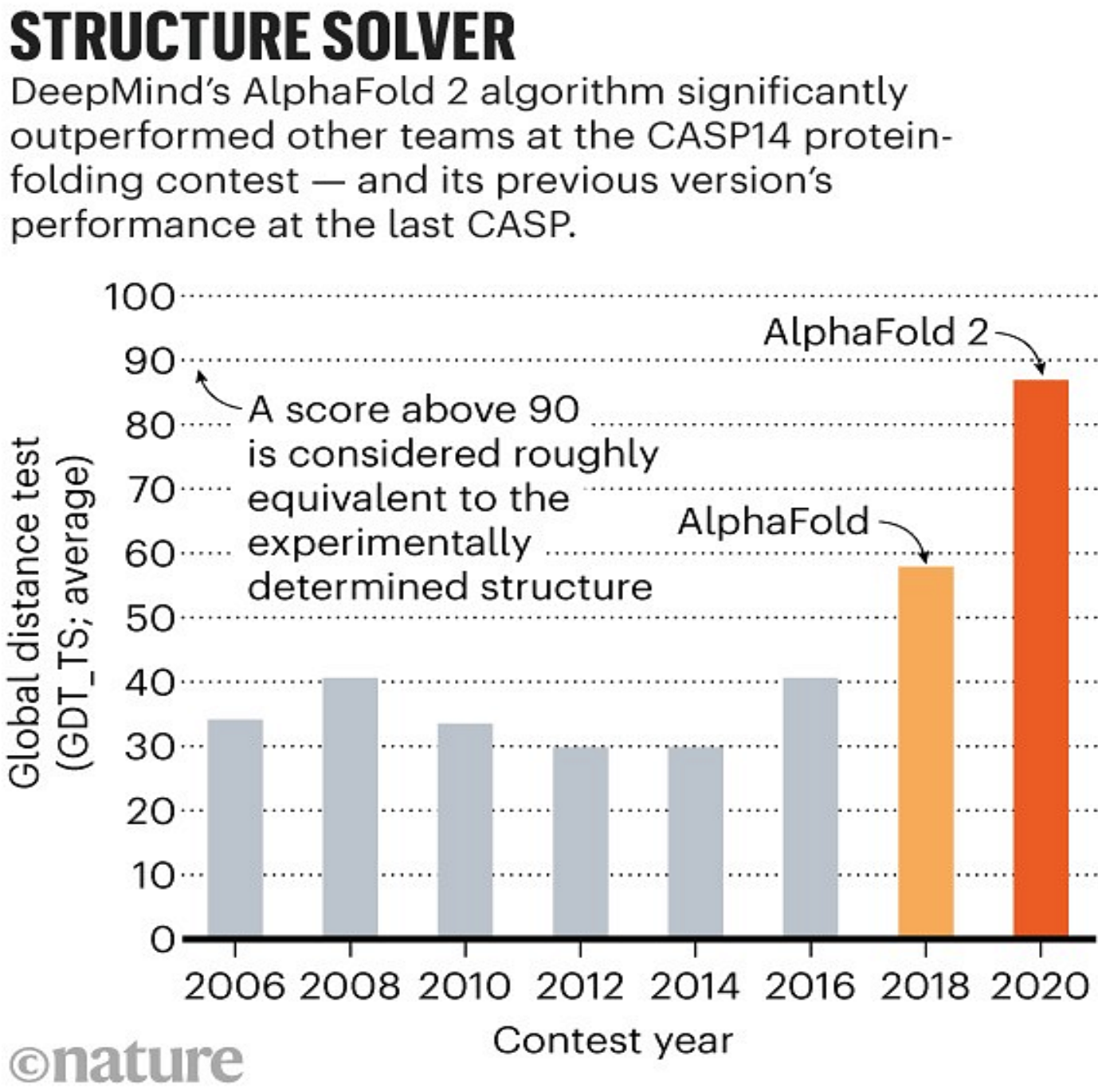
A Breakthrough in Protein Folding Unfolds

Author: Dr. Vishal Gulati

Posted on 01 Dec 2020

Just like there are Olympic Games for world's greatest athletes, for computational biologists, there is CASP (Critical Assessment of Structure Prediction). We just had the 14th competition (it takes place biannually). Until recently, it was just academic groups that played this game but now Tencent and Microsoft also show up. Like CASP13 in 2018, DeepMind's AlphaFold came top but that is not the big news. The big news this time is by how much.



The full paper is not yet published but AlphaFold's 2020 performance is the biggest advance in CASP ever. Last year the average performance of AlphaFold was around 60%, this year it is nearly 90%. The only problem for DeepMind now is that as far as CASP is concerned, there isn't much more left to do but if it wants to get started on solving some of the world's biggest problems, this is just the start.

