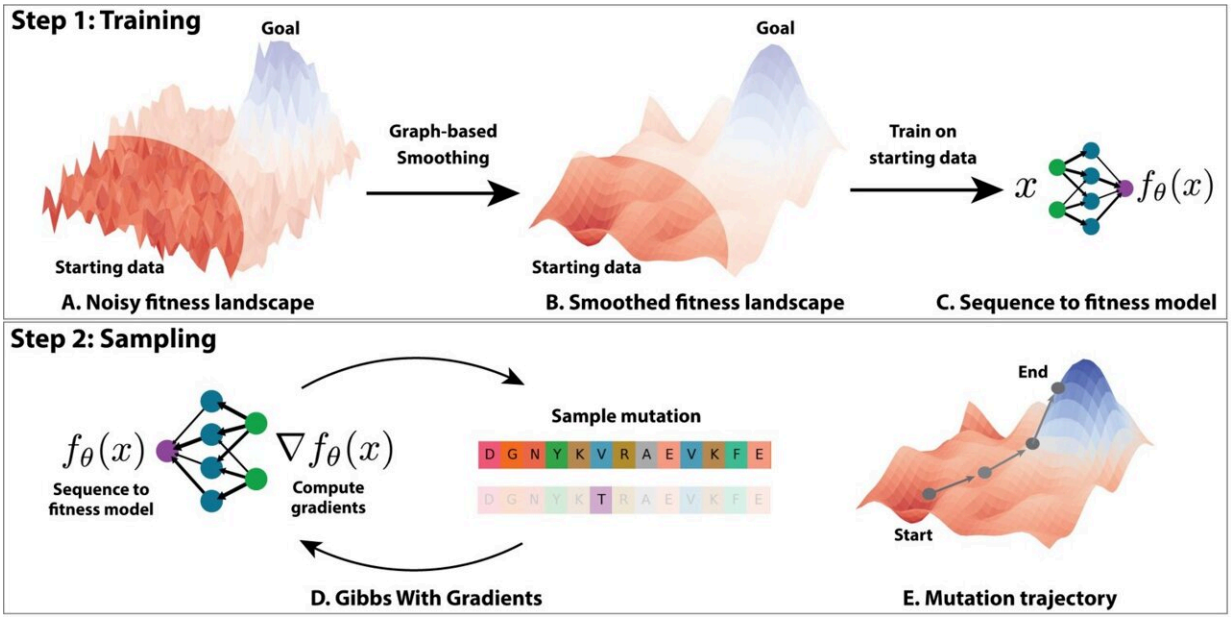
A new computational technique could make it easier to engineer useful proteins

April 3 2024, by Anne Trafton



Overview. (A) Protein optimization is challenging due to a noisy fitness landscape where the starting dataset (unblurred) is a fraction of the landscape with the highest fitness sequences hidden (blurred). (B) We develop Graph-based Smoothing (GS) to estimate a smoothed fitness landscape from the starting data. (C) A model is trained on the smoothed fitness landscape to infer the rest of the landscape. (D) Gradients from the model are used in Gibbs With Gradients (GWG) where on each step a new mutation is proposed. (E) The goal of sampling is for each trajectory to gradually head towards higher fitness. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2307.00494

To engineer proteins with useful functions, researchers usually begin with a natural protein that has a desirable function, such as emitting fluorescent light, and put it through many rounds of random mutation that eventually generate an optimized version of the protein.

This process has yielded optimized versions of many important proteins, including green fluorescent protein (GFP). However, for other proteins, it has proven difficult to generate an optimized version. MIT researchers have now developed a computational approach that makes it easier to predict mutations that will lead to better proteins, based on a relatively small amount of data.

Using this model, the researchers generated proteins with mutations that were predicted to lead to improved versions of GFP and a protein from adeno-associated virus (AAV), which is used to deliver DNA for gene therapy. They hope it could also be used to develop additional tools for neuroscience research and medical applications.

"Protein design is a hard problem because the mapping from DNA sequence to protein structure and function is really complex. There might be a great protein 10 changes away in the sequence, but each intermediate change might correspond to a totally nonfunctional protein.

"It's like trying to find your way to the river basin in a mountain range, when there are craggy peaks along the way that block your view. The current work tries to make the riverbed easier to find," says Ila Fiete, a professor of brain and cognitive sciences at MIT, a member of MIT's McGovern Institute for Brain Research, director of the K. Lisa Yang Integrative Computational Neuroscience Center, and one of the senior authors of the study.

