COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Marc Riedel (University of Minnesota), July 16, 2021

Title: EAGER: Computationally Predicting and Characterizing the Immune Response to Viral Infections

Marc Riedel Database Profile

NSF Award #: 2036064

YouTube Recording with Slides

July 2021 CIC Webinar Information

Transcript Editor: Macy Moujabber

Transcript

Slide 1

So, hello everyone. I'll be speaking today about computational work on the problem of peptide binding and this is a collaboration with George Vasmatzis from the Mayo Clinic.

Slide 2

So, with COVID-19, we've all heard that the disease is- it behaves like any disease. There is a range in severity of symptoms that people experience, but it's perhaps more pronounced with COVID-19 than other diseases. A good fraction of people who contract the virus show no symptoms whatsoever. Most show either no or mild symptoms, but a significant fraction shows severe symptoms. And of course, a small fraction show critical symptoms or even experience death as a result. So, there are many factors influencing the disease severity and experts can speak better about this than I can. Age, sex, and in particular comorbidities are very significant and past exposure to similar viruses plays a role, but a part of the equation here in terms of disease severity is also the innate differences in our immune systems. Essentially, the genes we've inherited from our parents play a role in how well our bodies fight off this disease or most diseases. So, our topic of study here is to computationally predict one aspect of the innate difference in an individual's immune system's so-called cellular response.

Slide 3

So, the cellular response is the first line of defense that our bodies have in response to any viral infection. It is the mechanism by which foreign peptides introduced into cells are chopped up into small fragments called peptides. These then can be transported to the cell surface where they can bind to cell surface receptors called MHC [Major Histocompatibility Complex] Class 1 molecules. If bound this way, the infected cells become targets for killer T cells which can come off and kill the infected cells- either killed from a viral infection or as it turns out, this is the most effective defense that we have against cancer. Cancerous cells are also killed off this way. This is the first line of defense because if it kicks in once we're infected, the immune system- the killer T cells can kill off all the infected cells before they have a chance to get going. But if the cellular response fails, then the infected cells become factories. They start churning out many many copies of the virus, and other aspects of the immune system have to take over.

Slide 4

So, the cellular response is- as with all aspects of immunology complex. Understanding it is very important. It's critical for understanding and predicting the severity of novel viruses such as SARS-CoV-2, for vaccine development targeting viruses, also understanding the impacts of viral mutations. How different viruses will affect different people through the- their innate cellular immune response, and as I keep returning to the topic for cancer immunotherapy, which is the area of expertise of my collaborator- like many people we turned our attention to COVID-19, and repurposed the skill sets. So, in this case, the computational predictions are originally targeting cancer immunotherapy. So, and of course all of this same theory applies to autoimmune diseases. So, the cellular immune response, at its core, is a computational problem. If we have the blue peptides in the figure on the right this is the protein fragment associated with the virus. The question is: will that blue peptide bind inside a groove or a cleft inside the yellow cell surface molecule the MHC 1 molecule? And the immune response then will be dependent upon whether this protein fragment binds, and whether it binds well enough for the killer T cells to recognize it. The blue fragments that come from the virus- those are novel. So, given a new, virus we will have a completely new set of peptides. The yellow cell surface molecules- those are determined by our genes and each individual has up to six different MHC 1 molecules determined by our inheritance- three from each parent. And the cell surface molecules, the MHC molecules, are among the most diverse in our genome, there are approximately 21,000 variants in the human population. This is no accident. Evolution has ensured that this aspect of our immune system is very diverse so that we've been able to survive past viral infections.

