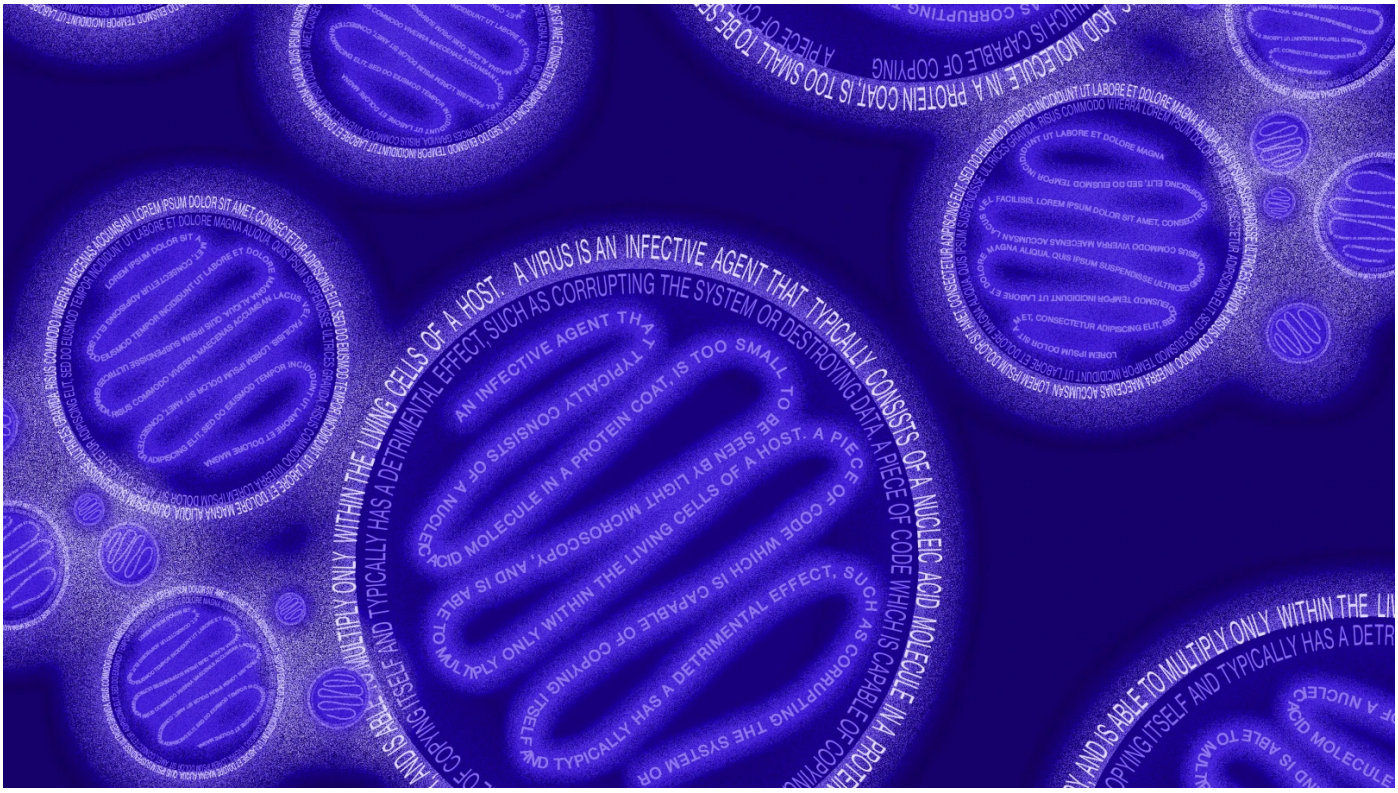
AIs that read sentences are now catching coronavirus mutations

NLP algorithms designed for words and sentences can also be used to interpret genetic changes in viruses—speeding up lab work to spot new variants.

Will Douglas Heaven



Galileo once observed that nature is written in math. Biology might be written in words. Natural-language processing (NLP) algorithms are now able to generate protein sequences and predict virus mutations, including key changes that help the coronavirus evade the immune system.

The key insight making this possible is that many properties of biological systems can be interpreted in terms of words and sentences. “We’re learning the language of evolution,” says Bonnie Berger, a computational biologist at the Massachusetts Institute of Technology.

In the last few years, a handful of researchers—including teams from geneticist George Church's lab and Salesforce—have shown that protein sequences and genetic codes can be modeled using NLP techniques.

In a study published in Science today, Berger and her colleagues pull several of these strands together and use NLP to predict mutations that allow viruses to avoid being detected by antibodies in the human immune system, a process known as viral immune escape. The basic idea is that the interpretation of a virus by an immune system is analogous to the interpretation of a sentence by a human.

