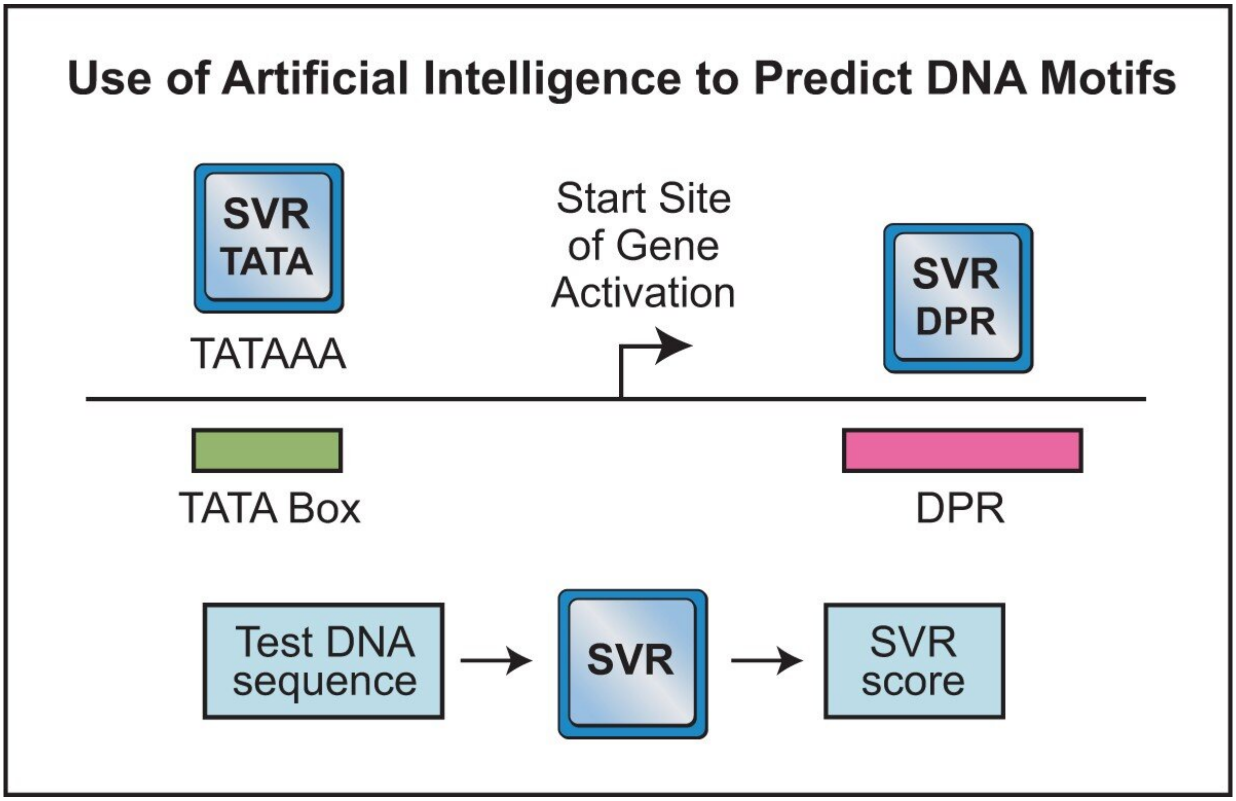
PHYS ORG

Machine learning aids gene activation discovery

September 9 2020



UC San Diego scientists have solved a long-standing puzzle in human gene activation. The discovery described in the journal Nature could be used to control gene activation in biotechnology and biomedical applications. Credit: Kadonaga Lab, UC San Diego

Scientists have long known that human genes spring into action through instructions delivered by the precise order of our DNA, directed by the four different types of individual links, or "bases," coded A, C, G and T.

Nearly 25% of our genes are widely known to be transcribed by sequences that resemble TATAAA, which is called the "TATA box." How the other three-quarters are turned on, or promoted, has remained a mystery due to the enormous number of DNA base sequence possibilities, which has kept the activation information shrouded.

Now, with the help of artificial intelligence, researchers at the University of California San Diego have identified a DNA activation code that's used at least as frequently as the TATA box in humans. Their discovery, which they termed the downstream core promoter region (DPR), could eventually be used to control gene activation in biotechnology and biomedical applications. The details are described September 9 in the journal *Nature*.

"The identification of the DPR reveals a key step in the activation of about a quarter to a third of our genes," said James T. Kadonaga, a distinguished professor in UC San Diego's Division of Biological Sciences and the paper's senior author. "The DPR has been an enigma—it's been controversial whether or not it even exists in humans. Fortunately, we've been able to solve this puzzle by using machine learning."

In 1996, Kadonaga and his colleagues working in fruit flies identified a novel gene activation sequence, termed the DPE (which corresponds to a portion of the DPR), that enables genes to be turned on in the absence of the TATA box. Then, in 1997, they found a single DPE-like sequence in humans. However, since that time, deciphering the details and prevalence of the human DPE has been elusive. Most strikingly, there have been only two or three active DPE-like sequences found in the tens of thousands of human genes. To crack this case after more than 20 years, Kadonaga worked with lead author and post-doctoral scholar Long Vo ngoc, Cassidy Yunjing Huang, Jack Cassidy, a retired computer scientist who helped the team leverage the powerful tools of artificial intelligence, and Claudia Medrano.

In what Kadonaga describes as "fairly serious computation" brought to bear in a biological problem, the researchers made a pool of 500,000 random versions of DNA sequences and evaluated the DPR activity of each. From there, 200,000 versions were used to create a machine learning model that could accurately predict DPR activity in human DNA.

The results, as Kadonaga describes them, were "absurdly good." So good, in fact, that they created a similar machine learning model as a new way to identify TATA box sequences. They evaluated the new models with thousands of test cases in which the TATA box and DPR results were already known and found that the predictive ability was "incredible," according to Kadonaga.

These results clearly revealed the existence of the DPR motif in human genes. Moreover, the frequency of occurrence of the DPR appears to be comparable to that of the TATA box. In addition, they observed an intriguing duality between the DPR and TATA. Genes that are activated with TATA box sequences lack DPR sequences, and vice versa.

Kadonaga says finding the six bases in the TATA box sequence was straightforward. At 19 bases, cracking the code for DPR was much more challenging.

"The DPR could not be found because it has no clearly apparent sequence pattern," said Kadonaga. "There is hidden information that is encrypted in the DNA sequence that makes it an active DPR element. The machine learning model can decipher that code, but we humans cannot."

Going forward, the further use of artificial intelligence for analyzing DNA sequence patterns should increase researchers' ability to understand as well as to control gene activation in human cells. This knowledge will likely be useful in biotechnology and in the biomedical sciences, said Kadonaga.

"In the same manner that machine learning enabled us to identify the DPR, it is likely that related artificial intelligence approaches will be useful for studying other important DNA sequence motifs," said Kadonaga. "A lot of things that are unexplained could now be explainable."

More information: Identification of the human DPR core promoter element using machine learning, *Nature* (2020). DOI: 10.1038/s41586-020-2689-7 , www.nature.com/articles/s41586-020-2689-7

Provided by University of California - San Diego

Citation: Machine learning aids gene activation discovery (2020, September 9) retrieved 8 May 2024 from https://phys.org/news/2020-09-machine-aids-gene-discovery.html