Great expectations - the potential impacts of AlphaFold DB

Stephen Cusack, Sebastian Eustermann, Gerard Kleywegt, Jan Kosinski, Julia Mahamid, José Antonio Marquez, Christoph Müller, Thomas Schneider, Janet Thornton, Jessica Vamathevan, Sameer Velankar, Matthias Wilmanns

Summary

DeepMind and EMBL's European Bioinformatics Institute (EMBL-EBI) have partnered, initially for a 2-year period, to make hundreds of thousands (and eventually many millions) of AlphaFold structure predictions freely available to the community through a new data resource, AlphaFold DataBase (AlphaFold DB). AlphaFold is an Artificial Intelligence (AI) system developed by DeepMind that predicts a protein's three-dimensional (3D) structure from its amino-acid sequence. The initial release of the resource provides structure predictions for most of the proteins in the human proteome as well as for the proteomes of 20 other species of significant biological or medical interest. In the coming months the resource will be expanded to cover a large proportion of all catalogued proteins (up to 90% sequence identity; i.e. UniRef90). This means that for every known sequence in the UniProt data resource there will be either an experimentally determined structure (in the Protein Data Bank, PDB), or a predicted structure in AlphaFold DB, or the structure can be readily modeled using traditional structure-prediction techniques based on models for similar sequences in PDB or AlphaFold DB. This development represents a step-change for molecular biology - for the first time in history, for almost every protein of known sequence, a high-quality 3D model will be readily available. Here, we will discuss some of the applications that AlphaFold DB may enable and consider the possible impact of the resource on science and, by extension, on society.

Background

Predicting a 3D structure of a protein based solely on its (one-dimensional) sequence of amino acids was first introduced in 1972 as the "protein-folding problem". It was subsequently realised that proteins (or domains) with similar amino-acid sequences have similar overall 3D structures, and that the degree of this similarity (measured by the root-mean-square distance, or RMSD, between corresponding atoms in the two models) is correlated with the degree of sequence similarity. In November 2020, more than 60 years after the first protein structures were determined experimentally, AlphaFold was recognised as the best-performing method for predicting 3D protein structure by the assessors of the 14th round of the biennial CASP experiment. The best-predicted 95% of residues in AlphaFold models had a median alpha carbon RMSD of 0.96 Å to the experimental models, compared to 2.83 Å for the next-best method. Thus, the AlphaFold predictions were consistently very similar to the experimentally determined structures of the proteins included in this round of CASP. The AlphaFold information for a specific protein also includes a predicted model-quality score for individual residues. This metric predicts the regions of sequence where the model is likely to be of high quality and the regions where the model is probably less reliable. A third output from AlphaFold predicts the uncertainty in the relative position, orientation and thus distance between pairs of residues. This allows assessment of the reliability of the positioning of secondary structure elements and domains with respect to one another.

The power of open data

Since the early 1970s, the structural biology community has archived its experimental structures in the PDB, a freely available global resource that now contains over 180,000 structures. AlphaFold builds on this huge body of experimental information and generates its predictions by analysing the relationship between these known protein structures and huge amounts of protein-sequence data. This protein-sequence information has also been generated by scientists all over the world, mainly through genome sequencing, and is made available through public resources, such as UniProt and Mgnify hosted at EMBL-EBI. As a result, AlphaFold is able to produce accurate structure predictions even for amino-acid sequences that it has never encountered before.

The EMBL-EBI/DeepMind collaboration

When DeepMind decided to predict the structures of a huge number of proteins and started to think about how to disseminate the 3D models freely and openly, an obvious partner to collaborate with was the European Bioinformatics Institute, EMBL-EBI, the European home of many biological data resources used by the AlphaFold team, including PDBe and UniProt. EMBL-EBI is part of the European Molecular Biology Laboratory (EMBL), an intergovernmental research and service organisation with laboratories in several countries. EMBL has more than four decades of experience in running large-scale facilities for protein-structure determination, including synchrotron beamlines in Hamburg and Grenoble and the brand-new Imaging Centre in Heidelberg, which includes state-of-the-art electron cryo-microscopes. EMBL also plays a leading role in several distributed European research infrastructures, most notably Elixir, Euro-BioImaging and Instruct-ERIC. Part of EMBL's mission is to promote open research data and open science, and EMBL-EBI's rich repertoire of biological databases is a key component of EMBL's strategy in this area. Moreover, EMBL-EBI is a major partner in many international efforts to define and apply standards for biological data.

