Could AlphaFold revolutionize chemical therapeutics?

To the Editor — Seeing is believing: the recent development of AlphaFold[1,2] allows protein-structure prediction with vastly enhanced accuracy that eclipses that of all past methods and is stirring excitement in the scientific community. However, how useful are these predicted models as is? Here we discuss the implications of this remarkable accomplishment in structure-based drug design (SBDD).

AlphaFold utilizes a 'knowledge-based' algorithm that takes advantage of the wealth of protein-sequence information available and the corpus of already solved structures in the Protein Data Bank, and bypasses the need to understand the deep rules that physically determine the protein-folding process. It relates co-evolution information from multiple sequence alignments with structures via a machine-learning algorithm to predict the structure of any given protein. What AlphaFold accomplished was revelatory, running circles around its competition at 2020's CASP14[3], the premier protein-structure-prediction competition. The algorithm's accuracy, its relative ease of use and its efficacy across species and protein type make it an unquestionable game-changer for structural biology—as it streamlines the structure-solving pipeline by providing good starting models—and for biology in general, allowing model-guided mutagenesis to delineate function. However, are these predictions of sufficient fidelity to be used directly in SBDD?

Accuracy of AlphaFold

The single most critical parameter in SBDD is accuracy[4]. In protein-structure determination, this is often represented by the 'resolution' of a structure. So what kind of accuracy can AlphaFold achieve? Let us ignore prediction of proteins with suitable homology to a known structure (≥40% identity), as the results generated from homology modeling are already deemed suitable for use in SBDD. Of the remaining targets for which their Protein Data Bank entries were not part of the training set, the median backbone r.m.s.d.$_{95}$ (Cα root-mean-square deviation at 95% residue coverage) between the AlphaFold prediction and the structure is 1.46 Å, and the first-quartile r.m.s.d.$_{95}$ is 0.79 Å. Thus, in the majority of cases, AlphaFold can reproduce the correct backbone even for proteins without good templates in the Protein Data Bank.

However, although Cα accuracy may be sufficient for the creation of a suitable model for the generation of testable hypotheses and as initial models for experimental structure determination, SBDD also requires accurate side-chain information. Defining binding pockets, modeling ligands and predicting water structures all require precise knowledge of side-chain positioning[5]. Switching now to all-atom r.m.s.d.$_{95}$ as a measure of side-chain accuracy, also for the template-reduced set, AlphaFold is correct within 2 Å and 1 Å in 52% and 17% of cases, respectively. Although this median all-atom accuracy may fall short of enabling SBDD in all cases, AlphaFold returns global and per-residue accuracy estimates that correlate strongly with actual accuracy. The per-residue accuracy is especially useful, as a poorer overall candidate might have regions of good accuracy upon which SBDD can be performed. Thus, AlphaFold users can employ this information to better interpret the predictions for their utility in SBDD.

What is next for AlphaFold?

AlphaFold has the potential to greatly expand the pool of targets for SBDD via the availability of potentially useful structures. On the basis of the Therapeutic Target Database (http://db.idrblab.net/ttd/)[6], the number of targets currently sought is only about 3,500, which pales in comparison to the estimated 50,000 unique proteins in the human proteome[7]. Providing public access to accurate structural information for a larger percentage of the proteome offers a potential path to the fulfillment of the decades-old promise of the field of 'chemical genetics'[8], wherein specific small-molecule modulators of a protein target are used as tools to elucidate the function of the protein in question. For completion of the circle of protein sequence to structure to ligand to function, however, a rapid, robust and inexpensive method for identifying ligands from the protein structure will be necessary. Virtual screening is well suited for this task in theory; however, enhancements in its robustness and computational efficiency may be necessary for probing the vastness of the chemical space in a feasible way. Will new virtual screening tools emerge to leverage this repository of protein structural information in an efficient manner? Were this possible, there is little doubt that these tools could lead to an explosion in the emergence of druggable targets for drug-discovery applications.

Although the advances made by AlphaFold are great, its usage is still in its infancy, and many improvements may be made to further increase the applicability of this technology. In addition to incremental improvements in accuracy, headway could be made in expanding the algorithm's compatibility. At present, AlphaFold can handle only unmodified single-chain proteins. In contrast, most proteins function in the cell as protein–protein complexes, protein–ligand complexes, post-translationally modified forms, or a distribution over possible states. Extending the applicability of AlphaFold to these other cases would further widen the list of targets and possibly assist large-molecule therapeutics by predicting binding modes of antibodies or other protein-based modalities. Indeed, RoseTTAfold, a machine-learning structure-prediction algorithm recently developed in David Baker's laboratory that incorporates an approach similar to that of AlphaFold, supports protein–protein complexes[9]. That being said, we are eager to see what the future will bring and hope that these strides toward a comprehensive solution to structure prediction will be translated into progress in human health through drug discovery. ❐

Alexander B. Tong[1], Jason D. Burch[2,3], Daniel McKay[2,3], Carlos Bustamante[4], Michael A. Crackower[2,3] and Hao Wu[5,6]✉

[1]*Jason L. Choy Laboratory of Single-Molecule Biophysics, Institute for Quantitative Biosciences-QB3, and Chemistry Graduate Group, University of California Berkeley, Berkeley, CA, USA.* [2]*Ventus Therapeutics, Waltham, MA, USA.* [3]*Ventus Therapeutics, Montreal, Canada.* [4]*Department of Molecular and Cell Biology, Department of Physics, and Department of Chemistry, Howard Hughes Medical Institute, University of California Berkeley, Berkeley, CA, USA.* [5]*Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA, USA.* [6]*Program in Cellular and Molecular Medicine, Boston Children's Hospital, Boston, MA, USA.*

✉*e-mail: wu@crystal.harvard.edu*

Published online: 24 September 2021
https://doi.org/10.1038/s41594-021-00670-x

References

1. Jumper, J. et al. *Nature* **596**, 583–589 (2021).
2. Tunyasuvunakool, K. et al. *Nature* **596**, 590–596 (2021).
3. Callaway, E. *Nature* **588**, 203–204 (2020).
4. Batool, M., Ahmad, B. & Choi, S. *Int. J. Mol. Sci.* **20**, 2783 (2019).
5. Huggins, D. J. & Tidor, B. *Protein Eng. Des. Sel.* **24**, 777–789 (2011).
6. Wang, Y. et al. *Nucleic Acids Res.* **48**, d1031–d1041 (2020).
7. Ponomarenko, E. A. et al. *Int. J. Anal. Chem.* **2016**, 7436849 (2016).
8. Schreiber, S. L. *Science* **287**, 1964–1969 (2000).
9. Baek, M. et al. *Science* **373**, 871–876 (2021).

Competing interests

J.D.B., D.M. and M.A.C. are employees of Ventus Therapeutics, a company that uses SBDD. H.W. is a co-founder of Ventus Therapeutics, and an institutional plan is in place to manage and monitor the conflicts of interest that arise from this relationship.