

Structural revolution?

A potential breakthrough in our ability to predict protein structure from sequence could have a major impact on evolutionary biology.

The recent announcement that AlphaFold2, a deep-learning program from Google AI's DeepMind, had won the biennial Critical Assessment of Structure Prediction (CASP) competition by a **substantial margin** caused a stir not only in scientific circles but in the mainstream media as well. Despite the wealth of genome sequence data we currently possess, there has been no reliable high-throughput way to turn this into information about the structure of proteins. This is because the physics of how a newly translated sequence of amino acids folds into a mature structure is computationally intractable. Instead, expensive and laborious techniques such as X-ray crystallography and cryo-EM, available only to specialist laboratories, are used to determine structures one at a time. But it is structures rather than sequences that determine how proteins, and therefore organisms, function, and hence there has been a bottleneck right at the root of biology. While much of the media attention focused on the biomedical consequences of this new development, the implications for studies of evolution could be just as profound.

Before we consider some of the new possibilities, however, it's important to note some words of caution. Foremost is that we don't yet have enough details for a rigorous assessment of AlphaFold2 by the scientific community. The code has not been released or peer reviewed, and we therefore do not know to what degree its impressive performance on the 2020 CASP structures would be replicated on other proteins. To some extent, deep-learning approaches always have an element of the black box about them, but full access to how the program is trained and the ability to test

it more widely is an important next step. Such testing will help to determine whether the predominance of human and medically relevant structures in existing databases has limited AlphaFold's performance on other types of protein. It has also been noted that it performs well on structures known from crystallography data, but there may be anomalies when determining what actually occurs in realistic biological solutions.

From an evolutionary standpoint, the parts of proteins we care most about are usually those where a small sequence change can have a major functional consequence, and it may be that these are precisely the regions where AlphaFold struggles the most, because its approach is based at least partially on the assumption that similar sequences fold in similar ways. It may therefore make good predictions at the level of a protein family, but distinguishing between close homologues that differ functionally may be more of a stretch. We also don't know how well it will work on large disordered proteins with low levels of secondary structure, or how good it will be at predicting interactions between proteins in multimers, which are often the actual unit of biological function.

But if AlphaFold lives up to just part of the hype, or if it spurs similar efforts that perform even better, evolutionary biologists have much to look forward to. The idea that we could quickly obtain medium-quality structures for thousands of proteins across the tree of life is an exciting one for obtaining broad insights into how phenotypes have diversified. At the very least, structural biology could become a part of the toolkit for more than a mere handful of specialist evolutionary laboratories. It

might be possible to rapidly predict the structural, and possibly functional, effects of a wide array of potential mutations, and reconstructing ancestral protein structures would not involve the years of painstaking work that has hitherto been pioneered by a few laboratories. Pinpointing the genetic basis of adaptation could get substantially easier if the functional consequences of mutations detected in genetic analyses could quickly be predicted. There are also potential insights to be had into how early proteins arose from simple sequences, and how *de novo* genes become functional, if we can understand the structural steps involved. And we may be able to identify homologies that are not obvious at the sequence level, because structure is often conserved even as sequence diverges.

It may be that deep-learning approaches are only able to get us so far, and that detailed studies of epistatic interactions over evolutionary time are still experimentally and computationally laborious. But even as a first pass at structure prediction, they are bound to make a large impact. Deep-learning approaches have already been applied to other areas of evolutionary biology such as phylogenetic reconstruction, and their use is likely to expand further. Determining how proteins interact would be a particularly welcome follow-up to the new work on the structure of individual proteins. While it is still early days for the DeepMind approach to structure, and the caveats above are very real, there is a sense of cautious optimism amongst structurally inclined evolutionary biologists. □

Published online: 15 December 2020
<https://doi.org/10.1038/s41559-020-01381-0>