
## HOW A LANGUAGE PROCESSING MODEL CAN PREDICT MUTATIONS IN VIRUSES

One of the biggest challenges in developing antiviral therapies or vaccines is the propensity of viruses to mutate. In many cases, the nucleic acid or protein target in the virus changes significantly so the vaccine or therapy is no longer effective. This viral escape effect requires a constant cycle of vaccine development and therapies which have a high impact on cost, timeline and clinical needs. A high throughput model to predict the mutation patterns in viruses has not been developed using conventional experimental methods. These techniques are time consuming where single virus strains are profiled and it is not possible to analyze the effect of multiple mutations on a virus. However, there has been a recent breakthrough by a group at MIT that used Natural Language Processing (NLP) to develop models to identify mutations in viruses and just as important, identify areas that are less likely to mutate[1].

Natural language processing (NLP) includes machine learning algorithms that were originally developed to understand human languages. Simply put, NLP is the ability of a computer to analyze and manipulate human languages. Human languages follow set grammar rules so the underlying hypothesis of the work done at MIT is that the same principles used in a language model can be used to analyze viral proteins. Some of the key hypotheses that were used in the study included a) semantics changes in the language model corresponded to antigenic changes; b) the grammaticality (or conformity to set grammar rules) of the language model translates to viral fitness and c) both these elements together help predict viral escape. The complex model was used to search for mutations in three well known viral proteins – influenza A hemagglutinin (HA) protein, the HIV-1 envelope protein and the SARS-CoV-2 spike protein. These proteins are responsible for binding to target cells and are widely studied drug targets. Therefore, it is important to understand the viral escape associated with these proteins in order to better design antiviral therapies.

The model showed striking results in all 3 proteins. The HA protein consists of a globular head and stalk domain and it has been experimentally shown that the head protein is more like to mutate compared to the stalk. The NLP model confirmed this finding and also supports the development of neutralizing antibodies that target the protein stalk. Similarly, the V1/V2 regions of the HIV-1 envelope protein were also confirmed by the model to be susceptible to mutations and consequently viral escape potential. One point to note is that the model in its current form detects genetic mutations leading to amino acid changes and does not account for post translational modifications like glycosylation or phosphorylation. The SARS-CoV-2 spike protein was predicted to have the highest mutation rates and escape potential in the N-terminal and receptor binding (S1) domains while the S2 subunit was predicted to be more stable and less likely to mutate. However, one question that remains unanswered the rate at which the SARS-CoV-2 virus mutates but given the reports on newly identified mutations, the findings from this model could be further refined to predict the best antigen targets for long lasting vaccine development.

The model relies on the fact that the evolution of viruses is based on maintaining viral fitness to continue the replication and infection cycle while escaping detection by the host's immune system – a viral Darwin's theory of the survival of the fittest. The NLP model has several advantages in that only genetic sequence information is required instead of complex tertiary protein structures and the amount of input data is modest. For this proof-of-concept study, 60,000 HIV sequences, 45,000 influenza A sequences and 4,000 SARS-CoV-2 sequences were used[2]. The potential for using this model is enormous as genetic sequence data is fast and relatively inexpensive to generate compared to complex protein analysis. The MIT group's is currently working on identifying targets for tumor vaccines[2] but as this model improves it can be used for AI based drug design of novel therapies to overcome drug resistance.

**References:**

[1] Hie *et al.* Learning the language of viral evolution and escape. Science (2021); 371:6526, 284-288.
[2] https://news.mit.edu/2021/model-viruses-escape-immune-0114