Listening in to how proteins talk and learning their language

October 21 2019

Wyss Institute researchers have created an approach to engineer proteins that

uses deep learning, and moves a lot of laborious laboratory experiments to the computer. Credit: Klinsley Stocum

Synthetic biologists have taken evolution of proteins into their own hands by changing some that occur in nature or even by synthesizing them from scratch. Such engineered proteins are used as highly efficacious drugs, components of synthetic gene circuits that sense biological signals, or in the production of high-value chemicals in ways that are more effective and sustainable than petroleum-based methods.

To engineer them, they use two very different approaches. In "directed evolution", they randomly vary the linear sequence of amino acid building blocks encoding a natural protein and screen for variants with the desired activity; or they use "rational design" to model proteins based on their actual 3-D structures to identify amino acids that likely will impact protein function. However, directed evolution can only cover a small part of the enormous space of possible protein sequences, while rational design approaches are limited by the relative scarcity of painstakingly resolved 3-D protein structures.

Now, a research team led by George Church, Ph.D. at Harvard's Wyss Institute for Biologically Inspired Engineering and Harvard Medical School (HMS) has created a third approach to engineering proteins that uses deep learning to distill the fundamental features of proteins directly from their amino acid sequence without the need for additional information. The approach robustly predicts the functions of both natural and de novo designed proteins, and moves a lot of laborious laboratory experiments to the computer, achieving up to two orders of magnitude cost reduction compared to existing approaches. The study is published in *Nature Methods*.

Church is a Founding Core Faculty member of the Wyss Institute and Lead of its Synthetic Biology platform. He also is the Robert Winthrop Professor of Genetics at Harvard Medical School and Professor of Health Sciences and Technology at Harvard University and the Massachusetts Institute of Technology (MIT).

"Instead of extensively characterizing proteins to understand their design principles, we used a neural network to learn those rules in an unbiased way, by systematically looking for patterns in a vast trove of raw protein sequences in public databases," said Surojit Biswas, one of the three co-first authors on the study who is a graduate student in Church's group. "The neural network learned a lot of the rules that we as humans have previously come to know through many painstaking studies, and beyond that, it also discovered new features in proteins."

The neural network approach, which the researchers named "unified representation" (UniRep), can be likened to learning a language where the learner builds a semantic understanding of how complex sentences are constructed from strings of letters and words. In protein language, UniRep was trained to predict the next amino acid in a protein sequence starting from its first one by exploring all the possibilities in protein sequences contained in public databases. Importantly, while proceeding through the remainder of the protein, one amino acid at a time, UniRep makes and draws on an internal "summary" of the sequence it has seen so far in the protein, which the team calls its "hidden state", to take into account its individual sequence and structural features. Feeding that information, and results from many other proteins, back into its algorithm, UniRep gradually revises the way it constructs hidden states, which improves its predictive capabilities over time. In the language analogy, the learner will be able to predict the next word of a sentence he is reading with increasing likelihood, based on a constantly improving understanding of syntax and choice of words.

"We trained UniRep on about 24 million protein sequences for roughly three weeks to enable it to predict sequences and their relationship to features like protein stability, secondary structure, and accessibility of internal sequences to surrounding solvents within proteins it had never seen before," said Grigory Khimulya, who was a student at Harvard College and is also a co-first author along with Biswas and Ethan Alley. "UniRep accurately described these features in proteins from very different protein families whose structures had been well-characterized in previous studies, even in synthetic proteins that don't have a counterpart in nature."

The team took UniRep a step further and used it as a tool to predict how single amino acid substitutions impact the function of proteins. The neural network robustly quantified the effects of single amino acid mutations in eight different proteins with diverse biological functions including enzyme catalysis, DNA binding, molecular sensing. In addition, using the Aequorea victoria green fluorescent protein (GFP) as a model, they tasked UniRep to analyze 64,800 variants of the protein, each carrying 1-12 mutations, which demonstrated that it could accurately anticipate how the distribution and relative burden of mutations changed the protein's brightness.

"Compared to other strategies, our data-driven approach reaches state-of-the-art or superior performance in predicting multiple properties of proteins at costs much lower than other methods," said Church. "This makes it a truly empowering tool for protein engineers in many areas."

"This new deep learning-based computational approach to protein engineering has the potential to accelerate the design of synthetic proteins with functions tailored for any desired application, whether it before for therapeutics, diagnostics, biomanufacturing, biocatalysis, or any other application. It literally can change the way we carry out molecular design in the future," said Wyss Founding Director Donald

Ingber, M.D., Ph.D., who is also the Judah Folkman Professor of Vascular Biology at HMS and the Vascular Biology Program at Boston Children's Hospital, as well as Professor of Bioengineering at Harvard's John A. Paulson School of Engineering and Applied Sciences.

More information: *Nature Methods* (2019). DOI: 10.1038/s41592-019-0598-1