Why Proteins and their Folding Matters

How proteins fold determines what proteins do. Millions have become infected and hundreds of thousands have died in the current pandemic because the spike protein on the surface of Cov-2 coronavirus binds very tightly to the ACE receptor in our nose and that is due to the 3D structure of these two proteins. If the 3D structure of the spike protein was even slightly different as it is in the case of other coronaviruses, the course of this pandemic would also be different. Certain 3D structures of (misfolded) proteins can cause mad cow disease, or Alzheimer's disease or cancer, and specific 3D structure of other proteins allows us to treat cancers.

Knowing 3D structures of proteins is a fundamental biology problem; thousands of scientists around the world have worked on determining the structures of thousands of proteins and thousands others work on finding new ways of determining these structures. Over the last few years, scientists have been able to determine the structure of ~90,000 high resolution protein structures. These are deposited in a database rather unimaginatively called the Protein DataBank (or PDB in short). The importance of 3D structures of proteins is so significant that it is my estimate between 30 to 50 Nobel laureates in Medicine, Physics and Chemistry were chosen for their role solving 3D structures of important proteins or advancing human knowledge on how to do so.These are critical molecules which are important as well as valuable. Trillions of dollars of market cap of the pharmaceutical industry is based on either drugs that bind to proteins or are proteins themselves.

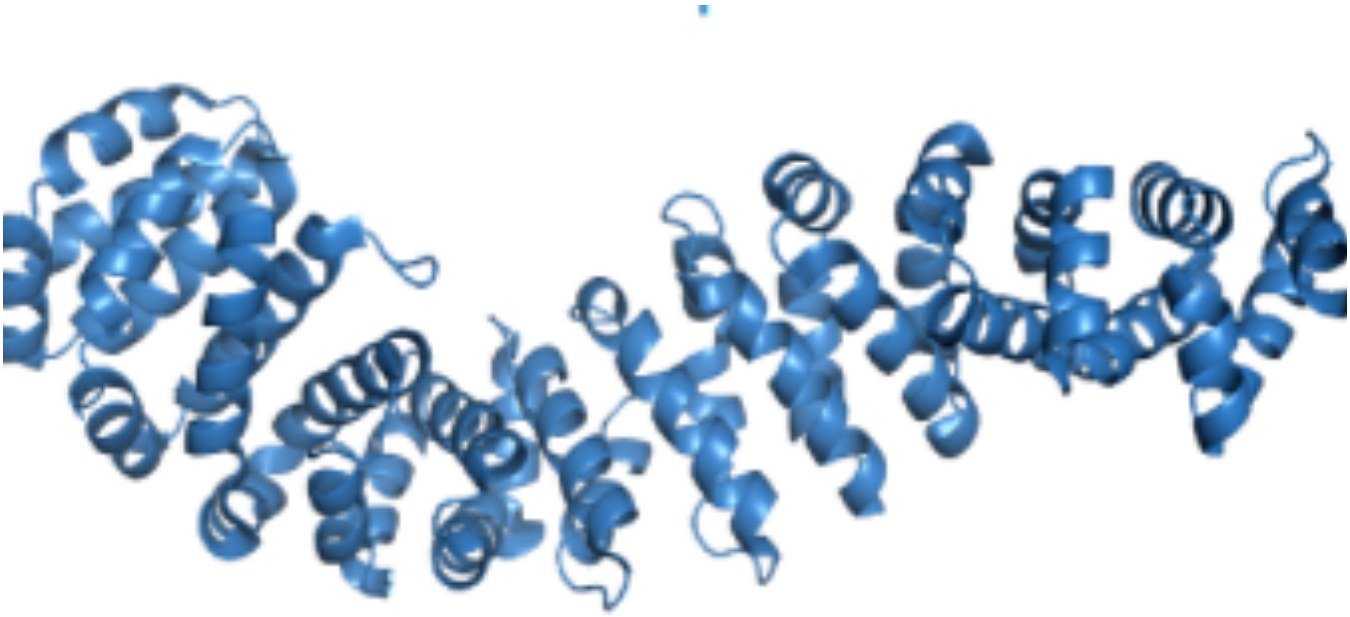
The "Protein Folding Problem"

To try and think of the protein folding problem, imagine you have a two meter piece of ribbon. And every time you throw this ribbon in the air, it lands on the ground in a very complex shape and every time you tossed a two meter piece of ribbon of that type, it always lands in the same complex shape. Not similar, not kinda, but within one millionth of a millimetre.

