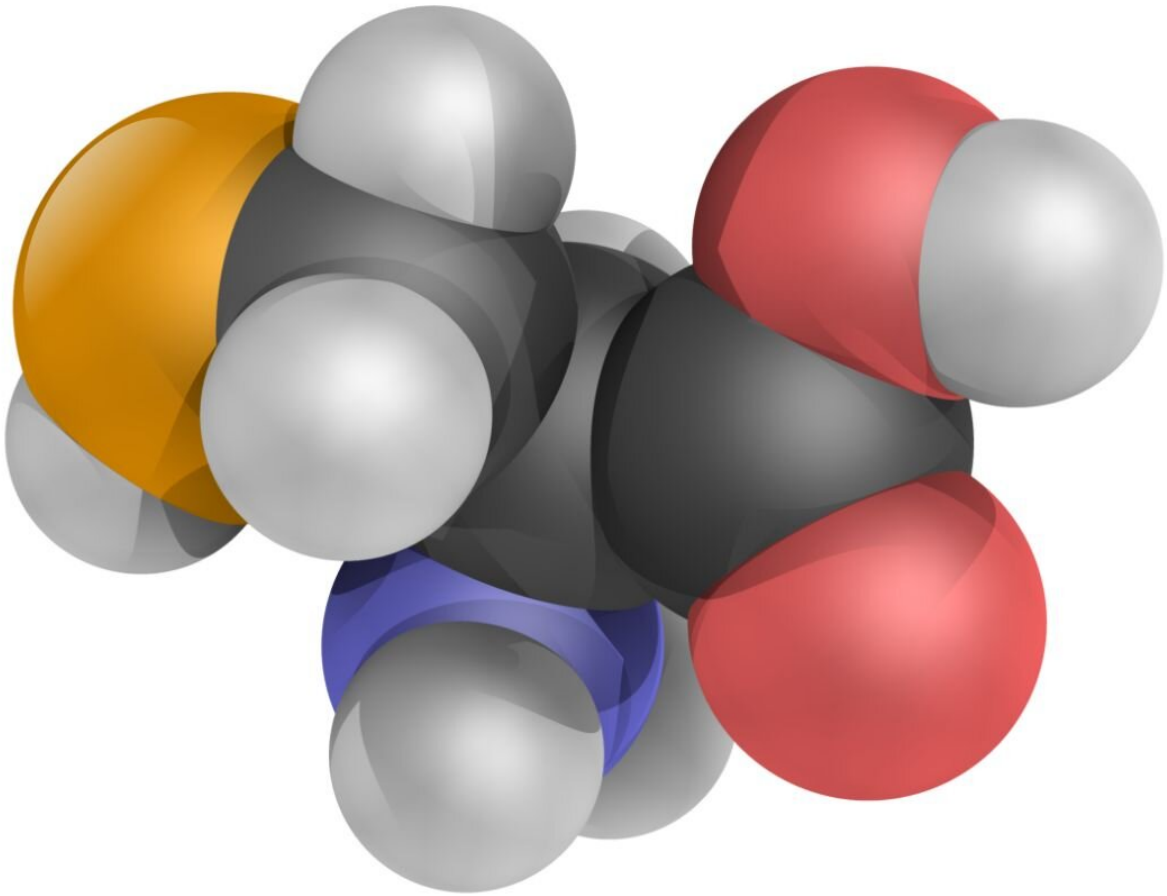
New deep-learning approach predicts protein structure from amino acid sequence

April 17 2019



The amino acid selenocysteine, 3D-balls model. Credit: YassineMrabet/CC BY 3.0/Wikipedia

Nearly every fundamental biological process necessary for life is carried out by proteins. They create and maintain the shapes of cells and tissues; constitute the enzymes that catalyze life-sustaining chemical reactions; act as molecular factories, transporters and motors; serve as both signal and receiver for cellular communications; and much more.

Composed of long chains of amino acids, proteins perform these myriad tasks by folding themselves into precise 3-D structures that govern how they interact with other molecules. Because a protein's shape determines its function and the extent of its dysfunction in disease, efforts to illuminate protein structures are central to all of molecular biology—and in particular, therapeutic science and the development of lifesaving and life-altering medicines.

In recent years, computational methods have made significant strides in predicting how proteins fold based on knowledge of their amino acid sequence. If fully realized, these methods have the potential to transform virtually all facets of biomedical research. Current approaches, however, are limited in the scale and scope of the proteins that can be determined.

Now, a Harvard Medical School scientist has used a form of artificial intelligence known as deep learning to predict the 3-D structure of effectively any protein based on its amino acid sequence.

Reporting online in *Cell Systems* on April 17, systems biologist Mohammed AlQuraishi details a new approach for computationally determining protein structure—achieving accuracy comparable to current state-of-the-art methods but at speeds upward of a million times faster.

"Protein folding has been one of the most important problems for biochemists over the last half century, and this approach represents a fundamentally new way of tackling that challenge," said AlQuraishi,

instructor in systems biology in the Blavatnik Institute at HMS and a fellow in the Laboratory of Systems Pharmacology. "We now have a whole new vista from which to explore protein folding, and I think we've just begun to scratch the surface."

Easy to state

While highly successful, processes that use physical tools to identify protein structures are expensive and time consuming, even with modern techniques such as cryo-electron microscopy. As such, the vast majority of protein structures—and the effects of disease-causing mutations on these structures—are still largely unknown.

Computational methods that calculate how proteins fold have the potential to dramatically reduce the cost and time needed to determine structure. But the problem is difficult and remains unsolved after nearly four decades of intense effort.

