The AlphaGo and AlphaFold Stir

CHAOK SEOK
PROFESSOR, DEPARTMENT OF CHEMISTRY, SEOUL NATIONAL UNIVERSITY

While the rest of the world was all abuzz over the clash between AlphaGo and Lee Sedol, I was actually quite indifferent. Part of it was my unfamiliarity with go, but mostly it was because I, as a scientist, was faced with a problem far more difficult than that game. I'm delighted to finally have this opportunity to admit that. In go, all you have to do is outdo your opponent; is it really so amazing in today's world that a computer can beat a human being? But when AlphaFold came about, I was quite shocked. To be sure, AlphaFold had by no means solved the problem I was trying to address. A closer look shows nothing truly staggering or new. Nevertheless, it is shocking. The reason may be that it upended the very framework of the natural sciences' traditional research approach and mindset that have been sustained for centuries.

Imagine, if you will, a problem incomparably more complex than a game of go. Go is more complex than chess, but the difference between go and this problem is far greater. The problem in question is not a game; it has no known rules like go. After you've solved something, you have no way of knowing how correct your answer was. Only when another scientist finds the right answer through experiment can you compare it with the answer you found. It is the most difficult kind of problem: one where you have to find the absolute correct answer in nature rather than simply outperforming someone else, and all without knowing the rules of the game.

So I will ask a question: if it's enough to find the answers through experiment, why do we insist on making calculations? To answer in simple terms, there's a tremendous ripple effect that arises when we are able to predict things by calculating. Google's DeepMind made an excellent decision in choosing this problem as its next target after go. It's a problem that has vexed scientists for over half a century, and it is closely tied to bio and pharmaceutical research. It's a complex problem of computational science, with a lot of related data that can be put to use. And, like a sport, it has regularly staged international competitions.

So let's talk in a bit more detail about the problem. The problem in question has to do with proteins. Proteins are tiny, nano-scale molecules that make up all life forms. The proteins that form organisms exist in large numbers and many different types. As they meet and react with each other, they allow various phenomena to occur, including embryogenesis, immunity, and digestion. It would not be overstating things to say that the different organisms that exist are determined by the proteins they form. Human beings and monkeys produce different proteins, although they are closer to each other than humans are to mice. People look different because their proteins slightly differ; they are more or less susceptible to certain diseases and react in different ways to the same medications. What proteins we make is decided in turn by our genes. That is why a family member's proteins are more similar to ours than a stranger's.

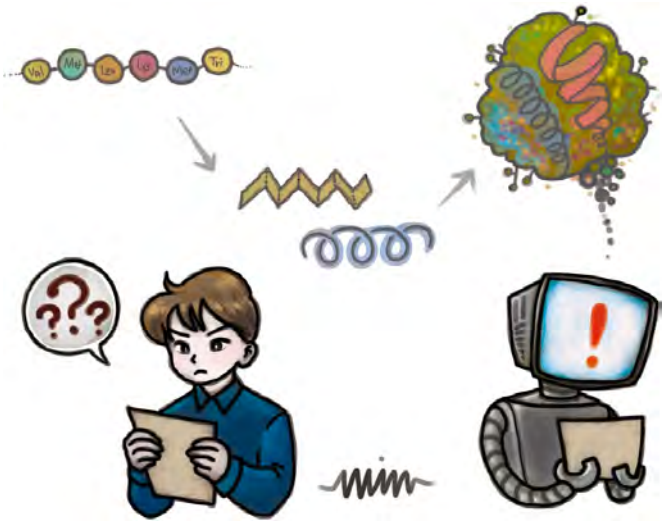
A single protein is typically made up of hundreds of amino acids. We can understand it as a long, chain-like molecule formed by stringing together amino acids like beads. There are 20 different amino acids in all, and a great many kinds of chains can be formed by stringing together hundreds of those 20 different beads in different combinations. The type of protein is determined by the order in which the amino acids are positioned to form it—the amino acid sequence. The amino acid sequences produced by humans and mice are sometimes quite similar and other times completely different. The proteins formed by different people are more or less the same, but

with very slightly differences in sequence. Those differences are what account for biological difference.

The problem we are talking about is how to predict the three-dimensional spatial structure of a protein from the one-dimensional information of its amino acid sequence. Once an amino acid sequence has been decided, since the atoms that form the individual amino acids are also set, the types of atoms in the protein and the sequence in which they are linked together are determined. Once the atoms forming a protein chain have been determined, the chain will bend at different angles in different locations, clumping together to form a kind of ball. The proteins in an organism usually clump into a single defined structure; the problem of protein structure prediction is one of predicting the position of each atom within this clumped structure. Imagine just how many times these thousands of atoms will clump together in continuous three-dimensional space to form structures. Certainly it is far more complex than go, which has only black and white pieces and empty spaces on a two-dimensional, 19×19 board.

Proteins also interact with other molecules in the body, and their three-dimensional structures offer an excellent clue toward understanding those interactions. The molecular structures of proteins provide us with useful information in gaining the most detailed understanding of biological phenomena and using it to design therapeutic agents. That's why no fewer than three Nobel Prizes to date have been awarded for methods of experimentally determining protein structures. A number of other Nobel Prizes have been given for using these methods to discover detailed information about specific phenomena. Could we yet see the day when the methods of protein structure prediction receive recognition on par with the three "big" experimental approaches?

