News & views

Computational biology

Protein-structure prediction revolutionized

Mohammed AlQuraishi

The full might of a world-leading artificial-intelligence laboratory has been brought to bear on protein-structure prediction. The resulting method, AlphaFold2, promises to transform our understanding of proteins. **See p.583 & p.590**

Most proteins self-assemble into specific 3D structures that, together with other biological molecules, determine the function and behaviour of cells. Over the past five decades, biologists have experimentally determined the structures of more than 180,000 proteins and deposited them in the Protein Data Bank[1], a freely available online resource. Despite this painstaking effort, the structures of hundreds of millions of proteins remain unknown, including more than two-thirds of those in the human proteome – the full set of proteins produced by our genome.

In two papers in this issue, scientists at DeepMind, Google's London-based sister company, describe a machine-learning method, AlphaFold2, that predicts protein structures with near-experimental accuracy[2], and report its application to the human proteome[3]. DeepMind has also announced that it has applied AlphaFold2 to the proteomes of 20 model organisms (see go.nature.com/2w6zhus). AlphaFold2 is free for academics to use and, in collaboration with the European Bioinformatics Institute in Hinxton, UK, DeepMind will make the predicted structures of almost all known proteins freely available to all.

AlphaFold2 – as the name implies – is the second iteration of a system that DeepMind introduced three years ago at the Thirteenth Critical Assessment of Structure Prediction (CASP13) competition. The first version of AlphaFold was technically impressive[4], and outperformed the other CASP13 entrants at the task of predicting protein structures from amino-acid sequences. However, it had a median accuracy of 6.6 ångströms for the most difficult set of proteins tested – that is, for the middle-ranked protein in the set, the atoms in the proposed structures were, on average, 6.6 Å away from their actual positions. This is much less accurate than experimental methods. Moreover, the original AlphaFold arguably represented only an incremental improvement over competing algorithms, in both design and performance.

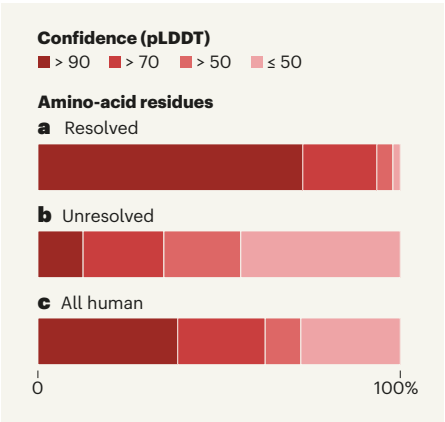

Figure 1 | The confidence of protein-structure predictions by AlphaFold2. Jumper *et al.*[2] report a machine-learning system, called AlphaFold2, that predicts the 3D structures of proteins from amino-acid sequences. Tunyasuvunakool *et al.*[3] used the same system to predict the structures of all human proteins that self-assemble into specific 3D structures. AlphaFold2 produces a confidence metric called the predicted local distance difference test (pLDDT) to estimate how well the predicted position of each amino-acid residue agrees with experimentally determined positions, on a scale of 1 to 100. The charts show the fractions of residues corresponding to different ranges of pLDDT for: **a**, residues that were previously resolved in structure-determination experiments (3,440,359 residues); **b**, residues that could not be resolved in experiments (589,079 residues); **c**, all of the residues in human proteins (10,537,122 residues). (Data from ref. 3.)

AlphaFold2 fundamentally changes this. Its median accuracy at CASP14, which was held in 2020, was 1.5 Å – comparable to the width of an atom and approaching the accuracy of experimental methods. Moreover, its design has few parallels with existing algorithms.

The prediction of protein structures is difficult for many reasons: the number of plausible shapes for any given protein is huge, but an algorithm must pick just one; the number of known structures is (relatively) small, limiting the data available for training structure-predicting systems; the rules underlying protein biophysics are only approximately known, and are expensive to simulate; and the forces that determine a protein's structure result not only from local interactions between nearby chemical groups in the protein molecule, but also from long-range interactions spanning the whole protein. Jumper *et al.*[2] (page 583) report a multitude of ideas to address these challenges in their design of AlphaFold2.

Central to this design is a machine-learning framework – known as an artificial neural network – that considers both local and long-range interactions in protein molecules. This differs from previous algorithms, which commonly considered only local interactions to reduce the computational burden of structure prediction. AlphaFold2 does not try to capture long-range interactions through computational brute force, which would be hopeless even with the resources available at Google. Instead, the authors introduced computational operations that efficiently capture long-range interactions on the basis of fundamental aspects of protein geometry. For example, the operations account for the fact that the coordinates of any three atoms in a protein must satisfy the triangle inequality rule (in other words, the sum of the lengths of any two sides of the triangle defined by the coordinates must be greater than or equal to the length of the remaining side).

AlphaFold2 applies these operations repeatedly (about 200 times) to gradually refine a model of a protein into its final 3D structure. Such iterative refinement, used millions of times, rather than hundreds, is a central component of physics-based approaches to protein-structure prediction[5]. But it is rarely used in machine-learning approaches – which instead predict structures by recognizing patterns of mutation in evolutionarily related proteins to detect co-evolving, and therefore spatially proximal, amino-acid residues[6]. AlphaFold2 breaks the mould by combining these two strategies. Crucially, it does not

impose known rules of protein biophysics or try to mimic the physical process of protein folding, as has previously been attempted[7,8]. Instead, it performs purely geometric refinements learnt from its repeated attempts to predict protein structures. In this sense, it exemplifies the learning-driven revolution that has swept the field of protein modelling[6,9].

In a companion paper, Tunyasuvunakool *et al.*[3] (page 590) report the use of AlphaFold2 to predict the structures of almost all human proteins that independently acquire well-defined 3D shapes, for a total of 23,391 proteins. Predictions at this scale were previously possible, but three features of the new system provide a big leap forward.

