PHYS.ORG

Machine learning method illuminates fundamental aspects of evolution

May 8 2023, by Adam Kohlhaas

ZOONOMIA
Diverse genomes reveal mammalian secrets p. 356
DAUBENTONIA MADAGASCARIENSIS
CHOLOEPUS HOFFMANNI
PHATAGINUS TRICUSPIS
ECHINOPS TELFAIRI

Credit: *Science*

A team of researchers in Carnegie Mellon University's Computational

Biology Department (CBD) have developed new methods to identify parts of the genome critical to understanding how certain traits of species evolved.

The work, published in *Science* and led by School of Computer Science Assistant Professor Andreas Pfenning, contributes to the Zoonomia Project, an effort to sequence the entire genomes of 240 mammals to shed light on fundamental aspects of genes and traits with important implications for protecting human health and conserving biodiversity. Making sense of these new, large data sets requires the latest in artificial intelligence (AI) and machine learning (ML) technology.

Certain parts of the genome known as coding DNA provide instructions for producing proteins, the indispensable regulators of cell function. Over time, slight differences arise in the instructions that coding DNA provides for protein production, becoming one of the driving forces behind evolution.

Yet these protein-producing DNA pieces account for a meager one percent of the three billion nucleotide pairs that make up the human genome. Other noncoding DNA regions, known as enhancers, determine when and where specific genes are active.

The CMU team created an ML approach called the Tissue-Aware Conservation Inference Toolkit (TACIT) to learn more about how these areas operate. While a traditional model of evolution might demonstrate changes in a species' brain size through a set of mutations in a group of genes, enhancers may simply turn genes on or off and achieve the same result.

Most research into the evolution of mammals focuses on the parts of the genome that have changed relatively little over millions of years. These conserved regions, especially genes, provide insight into fundamental elements in mammalian DNA that highlight unique traits in individual species.

The challenge for Pfenning and his team is that, over time, the DNA enhancer regions may change in sequence but not in function. For example, a well-studied Islet enhancer regulates gene levels in similar patterns across humans, mice, zebra fish and sponges, despite more than 700 million years of evolution. This makes them much more difficult to identify and track using traditional methods of examining individual nucleotides.

TACIT confronts this problem by accurately predicting if an enhancer will be active in a particular cell type or tissue. It allows scientists to identify these important enhancer regions in a newly sequenced genome without conducting a new laboratory experiment, offering potential applications in conservation biology. The toolkit can make predictions about how enhancers function in endangered or threatened species, where controlled laboratory experiments are impossible.

"TACIT provides an unprecedented opportunity to predict the function of parts of the genome outside of genes in species for which we cannot get primary tissue samples, such as the bottlenose dolphin and the critically endangered black rhinoceros," said Irene Kaplow, a lead author on the paper and a postdoctoral associate and Lane Fellow in CBD. "As ML methods and methods for identifying enhancers from specific cell types improve, I anticipate that we will be able to broaden the functions of TACIT to provide new kinds of insights into mammalian evolution."

After predicting the function of genomic sequences across the 240 mammals, the research team applied TACIT to identify the parts of the genome that have evolved in mammals for larger brains and found that those tended to be near genes whose mutations have been implicated in human brain-size disorders. They also identified an enhancer associated with social behavior across mammals that is specific to a particular subtype of neuron, the parvalbumin positive inhibitory interneuron.

"We think this is just the tip of the iceberg," said Pfenning, senior author of the study. "We found interesting relationships by applying TACIT to a small number of tissues and small number of traits, but there is still a lot more to discover."

More information: Irene M. Kaplow et al, Relating enhancer genetic variation across mammals to complex phenotypes using machine learning, *Science* (2023). DOI: 10.1126/science.abm7993

Provided by Carnegie Mellon University

Citation: Machine learning method illuminates fundamental aspects of evolution (2023, May 8) retrieved 28 April 2024 from https://phys.org/news/2023-05-machine-method-illuminates-fundamental-aspects.html