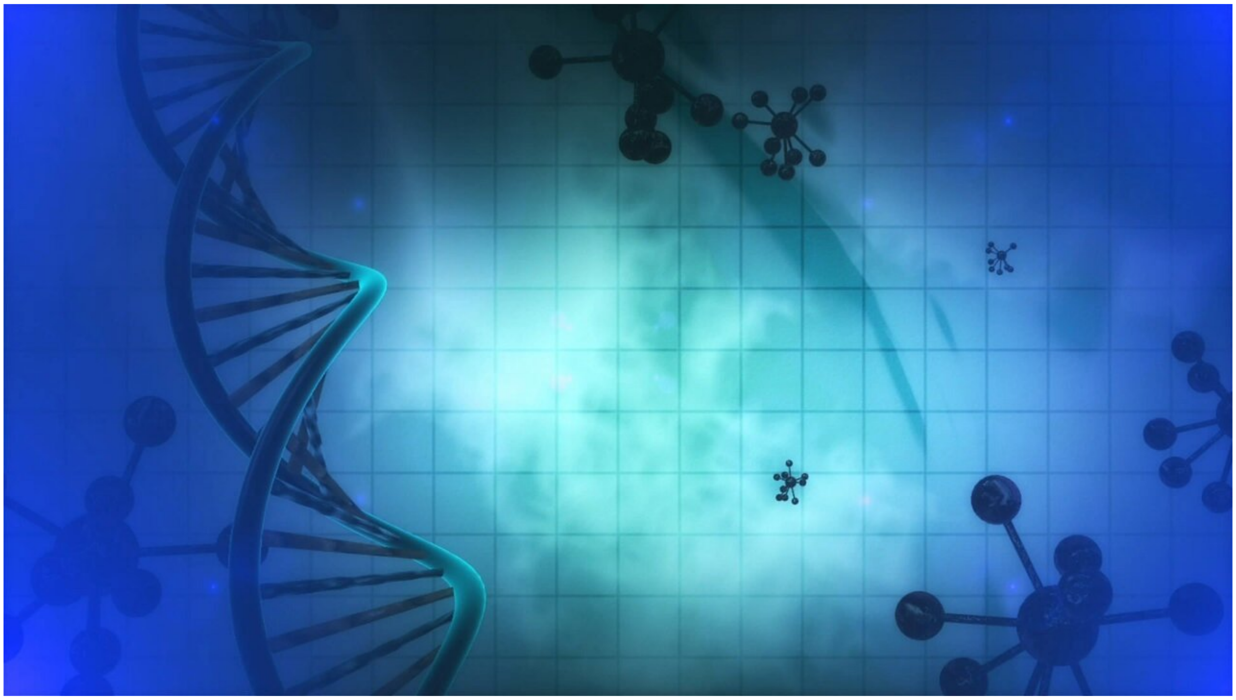
Medical Xpress

AI model shows remarkable capacity to interpret the meaning of gene variants in humans as benign or disease-causing

October 27 2021



Credit: CC0 Public Domain

No two human beings are the same, a biologic singularity encoded in the unique arrangement of the molecules that make up our individual DNA.

Variation is a cardinal feature of biology, the driver of diversity, and the engine of evolution, but it has a dark side. Alterations in DNA sequences and the resulting proteins that build our cells can sometimes lead to profound disruptions in physiologic function and cause disease.

But which gene alterations are normal or at least inconsequential, and which ones portend disease?

The answer is clear for a handful of well-known genetic mutations, yet despite dramatic leaps in genome sequencing technology over the past 20 years, our ability to interpret the meaning of millions of genetic variations identified through such sequencing still lags behind.

To make sense of it all, researchers at Harvard Medical School and Oxford University have designed an AI tool called EVE (Evolutionary model of Variant Effect), which uses a sophisticated type of machine learning to detect patterns of genetic variation across hundreds of thousands of nonhuman species and then use them to make predictions about the meaning of variations in human genes.

In an analysis published Oct. 27 in *Nature*, the researchers used EVE to assess 36 million protein sequences and 3,219 disease-associated genes across multiple species.

The results suggest that 256,000 previously identified human gene variants currently of unknown significance should, in fact, be reclassified as either benign or disease-causing.

The tool, the researchers said, can be used to augment current clinical methods used to determine the meaning of gene variants. And, when used in combination with such tools, EVE could boost the precision and accuracy of diagnosis, prognosis, and treatment choice.

"Increasingly, people have access to sequencing their genomes, but making sense of the data is not always straightforward. There is very little information about what it even means for likelihood of disease or disease progression," said study senior author Debora Marks, associate professor of systems biology in the Blavatnik Institute at HMS, who co-led the research with colleague Yarin Gal at Oxford University, co-first authors Jonathan Frazer and Mafalda Dias at Harvard Medical School, and Pascal Notin at Oxford.