First, the accuracy of the predictions is sufficiently high to generate biological insights and hypotheses that can be tested experimentally. Second, a calibrated self-assessment of each prediction provides a reliable estimate of correctness at the level of individual amino-acid residues (Fig. 1), enabling biologists to make inferences about confidently predicted regions. Third, AlphaFold2 is applicable to whole proteins, including large ones that have multiple, independently self-assembling units — a common feature of mammalian proteins. The resulting resource 'confidently' predicts nearly 60% of all human-protein regions; most of the remaining regions might be unable to acquire well-defined structures, or be able to do so only in the presence of other biomolecules.

AlphaFold2 has already helped structural biologists to solve crystallographic protein structures[10] and refine ones derived from cryo-electron microscopy experiments. It provides biophysicists studying protein motion with starting (static) structures, and those studying protein interactions with hypotheses about how protein surfaces bind to each other. AlphaFold2 also presents opportunities to formulate new algorithms for bioinformatics based on protein structures, and might help systems biologists to understand the behaviour of cellular pathways and molecular machines on the basis of the structures that comprise them. And the study of evolution, which has long relied on genetic sequences, can now more readily be formulated in terms of the onset of new classes of protein structure (folds) and their relationship to cellular function and organismal fitness.

It is tempting to compare the scale of this advance to that of the Human Genome Project, but there are important differences. In contrast to the human genome sequence, the predicted structures have not been experimentally verified; it will take time for evidence of their correctness to emerge, so that scientists can gain confidence in the predictions. Of course, experimental measurements can also be affected by 'noise', bias and incompleteness — 20 years passed between the publication of the first draft of the human genome and the complete sequence[11] — and modern structure-determination techniques routinely involve some computational inference. As predictions improve, disagreements between protein models and experiments could become difficult to resolve, a situation familiar to physicists[12] but largely unprecedented in biology.

Disordered protein regions, which do not have well-defined shapes but often encode functionally crucial parts of proteins, present an ongoing and fundamental challenge to AlphaFold2 and, therefore, to our understanding of protein structure. Future methods must take this disorder into account and begin to reflect the flexibility inherent in most proteins.

Other differences between the Human Genome Project and the present advance are in AlphaFold2's favour. Structure predictions are (relatively) cheap and will soon be available for all proteins, whereas genetic-sequencing technology took years to deploy and mature. Computational methods evolve rapidly, and it might therefore soon be possible to predict the structures of multi-protein complexes, alternative conformations of a protein (for proteins that adopt them) and the structures of designed proteins with a level of accuracy similar to that currently achieved by AlphaFold2. Finally, protein structures provide immediate biological insights, because they fit within established conceptual frameworks that relate a protein's structure to its function — unlike genetic sequences, which were largely inscrutable at the dawn of the genomics era. The fruits of this revolution might thus be more swiftly reaped.

Mohammed AlQuraishi is in the Columbia University Irving Medical Center, Columbia University, New York, New York 10032, USA.
e-mail: m.alquraishi@columbia.edu

1. Bernstein, F. C. *et al. J. Mol. Biol.* **112**, 535–542 (1977).
2. Jumper, J. *et al. Nature* **596**, 583–589 (2021).
3. Tunyasuvunakool, K. *et al. Nature* **596**, 590–596 (2021).
4. Senior, A. W. *et al. Nature* **577**, 706–710 (2020).
5. Kuhlman, B. & Bradley, P. *Nature Rev. Mol. Cell Biol.* **20**, 681–697 (2019).
6. AlQuraishi, M. *Curr. Opin. Chem. Biol.* **65**, 1–8 (2021).
7. Jumper, J. M., Faruk, N. F., Freed, K. F. & Sosnick, T. R. *PLoS Comput. Biol.* **14**, e1006578 (2018).
8. Ingraham, J., Riesselman, A., Sander, C. & Marks, D. *Int. Conf. Learning Representations* https://openreview.net/forum?id=Byg3y3C9Km (2019).
9. Wu, Z., Johnston, K. E., Arnold, F. H. & Yang, K. K. *Curr. Opin. Chem. Biol.* **65**, 18–27 (2021).
10. Millán, C. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2021.06.21.449228 (2021).
11. Nurk, S. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2021.05.26.445798 (2021).
12. Brumfiel, G. *Nature* https://doi.org/10.1038/nature.2012.10249 (2012).

The author declares no competing interests.

Biogeochemistry

Carbon stocks of African montane forests assessed

Nicolas Barbier

The inaccessibility of African montane forests has hindered efforts to quantify the carbon stored by these ecosystems. A remarkable survey fills this knowledge gap, and highlights the need to preserve such forests. **See p.536**

On page 536, Cuni-Sanchez *et al.*[1] report the assembly of a large database of tree inventories for 226 mature montane-forest plots in 12 African countries. The authors analyse the data to determine the amount of aboveground biomass and carbon stored in these highly diverse and threatened ecosystems. Their results suggest that African montane forests store more carbon than was previously thought, and the findings should help to guide efforts to conserve these ecosystems.

Cuni-Sanchez and colleagues measured trunk diameters and heights of the trees in plots, and identified the botanical species to deduce wood density — an approach that constitutes the gold standard for estimating the biomass, and thus the amount of carbon, contained per unit of forest area. This method involves the use of general statistical equations for describing tree form, called allometric models, and considers only the aboveground parts of trees. It therefore disregards several other pools of carbon, notably in the roots and soil. The overall approach might seem crude, but recognizing and measuring the many hundreds of tree species found on steep, cloud-shrouded slopes (Fig. 1), let alone the underground carbon, without visiting the sites,