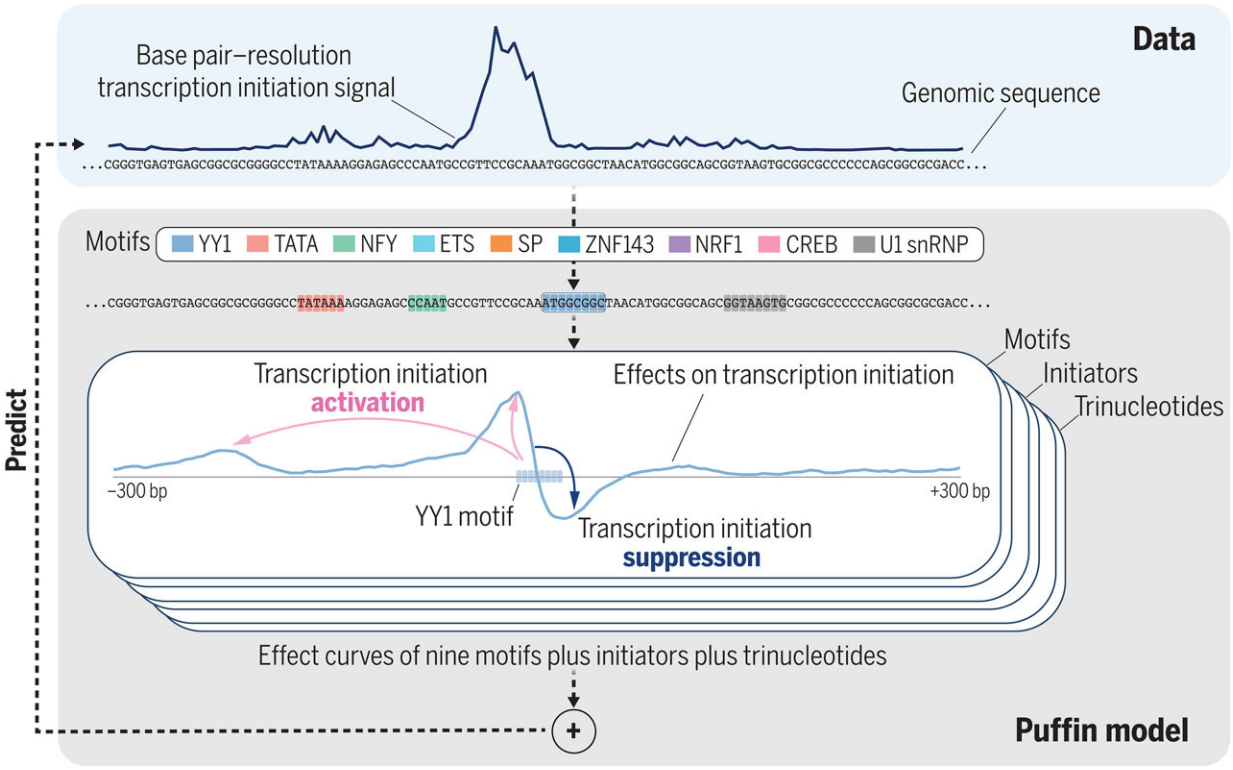
Machine learning sheds light on gene transcription

May 14 2024



A unified model that explains the sequence basis of transcription initiation in the human genome. Puffin predicts transcription initiation signals by first detecting sequence patterns that appear in the DNA sequence and then applying the effects of every sequence pattern on the transcription initiation signal. The model includes three types of sequence patterns: motifs, initiators, and trinucleotides. Strand-specific base pair–resolution transcription initiation signals are predicted by combining motif effects additively in log scale and then transforming to output scale. bp, base pairs. Credit: *Science* (2024). DOI: 10.1126/science.adj0116

A team led by researchers at UT Southwestern Medical Center has developed deep learning models to identify a simple set of rules that govern the activity of promoters—regions of DNA that initiate the process by which genes produce proteins.

Their findings, published in *Science*, could lead to a better understanding of how promoters contribute to gene regulation in health and disease.

"Although promoters are essential for the function of every gene, our understanding of how these genetic elements operate is incomplete despite decades of study that have defined many of their features. Our research sheds new light on how these sequences work in humans and other mammals," said Jian Zhou, Ph.D., Assistant Professor in the Lyda Hill Department of Bioinformatics at UT Southwestern.

Dr. Zhou co-led the study with first author Kseniia Dudnyk, a graduate student in the Zhou Lab, and Jian Xu, Ph.D., a former researcher at the Children's Medical Center Research Institute at UT Southwestern.

Creating the proteins that cells use to perform their activities starts with a process known as transcription. That's when an RNA polymerase protein latches onto a DNA strand and copies—or transcribes—the encoded information into an RNA molecule. The region where the RNA polymerase attaches to begin transcription is called the promoter.

In humans, promoters are typically composed of hundreds of base pairs, the units that make up DNA. Although researchers have identified common base pair sequences shared among some regions of DNA that are promoters, these sequences are often absent in human promoters, leaving the rules of how DNA sequence directs the transcription process unclear.

To better define promoters in humans and how they operate, the researchers developed a machine learning program they named Puffin. After analyzing data from tens of thousands of recognized human promoters, the program determined that they are made of three types of sequence patterns: motifs, initiators, and trinucleotides.

Puffin showed that depending on how these elements are arranged, they can activate or repress transcription of a gene. Puffin also can predict how the arrangement of these elements can direct RNA polymerase to preferentially transcribe a single strand of DNA or transcribe both strands simultaneously toward opposite directions. This bidirectional transcription is common in human genes.

The program further showed that mice and other mammals shared similar rule sets for governing promoter operation. In addition, Puffin allowed the researchers to predict whether and how transcription would occur if they mutated promoters, findings that closely matched those from experiments.

The study authors suggested that Puffin could help them understand how promoters work in healthy cells as well as how disease-associated alterations in promoters could lead to changes in gene transcription.

This program is available on a free web server (tss.zhoulab.io) so that other researchers can test any promoter sequence of interest. They added that using a similar machine learning approach could offer insights into other facets of the genome that are still not well understood.

More information: Kseniia Dudnyk et al, Sequence basis of transcription initiation in the human genome, *Science* (2024). DOI: 10.1126/science.adj0116

Provided by UT Southwestern Medical Center