Prediction results must be tested experimentally. John Moult, a professor at the University of Maryland, took a stab in 1994. He obtained amino acid sequences for various proteins whose structures would soon be determined experimentally by structural biologists. He then provided these sequences to contest participants and asked them to submit their structure predictions within a certain deadline for comparison with the experimentally determined structures. Known as CASP (Critical Assessment of techniques for Protein Structure Prediction), the event has been held every two years since then, with support from the US National Institutes of Health. Recently,



more than one hundred protein sequences each year have been presented as problems, with around 200 prediction teams participating. Two to three new sequences are presented each day during the summer of even-numbered years, when CASP is held. After the contest season is over, the prediction results are assessed by a third-party evaluation committee. Obviously, the names of the participating teams are kept hidden, and the predictions are evaluated strictly by their participant number. At the CASP conference in December, the ranking of participating teams is revealed; top-ranking teams are afforded the honor of giving an invited presentation, and everyone gathers to discuss the direction of future development.

I was both a participant and an evaluation committee member for the 13th CASP event in 2018. While the main focus of CASP is on predicting three-dimensional structure from amino acid sequences, it also includes a number of other areas. I was an evaluation committee member in one of those, the area of structural evaluation. (It is considered unusual for a CASP participant to be tasked with assessment for another area.) CASP organizing committee and evaluation committee members gathered in Switzerland last October for an evaluation meeting. We were there to examine the assessment results in different areas and select the top-placing groups to present at the conference that December. It was announced there that one participating team had placed first in the area of three-dimensional structural prediction by a large margin compared to previous years. During the meeting, the team was identified only by its participant number; we tried to predict which team it might be. Most people agreed it had to be Baker, while a few of them predicted it might be Google. David Baker

of the University of Washington is an outstanding scholar who conducts pioneering research in protein structure prediction as well as protein design. Google was also reported as having recently begun to research the protein structure prediction topic. People were quite stunned and excited when Google was revealed as the winner—it had swept into first place in so short a time.

It isn't really possible to talk in detail here about the methodological aspects of structure prediction. One thing I would like to stress, however, is that AlphaFold, the system created by Google's DeepMind, did not invent some new method that had not existed before. AlphaFold's focus was on one of the areas that has been making the greatest contributions recently to development of the field, namely the prediction of amino acid distances from sequence information. That approach was based on the groundwork laid by different methods that have been established in the field over many years. What AlphaFold did was to apply deep learning to this—which was not the first time this approach had been attempted. Similar methods had been developed by Professor David T. Jones of University College London, who also served as an adviser to DeepMind. Perhaps AlphaFold boasted some exceptional deep learning techniques, because it was more accurate in its predictions of amino acid distances, and it used the predicted distances between amino acids to develop a good overall structure. All of the ideas and methods that went into this had been things that previously existed.

As someone who studied chemistry and physics, I have romantic notions of predicting the three-dimensional structures of proteins from calculations of the physical and chemical interactions between their atoms. It may be similar to the ideas that gripped many theoretical scientists shortly after the first protein structure was elucidated in 1958. The reality, however, is that a more accurate means of prediction is to apply the experimentally determined structures of similar proteins as a template if they can be identified; the next most accurate approach is to do what AlphaFold did and extract distance information from the sequence information. In other words, information still plays a more crucial role than principles in the area of protein structure prediction. With current technology, it is possible to accurately predict structure to a certain extent (around 70%) about 60% of the time, so we can't say the protein structure prediction problem has yet been solved. AlphaFold may be seen as having contributed to lifting that number to around 60% from 50% before. It's the result of making very appropriate use of DeepMind's outstanding deep learning technology—something that was going to be put to use anyway—in the field of protein structure prediction.

While this may not appear like all that much on closer examination, AlphaFold nevertheless came as a shock to me. Back in October, when I was looking at the evaluation results with an unprejudiced eye—not knowing who team #043 were or what methods they had used—it was clear that they marked a big step forward from before. One of the big goals of the natural sciences, the field to which I belong, is an understanding of the physical and chemical principles that govern natural phenomena; with one general principle, it becomes possible for us to account for many different phenomena from a single perspective. But many of the problems in reality are too complex for us to approach solely through the natural science principles we have determined using mathematics as a tool. We now happen to live in an era where machine learning approaches based on data learning have developed enough to be applied to various problems that were not solved by other means. The research methods of the natural sciences are undergoing a change due to the quantitative growth of data obtained over their long history. The opening salvo for this was fired with AlphaFold's success at CASP. With AlphaGo coming a mere two years ahead of AlphaFold, I was quite foolish to think so little of it at the time. Data learning is completely changing the face of the natural sciences and their long-established culture.

Chaok Seok is a professor at the Department of Chemistry of Seoul National University. She received her BS in chemistry from Seoul National University and her Ph.D. from University of Chicago. Her research interests include developing methods for protein structure prediction and for predicting interactions of proteins with other molecules.