GENETICS

Deep learning spots natural selection at work

Scientists use artificial intelligence to hunt for genetic sequences moulded by evolution.

BY AMY MAXMEN

Pinpointing where and how the human genome is evolving can be like hunting for a needle in a haystack. Each person's genome contains three billion building blocks called nucleotides, and researchers must compile data from thousands of people to discover patterns that signal how genes have been shaped by evolutionary pressures.

To find these patterns, a growing number of geneticists are turning to a form of machine learning called deep learning. Proponents of the approach say that deep-learning algorithms incorporate fewer explicit assumptions about what the genetic signatures of natural selection should look like than do conventional statistical methods.

"Machine learning is automating the ability to make evolutionary inferences," says Andrew Kern, a population geneticist at the University of Oregon in Eugene. "There is no question that it is moving things forward."

One deep-learning tool called DeepSweep, developed by researchers at the Broad Institute of MIT and Harvard in Cambridge, Massachusetts, has flagged 20,000 single nucleotides for further study. These simple mutations might have helped humans to survive disease, drought or what Charles Darwin called the "conditions of life", researchers reported last month at the annual meeting of the American Society of Human Genetics in San Diego, California.

Since the 1970s, geneticists have created mathematical models to describe the fingerprint of natural selection in DNA. If a mutation arises that renders a person better able to survive and produce offspring than their neighbours, the percentage of the population with that gene variant will grow over time.

One example is the mutation that gives many adults the ability to drink cow's milk. It enables the body to produce lactase, an enzyme that digests the sugar in milk, into adulthood. By analysing human genomes using statistical methods, researchers discovered that the mutation spread rapidly through Europe thousands of years ago — presumably because nutrients in cow's milk helped people to produce healthy children[1]. Today, nearly 80% of people of European descent carry this variant.

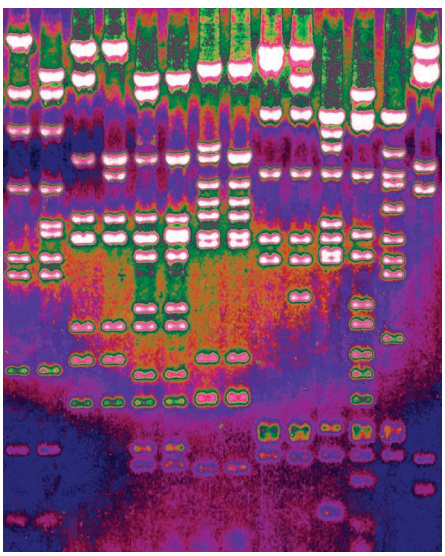
Yet geneticists have struggled to identify, and confirm, other specific snippets of the genome that spread throughout populations because they provided an adaptive edge. Deep learning excels at just this sort of task: discovering subtle patterns in large amounts of data.

But there is a catch. Deep-learning algorithms often learn to classify information after being trained by exposure to real data; Facebook, for example, primes algorithms to recognize faces using pictures that people have already labelled. Because geneticists don't yet know which parts of the genome are being shaped by natural selection, they must train their algorithms on simulated data.

To generate that simulated data, researchers need to imagine what the signature of natural selection looks like, says Sohini Ramachandran, a population geneticist at Brown University in Providence, Rhode Island. "We don't have ground-truth data, so the worry is that we may not be simulating properly."

And because deep-learning algorithms operate as black boxes, it's hard to know what criteria they use to identify patterns in data, says Philipp Messer, a population geneticist at Cornell University in Ithaca, New York. "If the simulation is wrong, it's not clear what the response means," he adds.

Researchers who use deep-learning algorithms do try to peek into the black box. DeepSweep's creators trained the algorithm on signatures of natural selection that they inserted into simulated genomes. When they tried it on real genomes, the algorithm zeroed in on the mutations that allow adults to drink milk. That bolstered the team's confidence in the tool, says Joseph Vitti, a computational geneticist at the Broad Institute who helped to develop DeepSweep.



GUY TEAR/WELLCOME COLL./CC BY

DNA fragments separated for analysis.

The researchers then sifted through data from the 1000 Genomes Project — an initiative that sequenced DNA from 2,504 people around the world — using a statistical method to identify regions that might be under evolutionary pressure. When DeepSweep examined these areas more closely, it delivered a list of 20,000 single mutations to explore.

In the coming months, Vitti and his colleagues will investigate what these mutations do by editing them in the DNA of living cells, to compare what happens when the mutations are present with when they are not.

Several other researchers are training deep-learning algorithms to search for signs of adaptation in genomes. A deep-learning model developed by Kern suggests that, at first, most mutations in humans are neither beneficial nor harmful[2]. Rather, they seem to drift along in populations, increasing natural genetic variability, and become more frequent only when a change in the environment gives people with a particular mutation an evolutionary edge.

In February, Ramachandran and her colleagues reported[3] on a deep-learning algorithm they developed, called SWIF(r). When they applied it to the genomes of 45 members of the Khomani San ethnic group in southern Africa, it flagged variations near genes associated with metabolism. The researchers speculate that the changes could have occurred thousands of years ago and helped members of the group to store fat when food was scarce.

The effects of the mutations still need to be tested. But, as with the variants spotted by DeepSweep, the candidates singled out by SWIF(r) provide scientists with a place to start.

"These are incredibly powerful methods for looking for the signals of natural selection," says Pardis Sabeti, a computational geneticist at the Broad Institute, and Vitti's PhD supervisor. "Some people didn't think you could pinpoint variants when I started. Some thought it was impossible." ■

1. Itan, Y. *et al. PLoS Comput. Biol.* **5**, e1000491 (2009).
2. Schrider, D. R. & Kern, A. D. *Mol. Biol. Evol.* **34**, 1863–1877 (2017).
3. Sugden, L. A. *et al. Nature Commun.* **9**, 703 (2018).