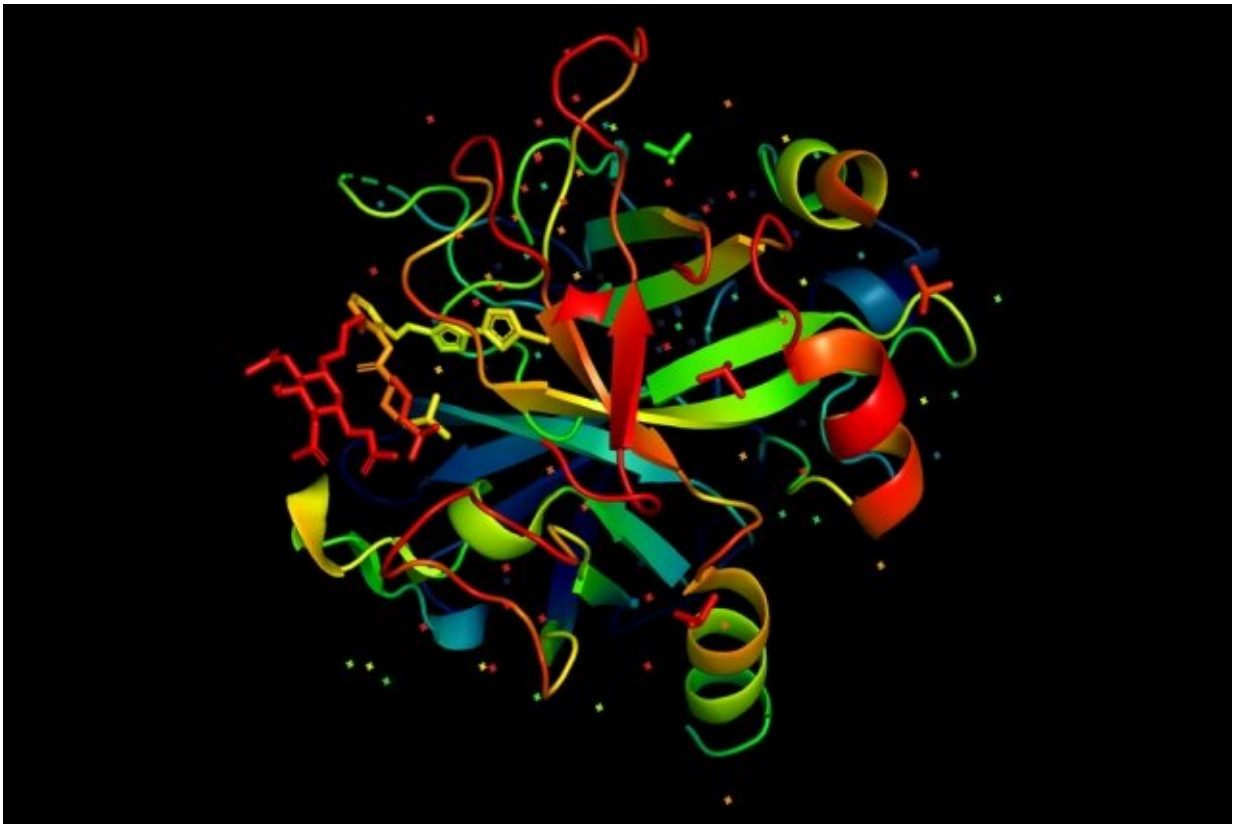
Model learns how individual amino acids determine protein function

March 25 2019, by Rob Matheson



A new model developed by MIT researchers creates richer, more easily computable representations of how individual amino acids determine a protein's function, which could be used for designing and testing new proteins. Credit: Massachusetts Institute of Technology

A machine-learning model from MIT researchers computationally

breaks down how segments of amino acid chains determine a protein's function, which could help researchers design and test new proteins for drug development or biological research.

Proteins are linear chains of amino acids, connected by peptide bonds, that fold into exceedingly complex three-dimensional structures, depending on the sequence and physical interactions within the chain. That structure, in turn, determines the protein's biological function. Knowing a protein's 3-D structure, therefore, is valuable for, say, predicting how proteins may respond to certain drugs.

However, despite decades of research and the development of multiple imaging techniques, we know only a very small fraction of possible protein structures—tens of thousands out of millions. Researchers are beginning to use machine-learning models to predict protein structures based on their amino acid sequences, which could enable the discovery of new protein structures. But this is challenging, as diverse amino acid sequences can form very similar structures. And there aren't many structures on which to train the models.

In a paper being presented at the International Conference on Learning Representations in May, the MIT researchers develop a method for "learning" easily computable representations of each amino acid position in a protein sequence, initially using 3-D protein structure as a training guide. Researchers can then use those representations as inputs that help machine-learning models predict the functions of individual amino acid segments—without ever again needing any data on the protein's structure.

In the future, the model could be used for improved protein engineering, by giving researchers a chance to better zero in on and modify specific amino acid segments. The model might even steer researchers away from protein structure prediction altogether.

"I want to marginalize structure," says first author Tristan Bepler, a graduate student in the Computation and Biology group in the Computer Science and Artificial Intelligence Laboratory (CSAIL). "We want to know what proteins do, and knowing structure is important for that. But can we predict the function of a protein given only its amino acid sequence? The motivation is to move away from specifically predicting structures, and move toward [finding] how amino acid sequences relate to function."

Joining Bepler is co-author Bonnie Berger, the Simons Professor of Mathematics at MIT with a joint faculty position in the Department of Electrical Engineering and Computer Science, and head of the Computation and Biology group.

Learning from structure

Rather than predicting structure directly—as traditional models attempt—the researchers encoded predicted protein structural information directly into representations. To do so, they use known structural similarities of proteins to supervise their model, as the model learns the functions of specific amino acids.

They trained their model on about 22,000 proteins from the Structural Classification of Proteins (SCOP) database, which contains thousands of proteins organized into classes by similarities of structures and amino acid sequences. For each pair of proteins, they calculated a real similarity score, meaning how close they are in structure, based on their SCOP class.

The researchers then fed their model random pairs of protein structures and their amino acid sequences, which were converted into numerical representations called embeddings by an encoder. In natural language processing, embeddings are essentially tables of several hundred

numbers combined in a way that corresponds to a letter or word in a sentence. The more similar two embeddings are, the more likely the letters or words will appear together in a sentence.

In the researchers' work, each embedding in the pair contains information about how similar each amino acid sequence is to the other. The model aligns the two embeddings and calculates a similarity score to then predict how similar their 3-D structures will be. Then, the model compares its predicted similarity score with the real SCOP similarity score for their structure, and sends a feedback signal to the encoder.

Simultaneously, the model predicts a "contact map" for each embedding, which basically says how far away each amino acid is from all the others in the protein's predicted 3-D structure—essentially, do they make contact or not? The model also compares its predicted contact map with the known contact map from SCOP, and sends a feedback signal to the encoder. This helps the model better learn where exactly amino acids fall in a protein's structure, which further updates each amino acid's function.

Basically, the researchers train their model by asking it to predict if paired sequence embeddings will or won't share a similar SCOP protein structure. If the model's predicted score is close to the real score, it knows it's on the right track; if not, it adjusts.

Protein design

In the end, for one inputted amino acid chain, the model will produce one numerical representation, or embedding, for each amino acid position in a 3-D structure. Machine-learning models can then use those sequence embeddings to accurately predict each amino acid's function based on its predicted 3-D structural "context"—its position and contact with other amino acids.

For instance, the researchers used the model to predict which segments, if any, pass through the cell membrane. Given only an amino acid sequence, the researchers' model predicted all transmembrane and non-transmembrane segments more accurately than state-of-the-art models.

"The work by Bepler and Berger is a significant advance in representing the local structural properties of a protein sequence," says Serafim Batzoglou, a professor of computer science at Stanford University. "The representation is learned using state-of-the-art deep learning methods, which have made major strides in protein structure prediction in systems such as RaptorX and AlphaFold. This work has ultimate application in human health and pharmacogenomics, as it facilitates detection of deleterious mutations that disrupt protein structures."

Next, the researchers aim to apply the model to more prediction tasks, such as figuring out which sequence segments bind to small molecules, which is critical for drug development. They're also working on using the model for protein design. Using their sequence embeddings, they can predict, say, at what color wavelengths a protein will fluoresce.

"Our model allows us to transfer information from known protein structures to sequences with unknown structure. Using our embeddings as features, we can better predict function and enable more efficient data-driven protein design," Bepler says. "At a high level, that type of protein engineering is the goal."

Berger adds: "Our machine learning models thus enable us to learn the 'language' of protein folding—one of the original 'Holy Grail' problems—from a relatively small number of known structures."

More information: Learning Protein Sequence Embeddings Using Information From Structure. openreview.net/pdf?id=SygLehCqtm