Regina Barzilay, the School of Engineering Distinguished Professor for AI and Health at MIT, and Tommi Jaakkola, the Thomas Siebel

Professor of Electrical Engineering and Computer Science at MIT, are also senior authors of an open-access paper on the work, which will be presented at the International Conference on Learning Representations (ICLR 2024) in May. It is available on the *arXiv* preprint server.

MIT graduate students Andrew Kirjner and Jason Yim are the lead authors of the study. Other authors include Shahar Bracha, an MIT postdoc, and Raman Samusevich, a graduate student at Czech Technical University.

Optimizing proteins

Many naturally occurring proteins have functions that could make them useful for research or medical applications, but they need a little extra engineering to optimize them. In this study, the researchers were originally interested in developing proteins that could be used in living cells as voltage indicators.

These proteins, produced by some bacteria and algae, emit fluorescent light when an electric potential is detected. If engineered for use in mammalian cells, such proteins could allow researchers to measure neuron activity without using electrodes.

While decades of research have gone into engineering these proteins to produce a stronger fluorescent signal, on a faster timescale, they haven't become effective enough for widespread use. Bracha, who works in Edward Boyden's lab at the McGovern Institute, reached out to Fiete's lab to see if they could work together on a computational approach that might help speed up the process of optimizing the proteins.

"This work exemplifies the human serendipity that characterizes so much science discovery," Fiete says. "It grew out of the Yang Tan Collective retreat, a scientific meeting of researchers from multiple

centers at MIT with distinct missions unified by the shared support of K. Lisa Yang. We learned that some of our interests and tools in modeling how brains learn and optimize could be applied in the totally different domain of protein design, as being practiced in the Boyden lab."

For any given protein that researchers might want to optimize, there is a nearly infinite number of possible sequences that could generated by swapping in different amino acids at each point within the sequence. With so many possible variants, it is impossible to test all of them experimentally, so researchers have turned to computational modeling to try to predict which ones will work best.

In this study, the researchers set out to overcome those challenges, using data from GFP to develop and test a computational model that could predict better versions of the protein.

They began by training a type of model known as a convolutional neural network (CNN) on experimental data consisting of GFP sequences and their brightness—the feature that they wanted to optimize.

The model was able to create a "fitness landscape"—a three-dimensional map that depicts the fitness of a given protein and how much it differs from the original sequence—based on a relatively small amount of experimental data (from about 1,000 variants of GFP).

These landscapes contain peaks that represent fitter proteins and valleys that represent less fit proteins. Predicting the path that a protein needs to follow to reach the peaks of fitness can be difficult, because often a protein will need to undergo a mutation that makes it less fit before it reaches a nearby peak of higher fitness. To overcome this problem, the researchers used an existing computational technique to "smooth" the fitness landscape.

Once these small bumps in the landscape were smoothed, the researchers retrained the CNN model and found that it was able to reach greater fitness peaks more easily. The model was able to predict optimized GFP sequences that had as many as seven different amino acids from the protein sequence they started with, and the best of these proteins were estimated to be about 2.5 times fitter than the original.

"Once we have this landscape that represents what the model thinks is nearby, we smooth it out and then we retrain the model on the smoother version of the landscape," Kirjner says. "Now there is a smooth path from your starting point to the top, which the model is now able to reach by iteratively making small improvements. The same is often impossible for unsmoothed landscapes."

Proof of concept

The researchers also showed that this approach worked well in identifying new sequences for the viral capsid of adeno-associated virus (AAV), a viral vector that is commonly used to deliver DNA. In that case, they optimized the capsid for its ability to package a DNA payload.

"We used GFP and AAV as a proof of concept to show that this is a method that works on data sets that are very well-characterized, and because of that, it should be applicable to other protein engineering problems," Bracha says.

The researchers now plan to use this computational technique on data that Bracha has been generating on voltage indicator proteins.

"Dozens of labs having been working on that for two decades, and still there isn't anything better," she says. "The hope is that now with generation of a smaller data set, we could train a model in silico and make predictions that could be better than the past two decades of

manual testing."

More information: Andrew Kirjner et al, Improving Protein Optimization with Smoothed Fitness Landscapes, *arXiv* (2023). DOI: 10.48550/arxiv.2307.00494

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology