AlphaFold2 and the future of structural biology

To the Editor — AlphaFold2 is a machine-learning algorithm for protein structure prediction that has now been used to obtain hundreds of thousands of protein models. The resulting resource is marvelous and will serve the community in many ways. Here I discuss the implications of this breakthrough achievement, which changes the way we do structural biology.

Imagine a website where you could download a reliable three-dimensional model of your protein of interest. Until recently, this was just a dream. Now such structure prediction has become reality, at least for many monomeric proteins. As a result of a collaboration between the company DeepMind and the European Molecular Biology Laboratory, hundreds of thousands of protein models were published online 22 July 2021.

It has been a long-term goal of the scientific community to provide structural information on the human proteome. However, despite decades of effort, only ~18% of the total residues in human protein sequences are covered by experimentally determined structures at this time. This coverage has now been widely expanded by structural modeling with DeepMind's machine-learning algorithm, AlphaFold2 (ref. [1]). As a result, the authors could double the number of residues in the human proteome that are covered with high-confidence three-dimensional information[2].

Machines learn to predict structures

AlphaFold2 incorporates empirical knowledge about protein structure into a deep-learning algorithm[1]. The algorithm also makes use of information from evolutionary conservation in the form of multiple-sequence alignment[3]. The resulting protein models are often as accurate as experimentally determined structures. Indeed, AlphaFold2 had outcompeted other prediction methods in a blind test, the 14th Critical Assessment of Protein Structure Prediction (CASP14)[4]. Whereas the original code was not public[5], the expanded code and the pretrained model for AlphaFold2 are now available for download via GitHub[1].

Also very recently, an academic team led by David Baker has provided an alternative machine-learning algorithm for structure prediction that is called RoseTTAFold[6]. This algorithm builds on the deep-learning approaches established during the development of AlphaFold2 and has already been applied to predict structures of several protein complexes. Like AlphaFold2, RoseTTAFold is available to the community and can now be used as an alternative route to predict protein structure from sequence.

AlphaFold2 and the community

Half a century ago, the structural biology community had decided that all experimentally resolved macromolecular structures should be collected in an open-access database, the Protein Data Bank (PDB)[7]. The PDB has been a great investment in the future and was essential for training the machine-learning algorithm of AlphaFold2. From the features learned during this training on experimentally determined structures, the algorithm could predict unknown structures with considerably higher accuracy than what has been achieved before.

The vast structural knowledge available in the PDB was thus a *conditio sine qua non* for developing the new prediction tools. Obtaining the many experimental structures that are collected in the PDB has required decades of hard work by the structural biology community and has remained challenging despite many advances made over the years. Now these efforts are paying off in a way that could not be imagined until recently. The new prediction algorithms distil the PDB to provide a tool that facilitates and accelerates structure determination and the use of structural knowledge in biomedicine.

How structural biology will change

The new algorithms will change how we do structural biology. First, they will facilitate structure solution of large assemblies by cryo-electron microscopy (cryo-EM). This approach generally requires detailed structures of the individual proteins or their domains as a starting point. It is expected that, if the individual structures are not available, they will now simply be downloaded as predicted models and fitted to the cryo-EM densities. The obtained fits may then be confirmed with the use of protein crosslinking and mass spectrometry[8].

Predicted structures may also be used as search models to solve X-ray crystal structures by molecular replacement[9], thereby making experimental phasing obsolete in many cases. Researchers using NMR may also benefit from the prediction algorithms. The time-consuming de novo solution of domain structures by NMR may be replaced by fast predictions so that the unique advantages of NMR in investigating protein folding and dynamics and the binding of ligands and nucleic acids can be utilized more readily.

The new prediction algorithms should also improve automated model building. This will not change the general approach in structural biology, which has always combined model building with experimental observations. The best-known example may be the DNA double helix, which was originally modeled to fit experimental observations that came from X-ray fiber diffraction and biochemistry[10]. Until today, structural models were built to explain experimental data, but soon machine-learning methods may be combined with classical refinement tools to largely automate model building, to the benefit of the community.

New challenges for computational biology

The new algorithms will be used to predict the structured proteome of any organism that is sequenced. Such predictions may help in the design of specific scientific projects, but they will also accelerate drug discovery and foster biotechnology applications. Large-scale predictions may additionally result in a new type of comparative structural proteomics, which, it is assumed, will lead to new discoveries. These developments, however, require that computational biologists stay closely connected to the experimental community.

In the near future, machine learning should be explored for predicting structures of protein–nucleic acid complexes, which are a notable blind spot of AlphaFold2 and RoseTTAFold. The PDB already contains nearly 10,000 entries for protein–nucleic acid complexes that should be used for training new algorithms. Whereas predicting protein–DNA complexes may be in reach, experimentally resolved protein–RNA complex structures remain low in number, and training sets are thus small, which may impair success at this time.

