as a preprint recently. So you can refer to the paper in detail [https://www.biorxiv.org/content/10.1101/2021.11.07.467640v2]. And for GraphSAGE here, still in advanced reading and preparation, but I'm going to show you, briefly show you, the results from GraphSAGE as well. Because this shows more than 70% accuracy which is plenty promising as well.

Slide 4

And here I'm going to show you a performance example for the best models for 20% of [inaudible] with this random seed. We see, in this case, when the Forest shows 60% accuracy, XBG was 67.7% accuracy. And if you look at computer metrics, again, I'm going to focus on this [EC3?] which implies mostly expanded PPIs. And if we look at the individual classes, focusing on the f1-score, the extra booster shows higher f1 scores across four individual classes.

Slide 5

Based on, based on this to [inaudible] boost model. The important features were identified using two alternating methods here. One by Gini index and the other by SHAP analysis, which based on SHAP, came through at the SHAP values. And interestingly enough, we see that cysteine and histidine are most - two most important features. Where this minus [C_minus and H_minus] means that the fraction of cysteine between virus and human. And the ratio means the ratio between the fractions cysteine and histidine reactions between virus and humans.

Slide 6

One control experiment we performed is to compare prediction of experimental PPIs and - with a prediction of text mining PPIs in the virus [inaudible]. Because the data size, the difference is pretty big here, six squared difference, but what we observed here is that XGBoost, in fact, shows higher accuracy. With 94% accuracy compared to 90% accuracy for the text mining case. So despite the data size difference, XGBoost it shows a good prediction performance. And this is the agreement between random force activities for ec3 and test binding as we expect shows mostly ec1 or ec2.

Slide 7

So based on those encouraging results, we applied those classifiers to SARS-CoV-2 for our second objective in two ways. So first, we apply that to IntAct database, which are a collection of experimental PPIs. And here I'm showing you the network by XGBoost with [inaudible] predicted evidence. So ec3 for blue, ec4, red. So this can be viewed as prioritizing a network. So although these links would be about 2,000 links from experimental data are equally meaningful, we can also prioritize those links based on this [inaudible] class predicted XGBoost in this case. Secondly, we also apply that to protein-wide interaction [inaudible] the old pairs of more than half a million between 27 SARS-CoV-2 proteins and about more than 20,000 human proteins. And here I'm showing you the subset of 22,000 PPIs with evidence class of at least 2. I either actually use XGBoost or Random Forest. And this is the another subset - 140 PPIs with the highest evidence class, 5, by XGBoost. And based on this interaction network we observed that many human proteins are enriching vascular smooth muscle contraction and the targets and also H2A components.

Slide 8

There are a few more applications of this work that have been found in the past month, actually. So Giuseppe Novelli, who is the renowned geneticist in Rome, in Italy, he reached out to me by email and by surprise, last month. He had read my preprint telling me about his quality therapeutics publication for HECT E3 ligases and his idea of using the results from this important work through this ongoing research. And we immediately realized that we can help each other based on my results on interactive network results. And we found that HECT-domain protein tend to interact with the SARS-CoV-2 proteins with an evidence class of greater than 2 with statistical significance. In other words, HECT-domain proteins are favored by SARS-CoV-2. Based on that observation you're asking whether there are other protein families favored by SARS-CoV-2. In addition we can also extend that to other virus species like human metapneumovirus, which Dr. Novelli is also working on as well.

Slide 9

So finally, I'm going to show you briefly about the graph neural networks using GraphSAGE and HinSAGE architecture. On the left are the accuracies by 15 different models using three different Java weights in the columns and five different edge embedding methods. As you see, without dropout rates, in fact, we see more than 70% accuracy values and accuracies which are very promising. This is a based on Viruses.STRING and if we apply that to SARS-CoV-2 IntAct database, you see the prediction is enriched with evidence class 2, or 3, in fact, which are mostly [inaudible] PPIs. And this consensus is number of agreements among these 15 different models. We see more consensus of, like, 8-9 for against two compared to 6-7 against 1 but I think this is also very important because - we will see.

Slide 10

Okay, with that I'd like to thank my collaborators for very helpful discussions and feedback and support. And the Yale Center for Research Computing for computational resources. And COVID HASTE community from Yale’s School of Engineering & Applied Science. And finally the Northeast Big Data Hub Seed Fund for supporting this work. Thank you so much.