Extremophiles Reveal a New Dimension of the Genome

Credit: iStock.

Extremophiles, as their name suggests, are organisms that can live in extreme conditions, many of which are inhospitable for other terrestrial organisms.

These fascinating organisms have been discovered deep within the Earth's crust, in extremely acidic or basic conditions, under high pressures and in environments with blisteringly hot or freezing cold temperatures.

Extremophiles have intrigued scientists for many years; how do they not only

survive, but *thrive* in such harsh environments? Advances in [next-generation sequencing](#) are helping to answer this question, providing insights into their genetic composition.

[Professor Lila Kari](#) from the Cheriton School of Computer Science has studied genetic signatures since the early 2000s. In a bid to understand whether an organism's genome might contain information *beyond* taxonomic and evolutionary insights – i.e., its ancestry and how it is related to other organisms – she turned to extremophiles.

Using machine learning algorithms, Kari and colleagues compared genetic signatures across 700 microbial extremophiles. Their findings, published in *[Scientific Reports](#)*, were so unexpected that the team "couldn't believe their eyes."[1]

Typically, the more similar an organism's DNA is to another organism, the more closely related those organisms are. But Kari and colleagues found some extremophiles had *very* similar DNA, despite being very distantly related.

This suggests that an environmental "signature" exists within the extremophiles' genome. Two extremophiles – though distantly related – could share similar genome signatures if they have adapted to survive in the same harsh conditions, such as extreme temperature or pH.

"Our study has revealed, in some sense, a new dimension of the genome: The DNA of extremophiles contains, in addition to ancestry information, information associated with the extreme environment where they live," said Kari.

Technology Networks recently had the pleasure of interviewing Kari, where we learned more about the backstory of this study and the significance of its findings.

Molly Campbell (MC): Why did you decide to focus on extremophiles in this study? Why are they so interesting, and are there any examples of extremophiles that you think are particularly fascinating?

Lila Kari (LK): Extremophiles are very interesting creatures, as they live at the edges of survivability and beyond. Besides being fascinating to young and old people alike (my younger daughter loves tardigrades!), it is of great interest to discover the biological mechanisms that allow them to survive and even thrive in incredibly hostile environments.

Cool examples of extremophiles are the super-cute tardigrades (also known as water bears or moss piglets) that can survive exposure to extreme temperatures, extreme pressures, air deprivation, radiation, dehydration and starvation. Or the microbe *Deinococcus radiodurans,* which can survive in outer space.

The study of extremophiles has become a hot topic in recent years, when humanity is exploring outer space and searching for organisms that can survive the extremely hostile conditions of outer space, to be sent to Mars or on other space missions.

We had different reason for choosing to study extremophiles. Our group has studied "genomic signatures" since the 2000s; that is, patterns in the genomes of organisms that allow us to identify them and classify and position them correctly on the Tree of Life.

To elaborate, from a mathematical point of view:

- A genome is a long string made up of "letters" from a four-letter alphabet (A, C, G and T).

- A "DNA word" is a sequence of such letters (for example, *GGAATC* a six-letter DNA word.
- A "pattern" refers to a pattern of frequencies of such DNA words in a given genome.

These frequency patterns form a so-called "genomic signature" of an organism, and they allow us to identify the species and the degree of relatedness to other species. Intuitively, this is akin to being able to tell the difference between a French book and an English book by noticing that the English one has a high frequency of the three-letter word "the", while the French one has a high frequency of the three-letter word, "les".

In the research to date, including our own, scientists have obtained highly accurate classifications of genomes of organisms based on this method, which led to the tentative conclusion that this "genomic signature" contains exclusively taxonomic information regarding biological relatedness.

However, we had the idea that perhaps the genome contains information other than taxonomic/evolutionary, and we wanted to explore this question. For example, could it be possible that the genome also has patterns that reflect the environment in which an organism lives?

We realized that if such an "environmental signal" would exist at all, it would probably be very faint, and that our only hope to discover such a signal would be to search for it in organisms that live in extreme environments, that is – you guessed it – in extremophiles.

MC: Can you explain the purpose of using machine learning methods in this study, and how you used them?

LK: Machine learning is a very powerful methodology for classification problems, such as the taxonomic classification problem we are exploring (given a DNA fragment, what species of organism does it belong to, and how closely it is related to other species).

In supervised classification, you "train" the machine learning algorithm with examples, each consisting of a DNA fragment and the species label of that organism. At the end of training, you present the algorithm with a new DNA fragment and ask it: what is its taxonomic label? The algorithm gives you the answer based on what it has "learned" during training.

Machine learning trained with genomic signatures of DNA fragments has proved to yield highly accurate classifications of very large biological datasets (tens of thousands, or even much larger). Note that taxonomic identification and classification is a very important problem for biodiversity research, given that 95% of the multicellular species on Earth have yet to be taxonomically classified.

