Cracking the code of life: New AI model learns DNA's hidden language

August 5 2024, by Magdalena Gonciarz

An artistic representation of the large language model trained on the DNA sequences. Credit: Magdalena Gonciarz, generated with Dall-E3

DNA contains foundational information needed to sustain life. Understanding how this information is stored and organized has been one of the greatest scientific challenges of the last century.

With GROVER, a new large language model trained on human DNA,

researchers could now attempt to decode the complex information hidden in our genome.

Developed by a team at the Biotechnology Center (BIOTEC) of Dresden University of Technology, GROVER treats human DNA as a text, learning its rules and context to draw functional information about the DNA sequences. This new tool, published in *Nature Machine Intelligence*, has the potential to transform genomics and accelerate personalized medicine.

Since the discovery of the double helix, scientists have sought to understand the information encoded in DNA. 70 years later, it is clear that the information hidden in the DNA is multilayered. Only 1–2 % of the genome consists of genes, the sequences that code for proteins.

"DNA has many functions beyond coding for proteins. Some sequences regulate genes, others serve structural purposes, most sequences serve multiple functions at once. Currently, we don't understand the meaning of most of the DNA. When it comes to understanding the non-coding regions of the DNA, it seems that we have only started to scratch the surface. This is where AI and large language models can help," says Dr. Anna Poetsch, research group leader at BIOTEC.

DNA as a language

Large language models, like GPT, have transformed our understanding of language. Trained exclusively on text, the large language models developed the ability to use the language in many contexts.

"DNA is the code of life. Why not treat it like a language?" says Dr. Poetsch. The Poetsch team trained a large language model on a reference human genome. The resulting tool named GROVER, or "Genome Rules Obtained via Extracted Representations," can be used to extract

biological meaning from the DNA.

"GROVER learned the rules of DNA. In terms of language, we are talking about grammar, syntax, and semantics. For DNA, this means learning the rules governing the sequences, the order of the nucleotides and sequences, and the meaning of the sequences. Like GPT models learning human languages, GROVER has basically learned how to 'speak' DNA," explains Dr. Melissa Sanabria, the researcher behind the project.

The team showed that GROVER can not only accurately predict the following DNA sequences but can also be used to extract contextual information that has biological meaning, e.g., identify gene promoters or protein binding sites on DNA. GROVER also learns processes that are generally considered to be "epigenetic," i.e., regulatory processes that happen on top of the DNA rather than being encoded.

"It is fascinating that by training GROVER with only the DNA sequence, without any annotations of functions, we are actually able to extract information on biological function. To us, it shows that the function, including some of the epigenetic information, is also encoded in the sequence," says Dr. Sanabria.

The DNA dictionary

"DNA resembles language. It has four letters that build sequences and the sequences carry a meaning. However, unlike a language, DNA has no defined words," says Dr. Poetsch. DNA consists of four letters (A, T, G, and C) and genes, but there are no predefined sequences of different lengths that combine to build genes or other meaningful sequences.

To train GROVER, the team had to first create a DNA dictionary. They used a trick from compression algorithms. "This step is crucial and sets

our DNA language model apart from the previous attempts," says Dr. Poetsch.

"We analyzed the whole genome and looked for combinations of letters that occur most often. We started with two letters and went over the DNA, again and again, to build it up to the most common multi-letter combinations. In this way, in about 600 cycles, we have fragmented the DNA into 'words' that let GROVER perform the best when it comes to predicting the next sequence," explains Dr. Sanabria.

The promise of AI in genomics

GROVER promises to unlock the different layers of genetic code. DNA holds key information on what makes us human, our disease predispositions, and our responses to treatments.

"We believe that understanding the rules of DNA through a language model is going to help us uncover the depths of biological meaning hidden in the DNA, advancing both genomics and personalized medicine," says Dr. Poetsch.

More information: Melissa Sanabria et al, DNA language model GROVER learns sequence context in the human genome, *Nature Machine Intelligence* (2024). DOI: 10.1038/s42256-024-00872-0

Provided by Dresden University of Technology