The DeepMind and EMBL-EBI collaboration to make the millions of AlphaFold predictions available through AlphaFold DB fits perfectly in this context. The new resource draws on the many years of experience available at EMBL-EBI in organising, visualising and presenting 3D structure data in resources such as PDBe, PDBe-KB and EMDB. The information available about the AlphaFold models can easily be enriched by integrating it with some of the many other EMBL-EBI databases containing for example functional or sequence-variation data. The initial release of AlphaFold DB contains over 350,000 structures which will eventually increase to an estimated 130 million 3D models (around 700 times more than currently in the PDB). The functionality of the website will also continuously be improved and extended.

Current limitations of the prediction method

Although the availability of predicted 3D models for the known "protein universe" is an exciting prospect with huge impact, there are nevertheless limitations to the AlphaFold method and resource, some of which may be addressed in the future:

- Many proteins function as **complexes** with other proteins, nucleic acids (DNA or RNA) or ligands. AlphaFold does not currently predict 3D structures for protein-protein or protein-DNA/RNA/ligand complexes. In some cases, the single-chain prediction may correspond to the structure adopted in a complex. In other cases (especially if the protein is structured only upon binding partner molecules) the missing context from surrounding molecules may lead to an uninformative prediction.
- Proteins are **dynamic systems** and adopt different structures depending on their environment or state within a functional cycle. Where a protein is known to have **multiple conformations** AlphaFold will usually only produce one of them. This leaves open many interesting questions about the conformational dynamics of proteins, crucial for understanding biological function, and this will remain a very active area of research.
- For **regions that are intrinsically disordered or unstructured** in isolation, AlphaFold is expected to produce a low-confidence prediction and the predicted structure will have an extended, ribbon-like appearance. AlphaFold may be of use as a tool for identifying such regions, but the prediction makes no statement about the relative likelihood of

different conformations (in biophysical terms: it is not a sample from the Boltzmann distribution). Furthermore, AlphaFold does not claim to predict the "folding pathway".

- AlphaFold has not been trained or validated for predicting the **effect of mutations**. In particular, it is not expected to capture the effect of point mutations that destabilise a protein.
- **Ligands are not included** in the structures since AlphaFold does not make any predictions about any of the non-protein components that are often observed in experimental structures (such as cofactors, metals, ligands including drug-like molecules, ions, carbohydrates and other post-translational modifications).
- As with experimental structures, predicted structures **may (or may not) lead to hypotheses about the function** of the protein and the mechanism underlying that function, but such hypotheses then have to be tested by further experimentation.

Benefits to the scientific community

The protein-structure predictions in AlphaFold DB will have an immediate impact on molecular structural biology research, and in a longer perspective, a significant scientific, medical and eventually economic impact. This step change will catalyse a huge amount of research in new areas, and the development of applications that were previously impossible, impractical or limited in their scope by the hitherto relatively restricted amounts of 3D structural information available. In the following, we examine the potential impact of AlphaFold DB on a number of different research areas and communities.

Opportunities for Structural Biology research

Structural biology is a branch of molecular biology that uses 3D structural information (ideally with atomic resolution) to answer biological questions, e.g. about the mechanism or function of a protein or complex. To this end, structural biologists usually determine multiple structures of the same protein, e.g. with ligands, with certain mutations, or in complex with other macromolecules including other proteins or nucleic acids.

