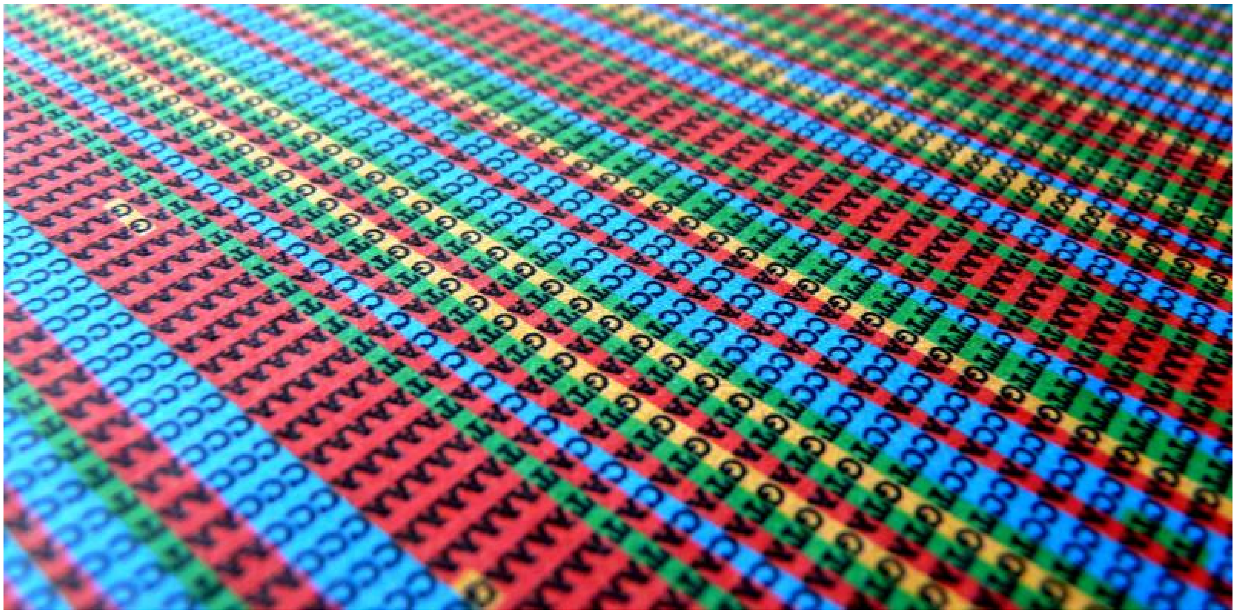
How computers help biologists crack life's secrets

December 17 2015, by Sri Krishna And Diego Chowell



It’s a lot for a person to puzzle out… call in the computers! Credit: Shaury Nash, CC BY-SA

Once the three-billion-letter-long human genome was sequenced, we rushed into a new "omics" era of biological research. Scientists are now racing to sequence the genomes (all the genes) or proteomes (all the proteins) of various organisms – and in the process are compiling massive amounts of data.

For instance, a scientist can use "omics" tools such as DNA sequencing to tease out which human genes are affected in a viral flu infection. But because the human genome has at least 25,000 genes in total, the number of genes altered even under such a simple scenario could potentially be in the thousands.

Although sequencing and identifying genes and proteins gives them a name and a place, it doesn't tell us what they do. We need to understand how these genes, proteins and all the stuff in between interact in different biological processes.

Today, even basic experiments yield big data, and one of the biggest challenges is disentangling the relevant results from background noise. Computers are helping us overcome this data mountain; but they can even go a step further than that, helping us come up with scientific hypotheses and explain new biological processes. Data science, in essence, enables cutting-edge biological research.

