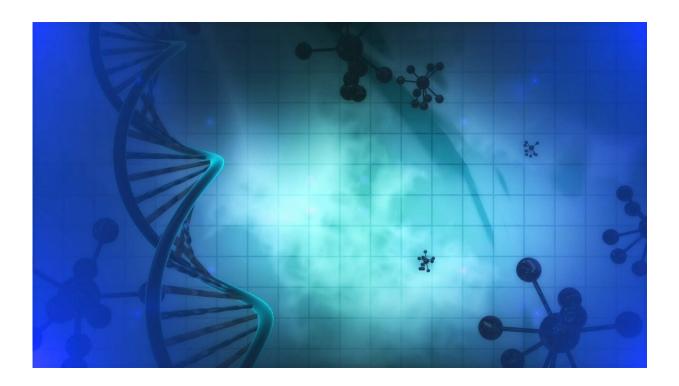# Medical Xpress

Biologists pioneer first method to decode gene expression

August 12 2019

Credit: CC0 Public Domain

Given the recent remarkable advancements in genetics, it's easy to assume that 21st century scientists have at their disposal a clear, quick way to run a genomic sequence scan and find out which genes among thousands can be expressed and which cannot. Gene expression is the process by which information encoded within genes leads to key products, such as proteins.

Surprisingly, that hasn't been possible until now. Biologists at the University of California San Diego have developed the first system for determining gene expression based on machine learning. Given the lack of such a method, the new process is considered a type of genetic Rosetta Stone for biologists.

"This paper represents the first method to distinguish genes that can be expressed from those that cannot," said Steve Briggs, a Division of Biological Sciences professor and senior author of the paper. "This is the basis for all of biology. Whether it's drug discovery or plant breeding or evolution, this touches the basic studies of biology."

The method, developed by graduate student Ryan Sartor, Briggs and their colleagues, is described August 12, 2019 in the *Proceedings of the National Academy of Sciences*.

Biologists have previously classified gene expression through experimental observations and scientific literature references. But the genomics field lacked a formalized process for revealing this information, called the "expressible gene set," or EGS, which comprises all protein-coding genes with the potential to be expressed.

"In biology, there is no method to do this," said Briggs. "In the past we've just had empirical approaches to making catalogs—we haven't had scientific criteria that classifies the genes based on their molecular features."

The new method leverages machine learning, the use of algorithms and other processes to analyze data, and is based on an example set of nearly 30,000 maize plant genes containing specific, detailed molecular features. An advanced algorithm was trained on the data and "learned" to classify gene expression at 99.4 percent accuracy.

The key to the advancement is bringing together chromatin biology, which contributes to regulating the DNA packaging within cells, with molecular features that are known to determine gene expression. Combining these with mathematical machine learning, the new method of determining the species-wide set of transcribed genes, or "expressome," then creates an atlas of expressible genes. The method may also be useful in understanding evolutionary mechanisms that silence certain genes.

Briggs is now applying the method to sorghum, an important grain for food and fodder, but says it can be useful beyond plant species. Ultimately, he says the new method is like a word decoder.

"The genome sequence is like a book," said Briggs. "The words are the genes. Until now, we couldn't tell which DNA sequences were real words and which merely resembled words. By removing non-words we now have a much more accurate reading of the book."

More information: Ryan C. Sartor el al., "Identification of the expressome by machine learning on omics data," *PNAS* (2019). www.pnas.org/cgi/doi/10.1073/pnas.1813645116

Provided by University of California - San Diego