New machine-learning tools should also be developed to analyse and predict conformational changes in proteins and to solve structures of polymorphic assemblies and protein fibers[11]. Machine-learning methods should also enable a better prediction of protein function and facilitate protein engineering and design[12]. Finally, up

to half of the human proteome is estimated to encode for intrinsically disordered regions, which often engage in multivalent interactions to form transient compartments in cells[13]. There is currently little structural information on such protein regions, but machine-learning tools may help us to better characterize such systems once more training data become available.

The future of structural biology

A long-term goal of structural biology remains the visualization of molecular structures in their natural context, which is often referred to as in-cell or in situ structure determination. Indeed, recent advances in cryo-electron tomography[14], data processing[15] and chemical crosslinking and mass spectrometry[16] demonstrate the feasibility of this approach. However, at this time, in-cell structural biology is limited to certain types of simple cells, parts of cells or exceptionally large and stable molecular complexes.

In-cell structural biology would benefit from the further development of machine-learning algorithms that would enable us to reliably predict structures of protein complexes that could then be used as templates to mine tomography data. However, experimental studies of such large complexes have revealed their transient nature and plasticity, showed that their integrity often depends on nucleic acids and small molecule cofactors and found that protein–protein interfaces are often very small in size. Therefore, accurate prediction of protein complexes will probably remain a formidable challenge for the foreseeable future and will rely on improved platforms to integrate information from various sources.

It will therefore probably require intermediate steps to achieve the transition from current state-of-the-art integrated structural biology to future in-cell structure determination. Over the coming years, structural biologists will probably try to resolve large endogenous assemblies and visualize isolated cellular compartments. Such studies will certainly benefit from the localization of proteins by fluorescence-labeling and high-resolution light microscopy[17].

Conclusions

The new prediction algorithms do not solve the protein folding problem in the sense that they do not reveal how a sequence encodes three-dimensional structure. However, they do solve the problem in practical terms, as they can reliably predict structure from sequence, at least in many cases. Although only time will tell, this advance is expected to represent a breakthrough in structural biology that is comparable to previous major advances, such as the introduction of synchrotron radiation[18] and selenomethionine phasing[19] for X-ray crystallography or the development of direct electron detectors for cryo-EM[20].

In summary, the recent advances in protein structure prediction that result from new machine-learning algorithms mark the beginning of a new era in structural biology. They will accelerate life science research and will facilitate many biomedical applications that require structural knowledge. The advances are also testimony to the power of artificial intelligence and open science, and they provide a seminal example of how transformative research may be done in the 21st century, to the benefit of science and society. ❐

Patrick Cramer
Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany.
✉*e-mail: patrick.cramer@mpibpc.mpg.de*

Published online: 10 August 2021
https://doi.org/10.1038/s41594-021-00650-1

References

1. Jumper, J. et al. *Nature* https://doi.org/10.1038/s41586-021-03819-2 (2021).
2. Tunyasuvunakool, K. et al. *Nature* https://doi.org/10.1038/s41586-021-03828-1 (2021).
3. Steinegger, M. et al. *BMC Bioinformatics* **20**, 473 (2019).
4. Pereira, J. et al. *Proteins* https://doi.org/10.1002/prot.26171 (2021).
5. AlQuraishi, M. *Bioinformatics* **35**, 4862–4865 (2019).
6. Baek, M. et al. *Science* https://doi.org/10.1126/science.abj8754 (2021).
7. Berman, H. M. *Nat. Struct. Mol. Biol.* **28**, 400–401 (2021).
8. Walzthoeni, T., Leitner, A., Stengel, F. & Aebersold, R. *Curr. Opin. Struct. Biol.* **23**, 252–260 (2013).
9. Rossmann, M. G. *Acta Crystallogr. D Biol. Crystallogr.* **57**, 1360–1366 (2001).
10. Cramer, P. *Cell* **182**, 787–789 (2020).
11. Masarati, G. et al. *J. Mol. Biol.* https://doi.org/10.1016/j.jmb.2021.167127 (2021).
12. Xu, Y. et al. *J. Chem. Inf. Model.* **60**, 2773–2790 (2020).
13. Borcherds, W., Bremer, A., Borgia, M. B. & Mittag, T. *Curr. Opin. Struct. Biol.* **67**, 41–50 (2021).
14. Turk, M. & Baumeister, W. *FEBS Lett.* **594**, 3243–3261 (2020).
15. Tegunov, D., Xue, L., Dienemann, C., Cramer, P. & Mahamid, J. *Nat. Methods* **18**, 186–193 (2021).
16. O'Reilly, F. J. et al. *Science* **369**, 554–557 (2020).
17. Bykov, Y. S., Cortese, M., Briggs, J. A. & Bartenschlager, R. *FEBS Lett.* **590**, 1877–1895 (2016).
18. Holmes, K. C. *Endeavour* **33**, 60–66 (1974).
19. Hendrickson, W. A., Horton, J. R. & LeMaster, D. M. *EMBO J.* **9**, 1665–1672 (1990).
20. McMullan, G., Faruqi, A. R. & Henderson, R. *Methods Enzymol.* **579**, 1–17 (2016).

Acknowledgements

I thank S. Dodonova, J. Söding and D. Tegunov for comments.

Competing interests

The author declares no competing interests.