Accelerating structure studies: the availability of predicted 3D models on a large scale is likely to significantly change the landscape of structural biology research, in some cases accelerating structural analysis. Currently, the PDB contains over 180,000 entries which cover ~55,000 unique proteins (UniProt accessions). The limited coverage in the PDB of the protein universe (~220 million sequences in UniProt or ~625 million sequences in MGnify) is an impediment for many areas of biology, including for structural biology itself. The availability of large numbers of predicted models can be used to kick-start experimental de novo structure determination (e.g., by providing phases through molecular replacement), even with low-quality or low-resolution data sets, which would otherwise be difficult or impossible to solve. In the short term, it will also help in determining structures for which experimental data was collected perhaps years ago but which have hitherto resisted all efforts to solve them. This includes over 5,700 cryo-EM maps in EMDB which could not be interpreted in terms of an atomistic model previously. Another issue is that experimental structural biologists often have difficulties expressing full-length proteins. Predicted models may help in dissecting the protein into functional domains, which can then be expressed individually or in combinations.

Cryo-EM has emerged as one of the main methods to determine the structures of large and flexible protein complexes and "molecular machines". It is expected that major and important complexes will not be resolved completely or in their entirety to high resolutions. The predicted models will be important for interpreting low-resolution regions by integrative modeling and simulations and for accelerating model building of the better resolved regions.

Filling in the components of protein complexes: AlphaFold DB will make it possible to study complex biological systems where experimental structural data at very high resolution or quality is not available, and may provide mechanistic hypotheses regarding the function of large macromolecular machines. Experimentalists could determine the shapes of proteins of interest in complex with other relevant proteins, DNA/RNA or small-molecule ligands to obtain a picture of the range of conformations and states such complexes adopt to carry out their function. Where experimental models of component proteins are not available, AlphaFold models could be used in a complementary manner. The models of such complexes could be used to generate hypotheses regarding relevant binding sites or interaction surfaces and subsequently to plan additional experiments, e.g. to find out which ligands (or fragments of ligands) could bind.

Generating hypotheses for analysis of protein dynamics: the availability of a 3D model for a protein may stimulate experimental analysis of its dynamics. Predicted models could be used to fit low-resolution data from methods such as small-angle scattering (SAS) or to inform time-resolved studies to understand enzyme catalysis mechanisms or conformational changes in side chains that bind ligands to confer specificity or selectivity.

Modelling of large macromolecular complexes using integrative and hybrid methods for structure determination (I/HM) will similarly benefit from the availability of predicted models for the building blocks of such complexes as well as from the anticipated increase in the number of macromolecular complex structures. Rapid advances in cryo-EM have made it possible to study macromolecular complexes in their biological context (the cell) using in-situ experiments. The predicted models may help to elucidate the identity of proteins that interact with large complexes in various contexts in a cell.

While AlphaFold DB will, in general, accelerate structural biology research, it will likely also induce a shift in emphasis from initial structural determination to the study of the more mechanistic and functional aspects of protein structures. Although this in turn may lead to an objective re-evaluation of the large-scale structural biology infrastructures devoted to structure determination (e.g. synchrotron X-ray crystallography beamlines), it is likely that for the foreseeable future they will be essential to validate and thus fully harness the potential of structure prediction, and to enable structural investigations for which no reliable predictions can be made at this time (structure of nucleic acids and large complexes, ligand and fragment screens, investigations of dynamics, etc.).

Opportunities for Structure Prediction research

Accurate prediction of the 3D atomic structure (or fold) of a protein from its sequence has been a “Holy Grail’’ of biology for many decades and a considerable global research effort has been dedicated to its pursuit. The success of AlphaFold in CASP14 constitutes a step change for this field.

Moving to new challenges in prediction: the availability of a 3D model for almost every protein will likely shift the focus of the structure-prediction community to different challenges, such as predicting the structure of (large) complexes and assemblies and predicting how drugs and other small molecules interact with proteins (docking, virtual screening), an area of intense research. Deep-learning technology may become a mainstay of future developments with application to these and other new challenges. One such challenge is that of studying intrinsically disordered and mobile regions in proteins, which are functionally important and may adopt a well-defined structure in the appropriate context, e.g. when interacting with a partner protein. About a third of the human proteome is predicted to contain intrinsically disordered regions and insights provided by the advances in prediction methods will advance their study. The success of deep-learning technology has also inspired applications to related problems, e.g. prediction of the 3D structure of RNA molecules.