"It's a neat paper, building off the momentum of previous work," says Ali Madani, a scientist at Salesforce, who is using NLP to predict protein sequences.

Berger's team uses two different linguistic concepts: grammar and semantics (or meaning). The genetic or evolutionary fitness of a virus—characteristics such as how good it is at infecting a host—can be interpreted in terms of grammatical correctness. A successful, infectious virus is grammatically correct; an unsuccessful one is not.

Similarly, mutations of a virus can be interpreted in terms of semantics. Mutations that make a virus appear different to things in its environment—such as changes in its surface proteins that make it invisible to certain antibodies—have altered its meaning. Viruses with different mutations can have different meanings, and a virus with a different meaning may need different antibodies to read it.

To model these properties, the researchers used an LSTM, a type of neural network that predates the transformer-based ones used by large language models like GPT-3. These older networks can be trained on far less data than transformers and still perform well for many applications.

Reading viruses

Instead of millions of sentences, they trained the NLP model on thousands of genetic sequences taken from three different viruses: 45,000 unique sequences for a strain of influenza, 60,000 for a strain of HIV, and between 3,000 and 4,000 for a strain of Sars-Cov-2, the virus that causes covid-19. “There’s less data for the coronavirus because there’s been less surveillance,” says Brian Hie, a graduate student at MIT, who built the models.

NLP models work by encoding words in a mathematical space in such a way that words with similar meanings are closer together than words with different meanings. This is known as an embedding. For viruses, the embedding of the genetic sequences grouped viruses according to how similar their mutations were.

The overall aim of the approach is to identify mutations that might let a virus escape an immune system without making it less infectious—that is, mutations that change a virus’s meaning without making it grammatically incorrect.

Take a language example. Changing just one word in the sentence "wine growers revel in good season" can produce the sentences "wine growers revel in strong season" or "wine growers revel in flu season." Both share the same grammatical structure but one has changed its meaning more than the other. The tool looks for similar changes in a virus, flagging those that change its meaning most.

To test their approach, the team used a common metric for assessing predictions made by machine-learning models that scores accuracy on a scale between 0.5 (no better than chance) and 1 (perfect). In this case, they took the top mutations identified by the tool and, using real viruses in a lab, checked how many of them were actual escape mutations. Their results ranged from 0.69 for HIV to 0.85 for one coronavirus strain. This is better than results from other state-of-the-art models, they say.

Looking ahead

Knowing what mutations might be coming could make it easier for hospitals and public health authorities to plan ahead. For example, asking the model to tell you how much a flu strain has changed its meaning since last year would give you a sense of how well the antibodies that people have already developed are going to work this year.

Still, this work is more about breaking new ground than making a real impact on public health—for now. Since doing the work published in Science, the team has been running models on new variants of the coronavirus, including the so-called UK mutation, the mink mutation from Denmark, and variants taken from South Africa, Singapore and Malaysia.

They have found a high potential for immune escape in all of them—although this hasn't yet been tested in the wild. But the model did miss another change in the South Africa variant that has raised concerns because it may allow it to escape vaccines. They are trying to understand why that is. "It consists of multiple mutations and we believe a combinatorial effect is coming into play," says Berger.

Using NLP accelerates a slow process. Previously, the genome of the virus taken from a covid-19 patient in hospital could be sequenced and its mutations re-created and studied in a lab. But that can take weeks, says Bryan Bryson, a biologist at MIT, who also works on the project. The NLP model predicts potential mutations straight away, which focuses the lab work and speeds it up.

"It's a mind-blowing time to be working on this," says Bryson. New virus sequences are coming out each week. "It's wild to be simultaneously updating your model and then running to the lab to test it in experiments. This is the very best of computational biology," he says.

But it's also just the beginning. Treating genetic mutations as changes in

meaning could be applied in different ways across biology. “A good analogy can go a long way,” says Bryson.

For example, Hie thinks that their approach can be applied to drug resistance. “Think about a cancer protein that acquires resistance to chemotherapy or a bacterial protein that acquires resistance to an antibiotic,” he says. These mutations can again be thought of as changes in meaning: “There’s a lot of creative ways we can start interpreting language models.”

“I think biology is on the cusp of a revolution,” says Madani. “We are now moving from simply gathering loads of data to learning how to deeply understand it.”

Researchers are watching advances in NLP and thinking up new analogies between language and biology to take advantage of them. But Bryson, Berger and Hie believe that this crossover could go both ways, with new NLP algorithms inspired by concepts in biology. “Biology has its own language,” says Berger.