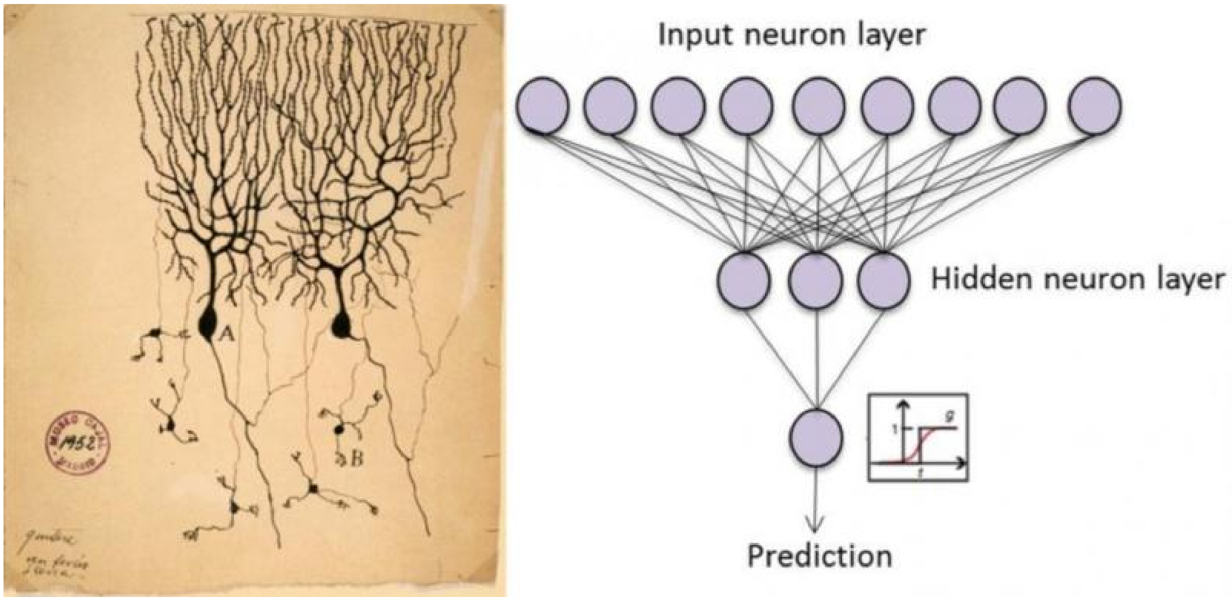
Computers to the rescue

Computers are uniquely qualified to handle massive data sets since they can simultaneously keep track of all the important conditions necessary for the analysis.

Though they could reflect human errors they're programmed with, computers can deal with large amounts of data efficiently and they aren't biased toward the familiar, as human investigators might be.

Computers can also be taught to look for specific patterns in experimental data sets – a concept termed machine learning, first proposed in the 1950s, most notably by mathematician Alan Turing. An algorithm that has learned the patterns from data sets can then be asked to make predictions based on new data it's never encountered before.

Machine learning has revolutionized biological research since we can now utilize big data sets and ask computers to help understand the underlying biology.



Left: Neurons as drawn circa 1899 by Santiago Ramón y Cajal, the father of neuroscience. Right: Schematic representation of an artificial neural network.

Training computers to "think" by simulating brain processes

We've used one interesting type of machine learning, called an artificial neural network (ANN), in our own lab. Brains are highly interconnected networks of neurons, which communicate by sending electric pulses through the neural wiring. Similarly, an ANN simulates in the computer a network of neurons as they turn on and off in response to other neurons' signals.

By applying algorithms that mimic the processes of real neurons, we can make the network learn to solve many types of problems. Google uses a powerful ANN for its now famous Deep Dream project where computers can classify and even create images.

Our group studies the immune system, with the goal of figuring out new therapies for cancer. We've used ANN computational models to study short surface protein-codes our immune cells use to determine whether something is foreign to our body and thus should be attacked. If we understand more about how our immune cells (such as T-cells) differentiate between normal/self and abnormal/foreign cells, we can design better vaccines and therapies.

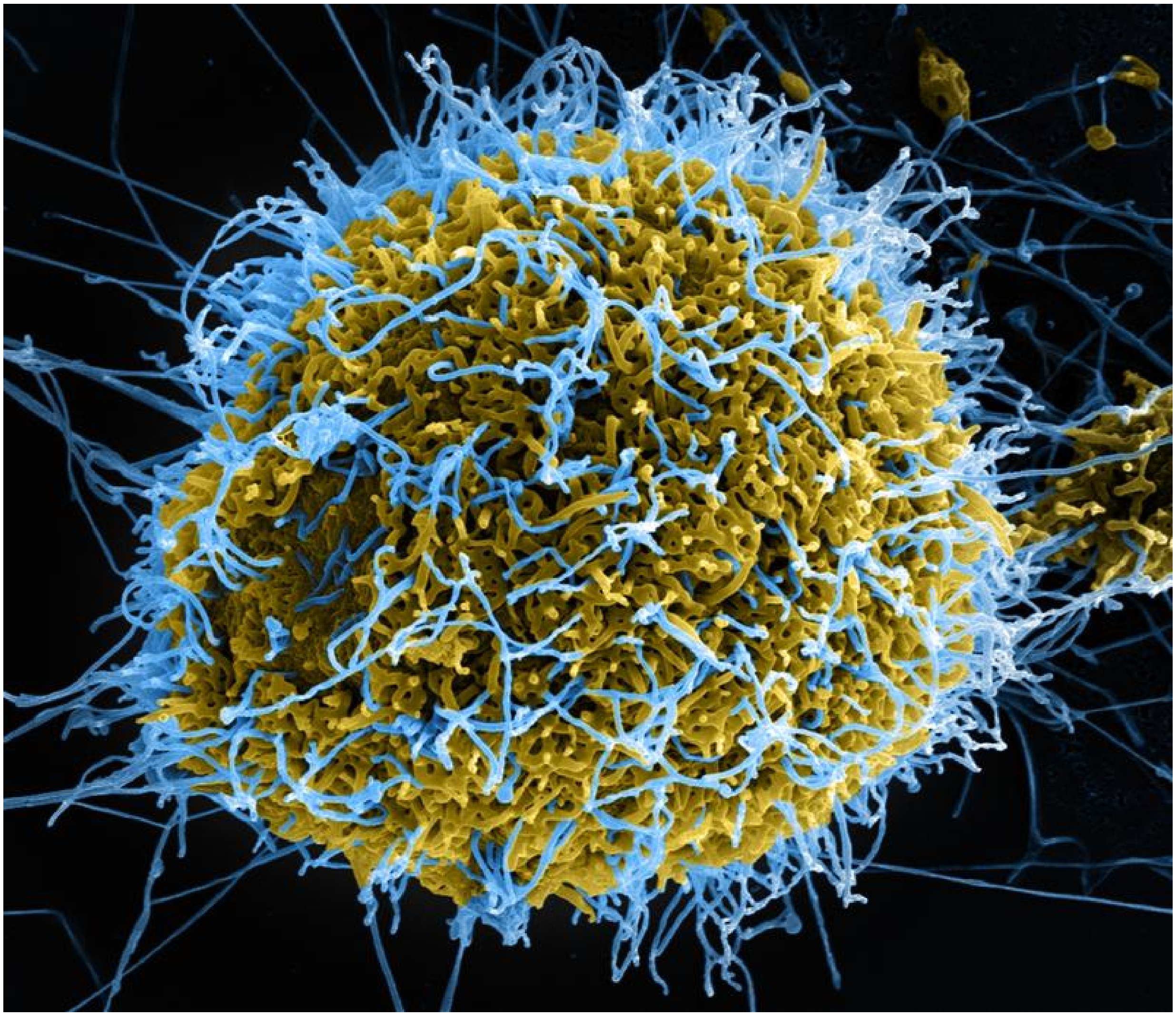
We scoured publicly available catalogs of thousands of protein-codes identified by researchers over the years. We divided this big data set into two: normal self-protein codes derived from healthy human cells, and abnormal protein-codes derived from viruses, tumors and bacteria. Then we turned to an artificial neural network developed in our lab.