Now imagine that everyone on this planet tried to toss this piece of ribbon and every time

it landed in exactly the same shape every time.

Please, stay with me here. Imagine that it is possible to know how a ribbon like this would have landed 1,000 years ago, 10,0000 years ago and even a million years ago and it always landed into the same complex mess.



Now imagine a world where the best scientists that knew the 1D structure of the ribbon very well but could not tell you how a piece of ribbon will land until they threw it up in the air. All they could do when they discovered a piece of ribbon was to throw it, collect it, and then use very expensive machines to measure it carefully to determine the structure of this mess. And each of these measurements took 3-4 years of a scientist to do. And then only learn about this piece of ribbon. For the next one, they had to start over. This is what we do for proteins. The ribbon is the 1D structure and the mess on the floor is the 3D structure.

In nature, what is truly mind blowing is that proteins don't try out the millions of different configurations before settling into the final one, they just 'know' which one is the right one and generations of scientists have been baffled by it. Given the complexity of different forces and constraints acting on different atoms in each amino acid in the chain: hydrogen bonds, van der Waals interactions, electrostatic interactions each step in the folding process changes the configuration such that the rest of the protein chain now operates in entirely different space. Ultimately, the 3D structure will be determined by what configuration of the amino acids is thermodynamically stable. This means that to solve the problem by brute force unless you know what happens first you can't guess

what will happen next. In many cases the protein folding takes place within seconds which makes the overall complexity of this problem overwhelming.

Is The Protein Folding Problem Really Solved?

Despite this being the biggest thing that has happened in protein folding, the protein folding problem is *not yet solved*. Compared to the problem of protein folding, CASP is a game. It is a very hard game but it is a reduced problem set which helps us train our tools and standardise performance. It is a necessary step but it is not sufficient. AlphaFold has aced CASP14 which is not a minor achievement and the start of something new and something big but it is not something where we can enter 1D data at one end and start getting 3D data on the other. We are a fair bit away from that. AlphaFold solves the 'what' of the problem in a limited way but not the 'how' of the problem yet but this is also probably the best chance for us to eventually solve the 'how'.

What is the best thing about AlphaFold is also what limits its value in other ways. A black box approach has the problem of not being able to know if the ab initio structure of a new protein predicted by AlphaFold is the right one. Scientists will continue to confirm AlphaFold's predictions in real work anyway.

And finally, while the results have been released the actual algorithm is not available for use by scientists at this moment. Given that DeepMind is not a publicly funded research institution, it may never be published.

How Will this Impact the Healthcare Industry?

Nothing will happen overnight and at least not directly in the classical drug discovery pathway. Even if we had a magical algorithm which would spit out 3D structure of every protein on demand, the major impact of this new invention will certainly not be on classical drug discovery. In the 90's we thought if only we had high resolution structures of more proteins we would have more novel drugs. In my previous fund, we even invested in a company doing just that. We now realise that just having more protein structures may not lead to more drugs because we have thousands of high resolution structures already and most of them remain 'undruggable' (i.e. not accessible to being interfered with for medical purposes). And many of our challenges are downstream from knowing protein structure.

That said, the possibilities this opens are far more exciting than just classical drug discovery. We expect algorithmes like this will evolve to help us understand protein-

protein interactions better. We are currently at a very early stage of understanding these interactions and this is an area we are very excited about. Fluidic Analytics developed a device for understanding protein interactions (I wrote about this before: Social Network, But For Proteins). They are applying it to understanding interactions between antibodies and their targets (including in COVID-19).

Another area where such an algorithm could help us make a big leap is to understand protein misfolding. Distorted, degraded and dysfunctional proteins cause many deadly diseases and play a role in ageing and inflammation.

The third area which is a bit longer term but could transform a whole industry is synthetic biology. By reverse engineering our knowledge of protein folding we could start synthesising proteins which analogues of which have never existed in nature but designed for a specific function. This has been tried but is currently it is too slow to be truly impactful. We have often imagined that many use cases for Evonetix (DNA synthesis in Cambridge) would include designing proteins on demand for a certain function rather than replicating what already exists in nature. As with any new technology the benefits are likely to be things we cannot even envisage today. But if you want to take a peek, try this TED Talk by David Baker (David's Rosetta Consortium used to win most CASP competitions before AlphaFold showed up... much like Arsenal before Manchester City)

To understand what this discovery means we have to take a step back and reflect. Like many things, the achievements of AlphaFold will be overestimated in the short term (hyperventilating press coverage is already here!) but underestimated in the long run. We often look back on the 20th Century for its achievements in biology, but It is truly amazing to be living in this century of advancements in biology and watching some of the biggest problems being solved in front of our very eyes.