THE CONVERSATION

Academic rigour, journalistic flair

Extremophile bacteria (yellow) can be seen in Yellowstone Park's Octopus Spring. (C. M. Helm-Clark/Wikimedia Commons)

Extreme environments are coded into the genomes of the organisms that live there

Published: February 22, 2024 1.16pm EST

Kathleen A. Hill

Associate Professor Biology, Western University

Lila Kari

Professor, Computer Science, University of Waterloo

An organism's genome is a set of DNA instructions needed for its development, function and reproduction. The genome of a present-day organism contains information from its journey on an evolutionary path that starts with the "first universal common ancestor" of all life on Earth and culminates with that organism.

Encoded within itself, an organism's genome contains information that can reveal connections to its ancestors and its relatives.

Other dimensions of the genome

Our research explores the hypothesis that an organism's genome could contain other types of information, beyond genealogy or taxonomy. We asked: Could the genome of an organism contain information that would allow us to determine the type of environment the organism lives in?

Extremophiles have been found in environments such as Pitch Lake in Trinidad and Tobago, the largest asphalt deposit in the world. (Shutterstock)

As unlikely as it seems, our team of computer science and biology researchers at the University of Waterloo and Western University found that to be the case for extremophiles — organisms that live and thrive in extremely harsh conditions. These environmental conditions range from extreme heat (over 100 C) to extreme cold (below -12 C), high radiation or extremes in acidity or pressure.

DNA as a language

We looked at genomic DNA as a text written in a "DNA language." A DNA strand (or DNA sequence) consists of a succession of basic units called nucleotides, strung together by a sugar-phosphate backbone. There are four such different DNA units: adenine, cytosine, guanine and thymine (A,C,G,T).

Viewed abstractly, a DNA sequence can be thought of as a line of text, written with "letters" from the "DNA alphabet." For example, "CAT" would be the three-letter "DNA word" corresponding to the three-unit DNA sequence cytosine-adenine-thymine.

In the 1990s, it was discovered that by counting occurrences of such DNA words in a short DNA sequence extracted from the genome of an organism, one could identify the species of the organism and the degree of its relatedness to other organisms in the evolutionary "tree of life."

The mechanism of this identification or classification of an organism based on DNA word counts is similar to the process that allows us to differentiate an English book from a French book: By taking one page from each book one notices that the English text has many occurrences of the three-letter word "the," while the French text has many occurrences of the three-letter word "les."

Note that the word-frequency profile of each book is not dependent on the particular page we chose to read and on whether we considered multiple pages, a single page or an entire chapter. Similarly, the frequency profile of DNA words in a genome is not dependent on the location and on the length of the DNA sequence that was selected to represent that genome.

A DNA strand consists of a succession of basic units: adenine, cytosine, guanine and thymine (ACGT). (Shutterstock)

That DNA word-frequency profiles can act as a "genomic signature" of an organism was a significant discovery and, until now, it was believed that the DNA word-frequency profile of a genome only contained evolutionary information pertaining to the species, genus, family, order, class, phylum, kingdom or domain that the organism belonged to.

Our team set out to ask whether the DNA word-frequency profile of a genome could reveal other kinds of information — for example, information regarding the type of extreme environment that a microbial extremophile thrives in.

Environment imprints in extremophile DNA

We used a dataset of 700 microbial extremophiles living in extreme temperatures (either extreme heat or cold) or extreme pH conditions (strongly acidic or alkaline). We used both supervised machine learning and unsupervised machine learning computational approaches to test our hypothesis.

In both types of environmental conditions, we discovered that we could clearly detect an environmental signal indicating the type of extreme environment a particular organism inhabited.

In the case of unsupervised machine learning, a "blind" algorithm was given a dataset of extremophile DNA sequences (and no other information about either their taxonomy or their living environment). The algorithm was then asked to group these DNA sequences in clusters, based on whatever similarities it could find among their DNA word-frequency profiles.

The expectation was that all the clusters discovered this way would be along taxonomic lines: bacteria grouped with bacteria, and archaea grouped with archaea. To our great surprise, this was not always the case, and some archaea and bacteria were consistently grouped together, no matter what algorithms we used.

The only obvious commonality that could explain their being considered similar by the multiple machine learning algorithms was that they were heat-loving extremophiles.

A shocking discovery

The tree of life, a conceptual framework used in biology that represents geneaological relationships between species, has three major limbs, called domains: bacteria, archaea and eukarya.

Eukaryotes are organisms that have a membrane-bound nucleus, and this domain includes animals, plants, fungi and the unicellular microscopic protists. In contrast, bacteria and archaea are single-cell organisms that do not have a membrane-bound nucleus containing the genome. What distinguishes bacteria from archaea is the composition of their cell walls.

 a figure showing the three branches of the tree of life

A schematic tree of life with the primary domains, archaea and bacteria, shown in purple and blue, respectively and the secondary domain, Eukaryotes, in green. (Tara Mahendrarajah), CC BY

The three domains of life are dramatically different from each other and, genetically, a bacterium is as different from an archaeon as a polar bear (eukarya) is from an *E. coli* (bacteria).

The expectation was therefore that the genome of a bacterium and of an archaeon would be as far apart as possible in any clustering by any genomic similarity measure. Our finding of some bacteria and archaea clustered together, apparently just because they are both adapted to extreme heat, means that the extreme temperature environment they live in caused pervasive, genome-wide, systemic shifts in their genome language.

This discovery is akin to finding a completely new dimension of the genome, an environmental one, existent in addition to its well-known taxonomic dimension.

Genomic impact of other environments

Besides being unexpected, this finding could have implications for our understanding of the evolution of life on Earth, as well as guide our thinking into what it would take to live in outer space.

 an orange sphere with a tail

Pyrococcus furiosus, a thermophilic archaeon that was surprisingly grouped with thermophilic bacteria. (Michelle Kropf/Wikimedia Commons), CC BY

Indeed, our ongoing research is exploring the existence of an environmental signal in the genomic signature of radiation-resistant extremophiles, such as *Deinococcus radiodurans*, which can survive radiation exposure, as well as cold, dehydration, vacuum conditions and acid, and was shown to be able to survive in outer space for up to three years.