# Researchers develop program to read any genome sequence and decipher its genetic code

November 9 2021

Credit: CC0 Public Domain

Yekaterina "Kate" Shulgina was a first year student in the Graduate School of Arts and Sciences, looking for a short computational biology project so she could check the requirement off her program in systems biology. She wondered how genetic code, once thought to be universal, could evolve and change.

That was 2016 and today Shulgina has come out the other end of that short-term project with a way to decipher this genetic mystery. She describes it in a new paper in the journal *eLife* with Harvard biologist Sean Eddy.

The report details a new computer program that can read the genome sequence of any organism and then determine its genetic code. The program, called Codetta, has the potential to help scientists expand their understanding of how the genetic code evolves and correctly interpret the genetic code of newly sequenced organisms.

"This in it of itself is a very fundamental biology question," said Shulgina, who does her graduate research in Eddy's Lab.

The genetic code is the set of rules that tells the cells how to interpret the three-letter combinations of nucleotides into proteins, often referred to as the building blocks of life. Almost every organism, from *E. coli* to humans, uses the same genetic code. It's why the code was once thought to be set in stone. But scientists have discovered a handful of outliers—organisms that use alternative genetic codes—exist where the set of instructions are different.

This is where Codetta can shine. The program can help to identify more organisms that use these alternative genetic codes, helping shed new light on how genetic codes can even change in the first place.

"Understanding how this happened would help us reconcile why we originally thought this was impossible… and how these really fundamental processes actually work," Shulgina said.

Already, Codetta has analyzed the genome sequences of over 250,000 bacteria and other single-celled organisms called archaea for alternative genetic codes, and has identified five that have never been seen. In all five cases, the code for the amino acid arginine was reassigned to a different amino acid. It's believed to mark the first-time scientists have seen this swap in bacteria and could hint at evolutionary forces that go into altering the genetic code.

The researchers say the study marks the largest screening for alternative genetic codes. Codetta essentially analyzed every genome that's available for bacteria and archaea. The name of the program is a cross between the codons, the sequence of three nucleotides that forms pieces of the genetic code, and the Rosetta Stone, a slab of rock inscribed with three languages.

The work marks a capstone moment for Shulgina, who spent the past five years developing the statistical theory behind Codetta, writing the program, testing it, and then analyzing the genomes. It works by reading the genome of an organism and then tapping into a database of known proteins to produce a likely genetic code. It differs from other similar methods because of the scale at which it can analyze genomes.

Shulgina joined Eddy's lab, which specializes in comparing genomes, in 2016 after coming to him for advice on the algorithm she was designing to interpret genetic codes.

Until now, no one has done such a broad survey for alternative genetic codes.

"It was great to see new codes, because for all we knew, Kate would do all this work and there wouldn't turn out to be any new ones to find," said Eddy, who's also a Howard Hughes Medical Investigator. He also noted the potential of the system to be used to ensure the accuracy of the many databases that house protein sequences.

"Many protein sequences in the databases these days are only conceptual translations of genomic DNA sequences," Eddy said. "People mine these protein sequences for all sorts of useful stuff, like new enzymes or new gene editing tools and whatnot. You'd like for those protein sequences to be accurate, but if the organism is using a nonstandard code, they'll be erroneously translated."

The researchers say the next step of the work is to use Codetta to search for alternative codes in viruses, eukaryotes, and organellar genomes like mitochondria and chloroplasts.

"There's still a lot of diversity of life where we haven't done this systematic screening yet," Shulgina said.

**More information:** Yekaterina Shulgina et al, A computational screen for alternative genetic codes in over 250,000 genomes, *eLife* (2021). DOI: 10.7554/eLife.71402

Provided by Harvard University

genome-sequence-decipher-genetic.html