The researchers emphasize that EVE is not a diagnostic test, but its computational prowess can augment current clinical tools used by geneticists and other physicians to make diagnoses, predict disease progression, and even choose treatment based on the presence of certain disease-causing genetic mutations.

"We believe our approach can be used as an added tool in current clinical assessments and offers a powerful new way to reduce uncertainty and clarify decision-making, particularly in the clinical setting," Marks added.

Indeed, the analysis showed that EVE outperformed other computational prediction models in predicting clinical effect and also scored as high as or better than current gold-standard high-throughput experiments that test the effect of a mutation on biologic function.

The stakes of accurately interpreting the meaning of genetic variation are enormous. Reading a benign variation as disease-causing could lead to erroneous diagnosis, fueling a cascade of further testing, anxiety, and even unnecessary medical interventions. Conversely, misinterpreting a disease-fueling change on one's DNA as inconsequential or benign could provide false reassurance when watchful observation, further testing, and preventive measures may be needed.

"What we hope this approach will do is generate powerful data that can empower the clinicians on the frontlines to make the right diagnostic, prognostic, and treatment decisions," Gal said.

With more data, more questions

The historic sequencing of the human genome in 2003 established a reference human genome against which newly sequenced ones are compared. Yet, this reference genome is not a standard or a baseline for a "normal" human genome. The rapidly growing amount of data from DNA sequencing renders the reference genome less of a standard and more of a fluid baseline that shifts over time as researchers better grasp the meaning of genetic variation.

Relating specific changes in the human genome to disease occurrence continues to bedevil the field of clinical genetics because the number of variants in the human population dwarfs the number scientists can investigate. Even though only a tiny fraction of the human population has been sequenced, researchers are already seeing millions of variants whose significance and meaning are unclear. Of those variants, only 2 percent are classified as benign, neutral, or pathogenic. The remaining 98 percent of the identified gene variants are currently deemed of "unknown significance."

In the human genome, protein-coding regions alone account for millions of observed 6.5 million mutations involving the position of a single amino acid in a protein made by a gene. These so-called missense mutations may have no effect on the function of a protein, or they may render the protein dysfunctional, causing disease. In fact, researchers estimate there may be a variant for every protein position—save for lethal ones—in the genomes of the 9 billion people inhabiting the planet. Every individual has many variants in their genome, compared with other people and with the reference human genome.

Adding another twist to an already compilated plot, humans inherit two versions of each gene—one from each parent. And, as people age, genes may acquire changes, known as somatic mutations.

"There's many ways in which one person doesn't just have one genome," Marks said. "You may have a different variant on one copy of a gene and, as we age, there are all sorts of somatic variations that occur—not only related to cancer development but to neurodegeneration, both of which are age-related processes driven by mutation."

To be sure, there are a number of disease-associated genes for which researchers have identified mutations that carry high risk of clinical disease, such as the *BRCA1* and *BRCA2* for breast and ovarian cancers and tumor-suppressor gene *p53* for a range of cancers. But even those genes have shown other unstudied mutations, the significance of which remains unclear.

All of this creates an urgent need to clarify the significance of genetic variations in humans—a process in which computation is going to play an increasingly important role in providing answers, Marks said.

Enter AI

A defining feature of neural networks is their capacity to continually reassess and update the probability of a hypothesis as new data become available. This means that neural networks can reevaluate evidence using new knowledge and therefore can detect patterns and meanings missed by traditional methods.

In the current study, the researchers used a sophisticated type of analysis known as unsupervised machine learning, a form of artificial intelligence that is not based on predefined parameters and rules but instead involves adaptive learning. What this means is that when presented with new data, a machine learning algorithm will become better at recognizing patterns over time. By contrast, in supervised machine learning, the algorithm learns to detect patterns from prelabeled data—its training has been supervised.

In a classic example given by informaticians, the algorithm is presented with cat and dog images and told which ones are which before it gets challenged to recognize unlabeled images of cats and dogs. In unsupervised machine learning, the algorithm is given a set of cat and dog images and not told which ones are which. It must discern the patterns on its own.

Both types offer advantages for specific tasks. One advantage of unsupervised models is that there is no chance of biasing their learning by feeding them prelabeled data. Also, they can adapt as the data change to perform more complex analyses. Most current computational methods used to assess the significance of gene variants employ supervised training based on clinical labels, which may bias these tools and cause inflated accuracy of prediction in the real world, the researchers said.

"Because the algorithm doesn't need to know in advance which images are cats, which images are dogs—it just needs a bunch of images of cats and dogs—there's no way of using information that it shouldn't know," Gal said.

