MIT News
ON CAMPUS AND AROUND THE WORLD

An “oracle” for predicting the evolution of gene regulation

Researchers create a mathematical framework to examine the genome and detect signatures of natural selection, deciphering the evolutionary past and future of non-coding DNA.

Raleigh McElvery | Department of Biology

March 11, 2022

Researchers devised a neural network model capable of predicting how changes to non-coding DNA sequences in yeast affect gene expression and reproductive fitness. The model creates maps, called fitness landscapes, shown here and rendered in the shape of fossilized birds and fish. These higher-order creatures evolved as a result of evolutionary changes to non-coding DNA sequences, like the ones depicted in the fitness landscapes.

Image: Martin Krzywinski

Despite the sheer number of genes that each human cell contains, these so-called “coding” DNA sequences comprise just 1 percent of our entire genome. The remaining 99 percent is made up of “non-coding” DNA — which, unlike coding DNA, does not carry the instructions to build proteins.

One vital function of this non-coding DNA, also called “regulatory” DNA, is to help turn genes on and off, controlling how much (if any) of a protein is made. Over time, as cells replicate their DNA to grow and divide, mutations often crop up in these non-coding regions — sometimes tweaking their function and changing the way they control gene expression. Many of these mutations are trivial, and some are even beneficial. Occasionally, though, they can be associated with increased risk of common diseases, such as Type 2 diabetes, or more life-threatening ones, including cancer.

To better understand the repercussions of such mutations, researchers have been hard at work on mathematical maps that allow them to look at an organism’s genome, predict which genes will be expressed, and determine how that expression will affect the organism’s observable traits. These maps, called fitness landscapes, were conceptualized roughly a century ago to understand how genetic makeup influences one common measure of organismal fitness in particular: reproductive success. Early fitness landscapes were very simple, often focusing on a limited number of mutations. Much richer datasets are now available, but researchers still require additional tools to characterize and visualize such complex data. This ability would not only facilitate a better understanding of how individual genes have evolved over time, but would also help to predict what sequence and expression changes might occur in the future.

In a new study published on March 9 in *Nature*, a team of scientists has developed a framework for studying the fitness landscapes of regulatory DNA. They created a neural network model that, when trained on hundreds of millions of experimental measurements, was capable of predicting how changes to these non-coding sequences in yeast affected gene expression. They also devised a unique way of representing the landscapes in two dimensions, making it easy to understand the past and forecast the future evolution of non-coding sequences in organisms beyond yeast — and even design custom gene expression patterns for gene therapies and industrial applications.

“We now have an 'oracle' that can be queried to ask: What if we tried all possible mutations of this sequence? Or, what new sequence should we design to give us a desired expression?” says Aviv Regev, a professor of biology at MIT (on leave), core member of the Broad Institute of Harvard and MIT (on leave), head of Genentech Research and Early Development, and the study’s senior author. “Scientists can now use the model for their own evolutionary question or scenario, and for other problems like making sequences that control gene expression in desired ways. I am also excited about the possibilities for machine learning researchers interested in interpretability; they can ask their questions in reverse, to better understand the underlying biology."

Prior to this study, many researchers had simply trained their models on known mutations (or slight variations thereof) that exist in nature. However, Regev’s team wanted to go a step further by creating their own unbiased models capable of predicting an organism’s fitness and gene expression based on any possible DNA sequence — even sequences they’d never seen before. This would also enable researchers to use such models to engineer cells for pharmaceutical purposes, including new treatments for cancer and autoimmune disorders.

To accomplish this goal, Eeshit Dhaval Vaishnav, a graduate student at MIT; co-first author Carl de Boer, now an assistant professor at the University of British Columbia; and their colleagues created a neural network model to predict gene expression. They trained it on a dataset generated by inserting millions of totally random non-coding DNA sequences into yeast, and observing how each random sequence affected gene expression. They focused on a particular subset of non-coding DNA sequences called promoters, which serve as binding sites for proteins that can switch nearby genes on or off.

“This work highlights what possibilities open up when we design new kinds of experiments to generate the right data to train models,” Regev says. “In the broader sense, I believe these kinds of approaches will be important for many problems — like understanding genetic variants in regulatory regions that confer disease risk in the human genome, but also for predicting the impact of combinations of mutations, or designing new molecules.”

Regev, Vaishnav, de Boer, and their coauthors went on to test their model’s predictive abilities in a variety of ways, in order to show how it could help demystify the evolutionary past — and possible future — of certain promoters. “Creating an accurate model was certainly an accomplishment, but, to me, it was really just a starting point,” Vaishnav explains.

First, to determine whether their model could help with synthetic biology applications like producing antibiotics, enzymes, and food, the researchers practiced using it to design promoters that could generate desired expression levels for any gene of interest. They then scoured other scientific papers to identify fundamental evolutionary questions, in order to see if their model could help answer them. The team even went so far as to feed their model a real-world population dataset from one existing study, which contained genetic information from yeast strains around the world. In doing so, they were able to delineate thousands of years of past selection pressures that sculpted the genomes of today’s yeast.

But, in order to create a powerful tool that could probe any genome, the researchers knew they’d need to find a way to forecast the evolution of non-coding sequences even without such a comprehensive population dataset. To address this goal, Vaishnav and his colleagues devised a computational technique that allowed them to plot the predictions from their framework onto a two-dimensional graph. This helped them show, in a remarkably simple manner, how any non-coding DNA sequence would affect gene expression and fitness, without needing to conduct any time-consuming experiments at the lab bench.

"One of the unsolved problems in fitness landscapes was that we didn’t have an approach for visualizing them in a way that meaningfully captured the evolutionary properties of sequences,” Vaishnav explains. “I really wanted to find a way to fill that gap, and contribute to the long-standing vision of creating a complete fitness landscape.”

Martin Taylor, a professor of genetics at the University of Edinburgh’s Medical Research Council Human Genetics Unit who was not involved in the research, says the study shows that artificial intelligence can not only predict the effect of regulatory DNA changes, but also reveal the underlying principles that govern millions of years of evolution.

Despite the fact that the model was trained on just a fraction of yeast regulatory DNA in a few growth conditions, he’s impressed that it’s capable of making such useful predictions about the evolution of gene regulation in mammals.

“There are obvious near-term applications, such as the custom design of regulatory DNA for yeast in brewing, baking, and biotechnology,” he explains. “But extensions of this work could also help identify disease mutations in human regulatory DNA that are currently difficult to find and largely overlooked in the clinic. This work suggests there is a bright future for AI models of gene regulation trained on richer, more complex, and more diverse datasets.”

Even before the study was formally published, Vaishnav began receiving queries from other researchers hoping to use the model to devise non-coding DNA sequences for use in gene therapies.

“People have been studying regulatory evolution and fitness landscapes for decades now,” Vaishnav says. “I think our framework will go a long way in answering fundamental, open questions about the evolution and evolvability of gene regulatory DNA — and even help us design biological sequences for exciting new applications.”

Paper: "The evolution, evolvability and engineering of gene regulatory DNA"