Having more methods to predict structures from sequence: experimentally determined structure models from the PDB and predicted models generated by a variety of methods will be freely available to the entire life-sciences community through a number of public resources (SWISS-MODEL, AlphaFold DB, Genome3D, etc.). 3D-Beacons is a portal for registering and locating such models, created by a consortium that has developed standards for accessing structure-model data using a distributed architecture (i.e., no single site hosts all the data). This makes it possible to programmatically access 3D models using common, standardised application programming interfaces (APIs), further facilitating and stimulating the use of these models.

AlphaFold DB will soon provide predicted models for all reference sequences from UniRef90 clusters, where every sequence has at least 90% sequence identity to the other members of its cluster. The importance of community-driven initiatives such as CASP in driving research efforts and the development of tools is emphasised by this development. Other such efforts (e.g., CAPRI for the assessment of structure-prediction methods for complexes) continue in this role and make computational structural biology a rich research area. The AlphaFold breakthrough will probably result in refocusing of this community effort, away from predicting individual protein structures to problems that remain challenging today, such as prediction of the structure of multidomain proteins and complexes, evaluation of prediction accuracy and quality metrics.

Opportunities for Structural Bioinformatics research

The availability of predicted 3D models on an unprecedented scale provides a veritable cornucopia of data to be exploited, analysed and mined by structural bioinformaticians.

Boosting development of tools for scientific discovery: it is envisaged that new methods will be developed for analysis of these structure models at scale (tens of millions rather than tens of thousands of models), e.g. comparing AlphaFold models to known experimental structures, looking for as yet experimentally unobserved folds, evolutionary analysis of domains, detection of instances of apparent convergent evolution in active sites, etc. AlphaFold DB will facilitate research on evolution of (multi-domain) protein structures and on the relation of structure and function, providing clues on engineering new functions and accelerating synthetic biology applications. The large-scale structure data will also facilitate research on computational approaches to predict the effects of sequence variation and ligand binding as well as to analyse conformational states and dynamics of protein structures.

Developing new tools for structure visualisation and interpretation: as the AlphaFold models become available, their users (who may not have specialist structural biology expertise) will need to be trained in how to critically assess and use these structures and to understand the limitations of any interpretation made using the predicted models. The need to represent and visualise structures, their dynamics and interactions, and to allow molecular biologists to understand their reliability and importance, will be a major challenge to structural bioinformaticians who will need to develop software tools with appropriate user-interfaces.

Improving tools for function prediction: structural bioinformaticians have devoted time to develop tools to exploit the limited number of experimentally determined structures in the PDB to annotate genomes (e.g., domain assignment), which can help to suggest functional assignments for proteins of unknown function. The performance of such tools can now be improved by using much bigger training sets including many proteins with known function but previously unknown structure.

Opportunities for the wider life-science community

When a high-quality 3D model for a protein becomes available, previous experimental observations can often be interpreted and new testable hypotheses formulated in light of the structure model, perhaps providing insights into why a mutation is deleterious, or how a protein interacts with another protein, etc.

In drug discovery, the use of 3D models can help understand why a certain drug compound is an inhibitor but not a related compound, or why certain proteins are “druggable” while others are not. The models will accelerate research efforts to identify new candidate drugs and even drug targets as a predicted model will now often be available to start looking for druggable sites, to kick-start structure determination, to design targeted ligand or fragment screens, etc. Virtual screening techniques may also suggest new uses for old drugs against targets for which no structure was previously available.

Outlook

The availability of AlphaFold DB, a huge resource of protein structures, can perhaps be compared to the release, two decades ago, of the entire human genome sequence in the public domain, which has resulted in substantial advances in biomedical research including in new unforeseen directions. The availability of structure models for most individual proteins similarly constitutes a step change in biology with a potentially massive impact. The models will provide new insights and understanding of fundamental processes related to health and disease, with applications in biotechnology, medicine, agriculture, food science and bioengineering. It will probably take one or two decades until the full impact of this development (scientific, medical and economic) can be properly assessed. As always with step changes in science or technology, some of the current scientific activities will need to evolve, but there will be an abundance of new and exciting opportunities, applications and spin offs, many of which we cannot even foresee today. Structural biology, and biology in general, will never be the same again and we can’t wait to see the impact of these new developments - it will be an exhilarating experience!