It is precisely the ability of unsupervised machine learning to detect new patterns from never-before encountered data that renders this approach especially suitable for analyzing genetic sequences from non-humans.

Clues from our evolutionary relatives

In this work, the researchers turned to an old hope—that by studying genetic variation across multiple species, they might glean clues about the significance of variation in humans.

Evolution tends to preserve features that are critical, or at least important, to function and survival across species. Thus, amino acid arrangements that recur across species are markers of biologic importance, indicating they are important to an organism's function and its evolutionary fitness. Thus, alterations to such highly conserved sequences likely spells trouble and are linked to pathogenicity.

"These species are a long way away evolutionarily speaking, and there are many genetic differences, but taken together, they give us information," Marks said. "This is why the model is so powerful about patterns that are relevant for humans and human variation."

EVE looked for evolutionarily conserved patterns to draw conclusions. It analyzed data from 140,000 species, including endangered and extinct organisms.

Scientists have used comparative genetics for many years to detect regions of similarity across DNA or protein sequences to draw meaning. The Harvard-Oxford team used a neural network to do so on a much greater scale.

Training EVE

After training on 250 million protein sequences, EVE estimated the likelihood of each single amino acid variant to be either benign or pathogenic. To determine whether EVE was making accurate predictions, the researchers compared its scores with established human mutations whose significance is known. The tool's results were remarkably consistent with the clinical data, the team found.

Next, the researchers applied EVE to a set of 3,219 human genes associated with disease. EVE made the right call on whether the mutation was pathogenic or benign across all genes, including 60 "clinically actionable" genes, the researchers said. When researchers compared EVE's performance with that of other supervised and unsupervised tools, it showed notably greater accuracy of prediction.

But how would EVE's predictions fare compared with findings made from actual clinical experiments, the gold standard of assessing how a genetic mutation affects physiologic function?

To answer this question, the team compared EVE's scores against results from clinical experiments involving well-studied mutations in five genes, among them genes related to various forms of cancer, several cancer syndromes, and heart rhythm disorders. EVE's predictions overlapped with current labels from experimental data.

"Our results turned out to be far better than we expected," Marks said. "It seems that by simply training a model to fit the distribution of sequences across evolution we extract information which enables us to make unexpectedly precise predictions about disease risk arising from a given genetic variant."

A matter of trust

A notable advantage that EVE has over current methods is that it assigns a continuous score rather than a binary score. This is because even when gene variants are labeled as benign or pathogenic, how a mutation might manifest physiologically is more nuanced.

"There's a whole continuum of pathogenicity," Marks said. "The continuous score is very important for predicting what the level of pathogenicity is. Does the mutation mean I am going to get pain in my little toe, or am I going to die tomorrow?"

Another important aspect of the tool is that it assigns a confidence-of-prediction score on a gene-by-gene basis. This can help clinicians contextualize the degree of certainty for any prediction. In other words, for each genetic variant, EVE tells the expert how much they can trust its call. This is a matter of trustworthiness, of confidence in the model, the researchers said.

"We're not providing clinicians merely with a number but also giving them the degree of uncertainty that comes with it," Gal said. "This is something that the expert can take and use in the decision-making process. The tool can say, 'I think that variant belongs to that pile, but I've never seen any variants like that before so take that with a grain of salt.' Or the tool can also say, 'I think that that other variant belongs to this pile, and I've seen very similar variants to that in the past, and I saw them belonging to this pile and therefore I'm going to assign it to this pile with high confidence.' Building trust between the tool and the expert is an important aspect of this work."

Looking to the future

This type of modeling is still in its infancy, and it's clear that evolution and genetic variation still have much to teach us about disease, the researchers said, adding that they plan to extend the work to other parts of the genome beyond protein-coding regions.

In the immediate future, however, the urgent task is to make clinical use of the genetic variation for which we do have some understanding. To do so, the researchers have already teamed up with a genome-sequencing company and are collaborating with various groups via the Chan Zuckerberg Initiative.

The team is also participating in the Atlas of Variant Effects Alliance, a global research effort whose mission is to map the effects of variation across the genome and create a comprehensive atlas of all possible human gene variants and their effects on protein function and physiology. The ultimate goal of the effort is to improve the diagnosis, prognosis, and treatment of human disease.

Study co-authors included Aidan Gomez of Oxford University and Joseph Min and Kelly Brock of Harvard Medical School.

More information: Debora Marks, Disease variant prediction with deep generative models of evolutionary data, *Nature* (2021). DOI: 10.1038/s41586-021-04043-8. www.nature.com/articles/s41586-021-04043-8

Provided by Harvard Medical School