Slide 5

But this also poses a very significant computational challenge as I'll describe. So, the core problem that we're tackling is how to apply computer science to predict whether the blue peptide will bind inside the yellow cell surface molecule, and to do this for a very wide range of peptides, all of those associated with the virus, and for the very wide range of cell surface receptors the MHC 1 molecules. Now there's been a lot of prior work- computational prior work. There's been experimental work that of course has

provided molecular information on the structure of these molecules, and the computational work that's been very successful has been to apply a machine learning- no surprise there for those who have a background in computer science and neural networks. So, there's a package called NetMHC that has been trained extensively on the experimental data with the binding strength for known pairs of the blue yellow molecules- the MHC 1 peptide pairs. And based upon a neural network structure, this program can predict, given a new peptide, how strong it will predict. And this inference is based upon simply peptide sequence, so the sequence of letters, these are the amino acids in the peptide are paired with a label that corresponds to the cell surface molecule the, MHC 1 molecule. There's a strength. That's all the experimental data. And so, given many thousands such pairs, the neural network can be trained. And once it's trained, given a novel peptide sequence, a novel sequence of amino acids, it can predict how well that peptide will bind to a given MHC 1 molecule. So here, the peptide sequence would be the blue peptide molecule in the drawing, and the label would correspond to the yellow cell surface receptor, the MHC 1 molecule. This is great and neural networks are powerful in the sense that they can be easily trained and they can very effectively make predictions based upon the data they're given. But the limitation here is there is absolutely no molecular data whatsoever, we're simply training labels and letters, and also the training data is taken for a very diverse set of experimental data. A lot of it for instance comes from HIV, a different virus, and the peptides associated with HIV, and the problem is given a brand-new virus like SARS-CoV-2, most of the peptides have never been seen before, and a neural network will make predictions that are spurious because it's making inferences from a very different region of the peptide space.

Slide 6

And so, our approach would- is to do molecular-level simulations, and of course there's been a lot of prior work on this topic. Very sophisticated molecular simulation techniques are known and widely used. One is called Molecular Dynamics. There's also Monte Carlo based simulations. They use a technique called Simulated Annealing, and Molecular Docking is another approach. Software available for such molecular level simulations are widely used, but they're special. They're general purpose. They've been developed for broad classes of molecules binding and they're extremely computationally intensive. To take a peptide, an MHC 1 pair, and to use existing software to simulate it, it takes days, sometimes weeks to simulate a single binding event. So, weeks of actually supercomputing time to make a single prediction. And the scope of the problem we're confronting is we have 21,000 variants of the yellow cell surface molecules, the MHC1 molecules, and for SARS-CoV-2, if we focus just on the spike protein and we chop that up into little bits for the peptides, we have about 38,000 of those. So, we're talking about 1 billion combinations that we want to simulate in terms of the strain, and if it takes a week of supercomputing time each we obviously don't have a billion weeks to study this.

Slide 7

So, our approach is twofold. On the one hand, we're creating highly customized software for molecular simulation and we're using the details of the domain we're working in. So, we're starting with the peptide. The peptides don't vary so much in terms of their length or their shape. We're starting with the peptides correctly aligned inside the cleft of the cell surface molecule, the MHC 1 molecule, so we don't

spend a lot of time just rotating the entire peptide in space. We place it exactly where it should be, and we perform the entire search in the torsional space. So instead of moving the molecule around three-dimensional space we just twist and turn its bonds to try to find the optimal configuration. The other contribution is we're deploying this at scale. So, we're using GPUs, and then eventually cloud computing infrastructure, to really throw computing power at the problem. And our goal as I stated in the title is to turn a billion days into a million minutes or perhaps one month of cloud computing time.

Slide 8

So, the challenge is here and I'll gloss over this- I'm almost out of time. One thing: the experimental data is not complete. We don't have full molecular models, not only of all 21,000 variants, but we don't even have a good geographic spreads. This ties in with some of the previous talks. Most of the experimental data is for the variance of the MHC 1 molecules from Western Caucasian demographics. So, the other approach is we must infer the structure of the molecules and then simulate them.

Slide 9

And so, final slide: the impact of this work. Well it's relatively straightforward to determine for an individual what genes they have- that code for the relevant molecules, the MHC 1 molecules. It can be done through HLA typing, which is done for paternity testing. Given that information and given our computational infrastructure, we'll be able to predict for an individual, how that individual will respond to a new pathogen. to a virus, to variants of the virus for different individuals, and for the effect of different vaccines on different viruses for different individuals. And of course, as I stated at the outset, this work will apply not only to viruses, but also potentially the cancer immunotherapy and autoimmune diseases. So, I'll stop there. Thank you very much.