Proteins are built from a library of 20 different amino acids. These act like letters in an alphabet, combining into words, sentences and paragraphs to produce an astronomical number of possible texts. Unlike alphabet letters, however, amino acids are physical objects positioned in 3-D space. Often, sections of a protein will be in close physical proximity but be separated by large distances in terms of sequence, as its amino acid chains form loops, spirals, sheets and twists.

"What's compelling about the problem is that it's fairly easy to state: take a sequence and figure out the shape," AlQuraishi said. "A protein starts off as an unstructured string that has to take on a 3-D shape, and the possible sets of shapes that a string can fold into is huge. Many proteins are thousands of amino acids long, and the complexity quickly exceeds the capacity of human intuition or even the most powerful computers."

Hard to solve

To address this challenge, scientists leverage the fact that amino acids interact with each other based on the laws of physics, seeking out energetically favorable states like a ball rolling downhill to settle at the bottom of a valley.

The most advanced algorithms calculate protein structure by running on supercomputers—or crowd-sourced computing power in the case of projects such as Rosetta@Home and Folding@Home—to simulate the complex physics of amino acid interactions through brute force. To reduce the massive computational requirements, these projects rely on mapping new sequences onto predefined templates, which are protein structures previously determined through experiment.

Other projects such as Google's AlphaFold have generated tremendous recent excitement by using advances in artificial intelligence to predict a protein's structure. To do so, these approaches parse enormous volumes of genomic data, which contain the blueprint for protein sequences. They look for sequences across many species that have likely evolved together, using such sequences as indicators of close physical proximity to guide structure assembly.

These AI approaches, however, do not predict structures based solely on a protein's amino acid sequence. Thus, they have limited efficacy for proteins for which there is no prior knowledge, evolutionary unique proteins or novel proteins designed by humans.

Training deeply

To develop a new approach, AlQuraishi applied so-called end-to-end differentiable deep learning. This branch of artificial intelligence has

dramatically reduced the computational power and time needed to solve problems such as image and speech recognition, enabling applications such as Apple's Siri and Google Translate.

In essence, differentiable learning involves a single, enormous mathematical function—a much more sophisticated version of a high school calculus equation—arranged as a neural network, with each component of the network feeding information forward and backward.

This function can tune and adjust itself, over and over at unimaginable levels of complexity, in order to "learn" precisely how a protein sequence mathematically relates to its structure.

AlQuraishi developed a deep-learning model, termed a recurrent geometric network, which focuses on key characteristics of protein folding. But before it can make new predictions, it must be trained using previously determined sequences and structures.

For each amino acid, the model predicts the most likely angle of the chemical bonds that connect the amino acid with its neighbors. It also predicts the angle of rotation around these bonds, which affects how any local section of a protein is geometrically related to the entire structure.

This is done repeatedly, with each calculation informed and refined by the relative positions of every other amino acid. Once the entire structure is completed, the model checks the accuracy of its prediction by comparing it against the "ground truth" structure of the protein.

This entire process is repeated for thousands of known proteins, with the model learning and improving its accuracy with every iteration.

New vista

Once his model was trained, AlQuraishi tested its predictive power. He compared its performance against other methods from several recent years of the Critical Assessment of Protein Structure Prediction— an annual experiment that tests computational methods for their ability to make predictions using protein structures that have been determined but not publicly released.

He found that the new model outperformed all other methods at predicting protein structures for which there are no preexisting templates, including methods that use co-evolutionary data. It also outperformed all but the best methods when preexisting templates were available to make predictions.

While these gains in accuracy are relatively small, AlQuraishi notes that any improvements at the top end of these tests are difficult to achieve. And because this method represents an entirely new approach to protein folding, it can complement existing methods, both computational and physical, to determine a much wider range of structures than previously possible.

Strikingly, the new model performs its predictions at around six to seven orders of magnitude faster than existing computational methods. Training the model can take months, but once trained it can make predictions in milliseconds compared to the hours to days it takes using other approaches. This dramatic improvement is partly due to the single mathematical function on which it is based, requiring only a few thousand lines of computer code to run instead of millions.

The rapid speed of this model's predictions enables new applications that were slow or difficult to achieve before, AlQuraishi said, such as predicting how proteins change their shape as they interact with other molecules.

"Deep-learning approaches, not just mine, will continue to grow in their predictive power and in popularity, because they represent a minimal, simple paradigm that can integrate new ideas more easily than current complex models," he added.

The new model is not immediately ready for use in, say, drug discovery or design, AlQuraishi said, because its accuracy currently falls somewhere around 6 angstroms—still some distance away from the 1 to 2 angstroms needed to resolve the full atomic structure of a protein. But there are many opportunities to optimize the approach, he said, including further integrating rules drawn from chemistry and physics.

"Accurately and efficiently predicting protein folding has been a holy grail for the field, and it is my hope and expectation that this approach, combined with all the other remarkable methods that have been developed, will be able to do so in the near future," AlQuraishi said. "We might solve this soon, and I think no one would have said that five years ago. It's very exciting and also kind of shocking at the same time."

To help others participate in method development, AlQuraishi has made his software and results freely available via the GitHub software sharing platform.

"One remarkable feature of AlQuraishi's work is that a single research fellow, embedded in the rich research ecosystem of Harvard Medical School and the Boston biomedical community, can compete with companies such as Google in one of the hottest areas of computer science," said Peter Sorger, HMS Otto Krayer Professor of Systems Pharmacology in the Blavatnik Institute at HMS, director of the Laboratory of Systems Pharmacology at HMS and AlQuraishi's academic mentor.

"It is unwise to underestimate the disruptive impact of brilliant fellows

like AlQuraishi working with open-source software in the public domain," Sorger said.

More information: *Cell Systems* (2019). DOI: 10.1016/j.cels.2019.03.006