For our extremophile research, we used both supervised and unsupervised machine learning. In the latter, the algorithm is simply given a large dataset of DNA fragments and is asked to discover what various DNA fragments have in common and group them in clusters based on their similarities, whatever those may be. In other words, unsupervised machine learning is a "blind" algorithm that is given some unlabeled DNA fragments and no other information, and is told to "go there and find something," and group the organisms in clusters based on whatever similarities it finds.

Want more breaking news?

MC: Your data show that the environment creates a "mark" in the genome of an organism living in it, meaning organisms that are distantly related might have similarities in their genomes if they occupy similar "harsh" environments. Can you talk more about this finding and its significance?

LK: To our great astonishment, we discovered that besides the expected taxonomic signature (DNA word patterns that allow us to identify a species, and how it is related to other species), the extremophile genomes also exhibit an environmental signature. That is, word patterns that are associated with the type of environment that the extremophile lives in (e.g., certain common DNA patterns are found in extremophiles who live in very hot environments, even though taxonomically they are as distant as they can be in the Tree of Life).

It is as if, until now, we thought that the DNA/genome of an organism is like a "book". Now we discovered that it is actually like sheet music where the lyrics (the taxonomic information) are interwoven with music (the environmental information), both encoded with the same alphabet. Moreover, we discovered that, at times (e.g., in some extremophiles), the environmental signal is louder than the evolutionary signal.

This is a bit like, say, trying to classify songs of different genres, sung in different languages. Most of the time what stands out in a song is the language of the lyrics, but sometimes the music is so loud that you can only tell that both belong to the same genre (e.g. rock, or heavy metal) but you cannot tell in which language they

are sung.

Overall, I cannot stress enough how unexpected this discovery is: It is akin to discovering a completely new "dimension" of the genome.

MC: Are there any examples of extremophiles with similar genome patterns that you think are particularly "surprising" or exciting?

LK: Yes, we found several concrete examples of completely unrelated organisms with similar genomic signatures caused by the environment. Now I am exaggerating a bit here, since all life on Earth is related, as descendants of the "Last Universal Common Ancestor". What I mean is that they are as distantly related as they can be, i.e., some being bacteria and some archaea, with bacteria and archaea being two of the only three distinct "domains" of life" alongside eukaryotes.

These unrelated organisms were grouped together as similar by all the machine learning algorithms we used, including the "blind" unsupervised ones, in spite of being more different taxonomically from each other than a lichen is from a polar bear.

The only feature that these grouped-together extremophiles had in common was that they were all thermophiles, that is, they lived in extremely hot environments. Among them is one of my favourite extremophiles, the archaeon *Pyrococcus furiosus*. You have to love that name, which comes from the Greek "pyrococcus", meaning fireball, and the Latin "furiosus", meaning furiously, and refers to its furious swimming at temperatures of over 100 °C!

P. furiosus was grouped together with three thermophile bacteria, which was completely jaw-dropping given how different they are taxonomically.

MC: Are there any limitations to this research that you think it is important to highlight? If so, how could future work overcome such limitations?

LK: Every method has limitations, including machine learning algorithms. First, their classification/clustering accuracies, while high, are almost never 100%, so we are working on improving the accuracy of the machine learning methods that we used in this study.

In addition, a limitation of machine learning algorithms is that they are "black box" methods, in the sense that while they output a classification or clustering, they do not offer a rationale for their output. More research is needed to be able to understand and interpret the results of the machine learning classifications/clustering.

MC: What are your next research steps?

LK: We are currently working on several research directions, including the exploration of the existence of an environmental signature in radiation-resistant extremophiles, such as *Deinococcus radiodurans*, which was recently proved by scientists to be able to survive outer space conditions for one to three years.[2,3]

The next steps would then be to extend our exploration to polyextremophiles (organisms that thrive under multiple stresses in extreme environments, e.g., both high temperature and high acidity), and to multicellular extremophilic organisms such as tardigrades or extremophilic plants.

References:

1. Arias PM, Butler J, Randhawa GS, Soltysiak MPM, Hill KA, Kari L. Environment and

taxonomy shape the genomic signature of prokaryotic extremophiles. *Sci Rep.* 2023;13(1):16105. doi: [10.1038/s41598-023-42518-y](https://doi.org/10.1038/s41598-023-42518-y)

2. Ott E, Kawaguchi Y, Kölbl D, et al. Molecular repertoire of *Deinococcus radiodurans* after 1 year of exposure outside the International Space Station within the Tanpopo mission. *Microbiome*. 2020;8(1):150. doi: [10.1186/s40168-020-00927-5](https://doi.org/10.1186/s40168-020-00927-5)

3. Kawaguchi Y, Shibuya M, Kinoshita I, et al. DNA damage and survival time course of deinococcal cell pellets during 3 years of exposure to outer space. *Front Microbiol*. 2020;11. doi: [10.3389/fmicb.2020.02050](https://doi.org/10.3389/fmicb.2020.02050)