Once we fed the protein-codes into the ANN, the algorithm was able to identify fundamental differences between normal and abnormal protein-codes. It would be tough for people to keep track of these kinds of biological phenomena – there are literally thousands of these protein codes to analyze in the big data set. It takes a machine to wrangle these complex problems and define new biology.

Predictions via machine learning

The most important application of machine learning in biology is its utility in making predictions based on big data. Computer-based predictions can make sense of big data, test hypotheses and save precious time and resources.

For instance, in our field of T-cell biology, knowing which viral protein-codes to target is critical in developing vaccines and treatments. But there are so many individual protein-codes from any given virus that it's very expensive and difficult to experimentally test each one.



Viruses have distinct patterns on their surfaces that our immune systems want to read and act on. Credit: National Institute of Allergy and Infectious Diseases, National Institutes of Health, CC BY

Instead, we trained the artificial neural network to help the machine learn all the important biochemical characteristics of the two types of protein-codes – normal versus abnormal. Then we asked the model to "predict" which new viral protein codes resemble the "abnormal" category and could be seen by T-cells and thus, the immune system. We tested the ANN model on different virus proteins that have never been studied before.

Sure enough, like a diligent student eager to please the teacher, the neural network was able to accurately identify the majority of such T-cell-activating protein-codes within this virus. We also experimentally tested the protein codes it flagged to validate the accuracy of the ANN's predictions. Using this neural network model, a scientist can thus rapidly predict all the important short protein-codes from a harmful virus and test them to develop a treatment or a vaccine, instead of guessing and testing them individually.

Implementing machine learning wisely

Thanks to constant refining, big data science and machine learning are increasingly becoming indispensable for any kind of scientific research. The possibilities for using computers to train and predict in biology are almost endless. From figuring out which combination of biomarkers are best for detecting a disease to understanding why only some patients benefit from a particular cancer treatment, mining big data sets using computers has become a valuable route for research.

Of course, there are limitations. The biggest problem with big data science is the data themselves. If data obtained by -omics studies are faulty to begin with, or based on shoddy science, the machines will get trained on bad data – leading to poor predictions. The student is only as good as the teacher.

Because computers are not sentient (yet), they can in their quest for patterns come up with them even when none exist, giving rise again, to bad data and nonreproducible science.

And some researchers have raised concerns about computers becoming black boxes of data for scientists who don't clearly understand the manipulations and machinations they carry out on their behalf.

In spite of these problems, the benefits of big data and machines will continue to make them valuable partners in scientific research. With caveats in mind, we are uniquely poised to understand biology through the eyes of a machine.

This story is published courtesy of The Conversation *(under Creative Commons-Attribution/No derivatives).*

Source: The Conversation

Citation: How computers help biologists crack life's secrets (2015, December 17) retrieved 25 April 2024 from https://phys.org/news/2015-12-biologists-life-secrets.html