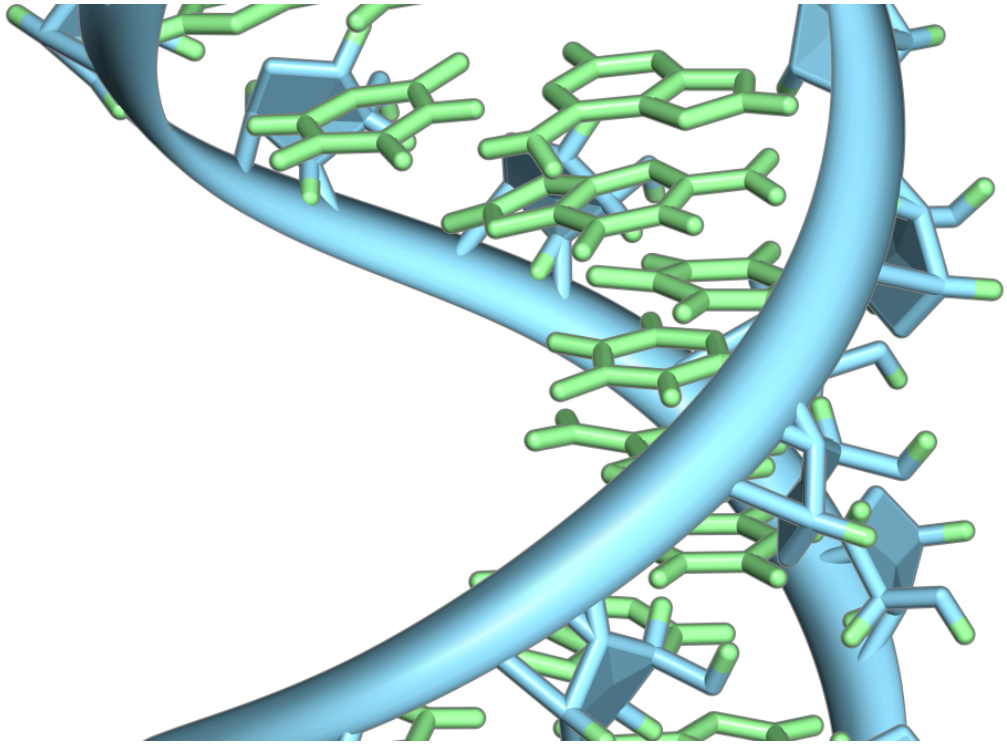
Deep learning cracks the code of messenger RNAs and protein-coding potential

July 23 2018



A hairpin loop from a pre-mRNA. Highlighted are the nucleobases (green) and the ribose-phosphate backbone (blue). Note that this is a single strand of RNA that folds back upon itself. Credit: Vossman/ Wikipedia

Researchers at Oregon State University have used deep learning to decipher which ribonucleic acids have the potential to encode proteins.

The gated recurrent neural network developed in the College of Science and College of Engineering is an important step toward better understanding RNA, one of life's fundamental, essential molecules.

Unlocking the mysteries of RNA means knowing its connections to human health and disease.

Deep learning, a type of machine-learning not based on task-specific algorithms, is a powerful tool for solving the puzzle.

"Deep learning may seem scary to some people, but at the end of the day, it's just crunching numbers," said David Hendrix, the study's lead author. "It's a tool just like calculus or linear algebra, one that we can use to learn biological patterns. The amount of sequencing data we have now is huge, and deep learning is well suited to face the challenges associated with the vast amount of data and to learn new biological rules that characterize the function of these molecules."

RNA is transcribed from DNA, the other nucleic acid—so named because they were first discovered in the cell nuclei of living things—to produce the proteins needed throughout the body.

DNA contains a person's hereditary information, and RNA acts as the messenger that delivers the information's coded instructions to the protein-manufacturing sites within the cells.

Some RNAs are functional molecules transcribed from DNA that aren't translated into proteins. These are known as non-coding RNAs.

Every day, new RNAs are discovered, and gene sequencing technology has advanced to the point that molecular biologists are facing a "torrent" of new transcript annotations to glean information from, Hendrix said.

These vast datasets require new approaches, said the researcher, an assistant professor with joint appointments in biochemistry/biophysics and computer science.

Hendrix and colleagues gave a gated neural network training on both noncoding and messenger RNA sequences, then turned it loose on the data to "learn the defining characteristics of protein-coding transcripts on its own."

It did, with remarkable improvement over existing state-of-the-art methods for predicting protein-coding potential.

"It's really exciting," Hendrix said. "With the competing programs, developers would tell the program what an open reading frame is, what a start codon is, what a stop codon is. We thought it would be better to have a more de novo approach where the neural network can learn independently."

A codon is a sequence of three nucleotides, the basic structural unit of nucleic acids. Codons act like a translator between the nucleotides in DNA and RNA and the 20 amino acids behind protein synthesis.

Compared to other approaches, the model that the OSU team developed, known as mRNN, was better by a statistically significant margin in nearly every available metric.

"It not only found stop codons, it distinguished real stop codons from other trinucleotides that match stop codons and recognized long-range dependencies in the sequences," Hendrix said. "It doesn't wait to see a stop codon—we found it makes its decision long before the stop codon, 200 nucleotides from the start codon. And it learned a subset of codons that were highly predictive of protein-coding potential when observed in a potential open reading frame."

Hendrix and colleagues dubbed these special codons "TICs—translation-indicating codons.

More information: Steven T Hill et al, A deep recurrent neural network discovers complex biological rules to decipher RNA protein-coding potential, *Nucleic Acids Research* (2018). DOI: 10.1093/nar/gky567

Provided by Oregon State University

Citation: Deep learning cracks the code of messenger RNAs and protein-coding potential (2018, July 23) retrieved 6 May 2024 from https://phys.org/news/2018-07-deep-code-messenger-rnas-protein-coding.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.