Interactomes in the era of deep learning

Joana Pereira and Torsten Schwede

Biozentrum, University of Basel, Basel, Switzerland.

SIB Swiss Institute of Bioinformatics, Biozentrum, University of Basel, Basel, Switzerland.

Science (2021) **374**, 1319-1320

DOI: 10.1126/science.abm8295

https://www.science.org/doi/10.1126/science.abm8295

Characterizing macromolecular interactions allows better understanding of the inner workings of a cell. However, all methods available today have limitations: Some tell us whether two macro molecules interact, others provide atomic detail about the interaction partners or, at best, the structures of isolated assemblies without cellular context. On page 1340 of this issue, Humphreys et al. (1) describe a new computational approach, founded on the ongoing deep learning revolution in structure bioinformatics (2, 3), to predict the composition and model the three-dimensional (3D) structure of protein-protein interactions at the same time. They apply their approach to a eukaryotic system—baker's yeast (Saccharomyces cerevisiae)—and predict and accurately model more than 1500 protein-protein interactions, 106 of which were not seen before, paving the way to high-throughput, high-accuracy modeling of entire cells.

Determining the 3D structures of macromolecules and their interactions provides important information about macromolecular mechanisms, which can be used, for example, in drug development or exploited in biotechnology. Experimental structural biology methods such as macromolecular crystallography (MX) and high-resolution cryo–electron microscopy (cryo-EM) provide atomic-level detail of macromolecular structures and their assemblies (4). Such experiments are laborious and require purification of the macromolecules from their cellular context. Although techniques such as yeast two-hybrid (Y2H) and cross-linking mass spectrometry (XL-MS) allow for large-scale detection of interaction partners, methods such as site-directed mutagenesis or Förster energy resonance transfer (FRET) experiments characterize individual interactions and interfaces. This information can be used to guide the modeling of assemblies by, for example, macromolecular docking in integrative (also known as hybrid) approaches that combine a variety of data types from low-resolution experiments with computational modeling to generate 3D representations of macromolecular assemblies (5).

In recent years, structural biology has seen its horizons drastically expanded by computational techniques for structure prediction (see the figure), fueled by the evolution of machine learning algorithms (6) as well as a rapid increase of experimental information in open databases such as the Protein Data Bank, which celebrates its 50th anniversary this year. The Critical Assessment of Structure Prediction (CASP) experiment has, since 1994, provided a platform for testing protein structure prediction methods and, during its history, has lived through (and stimulated) several revolutions (7). For example, the development of sensitive methods for the detection of remote homologous relationships boosted homology-based modeling, and the use of coevolution information further improved the modeling of proteins without homologs of known structure. This latter method is based on the idea that residues close in space are evolutionarily coupled, and that coupling signals extracted from multiple sequence alignments can be used to predict close contacts in 3D. This not only proved to be useful for the prediction of protein 3D structures, but also readily expanded to the realm of

intermolecular interactions, acting as a fast and accurate method to screen and predict protein-interacting pairs in, for example, the proteome of a bacterium (Escherichia coli) (8, 9).

This year, a new breakthrough occurred and a new era in structural bioinformatics started (2, 3): DeepMind's AlphaFold2 algorithm (6) became the first computational method to reach close-to-experimental atomic accuracy for individual protein structures in CASP (10). The basis of this success was the combined use of state-of-the-art deep learning methods with massive amounts of computing power and the vast structure and sequence data accumulated over the past five decades. This promoted a quick and intense activity in the community, with RoseTTAFold rising shortly as a close academic competitor of AlphaFold2 (11). Both methods make use of state-of-the-art deep learning approaches but differ in their core architecture. Still, an important part of both is the use of evolutionary couplings from multiple sequence alignments, which are efficiently handled within their underlying networks to predict interatomic contacts and accurately compute 3D coordinates for the atoms in a target protein from its amino acid sequence. Given the previous success of such signals for the identification of protein-protein interactions (8, 9), it makes sense to explore such methods to improve the prediction and modeling of protein-protein interactions and their assemblies at the atomic level.

Although most efforts focused on adapting the AlphaFold2 and RoseTTAFold workflows to model protein complexes of known composition and stoichiometry (12), Humphreys et al. combined the speed of RoseTTAFold's contact prediction algorithm with the high accuracy of AlphaFold2's folding engine and suggest a new method to accurately predict and model at the same time protein pairs across the baker's yeast proteome, the first eukaryote to have its interactome modeled in such a high-throughput fashion. Scanning through ∼8 million putative protein pairs, Humphreys et al. predicted those more likely to interact on the basis of strong coevolutionary signals and replaced macromolecular docking by protein structure prediction of the joint pair to model the 3D structure of the assembly. The method was able to accurately predict the composition and model the structure of more than 1500 interacting pairs spanning almost all key eukaryotic cell processes, including 106 undescribed assemblies that may highlight previously unknown processes, as well as more than 600 previously known interacting pairs (according to low-resolution biophysical data).

The work by Humphreys et al. is a step closer to the modeling of entire cells at high resolution and has already inspired further studies into the interactome of the human mitochondrion (13). Currently, methods such as MX and electron microscopy (EM) provide high-resolution atomic representations of macromolecular machines in isolation. Cellular cryo–electron tomography (cryo-ET) has the potential to provide a detailed snapshot of the network of macromolecular interactions, but so far only subnanometer resolution can be obtained (14). Artificial intelligence (AI)–based highly accurate proteome-wide modeling of interactions may be able to compensate that resolution gap in a timely manner, especially for more complex organisms. Notwithstanding, methods such as AlphaFold2 and RoseTTAFold provide a static model; incorporating the transient and dynamic nature of macromolecular assemblies will need to be addressed in the future.

This work also highlights the success of open science and community-based method development. AlphaFold2, developed by a commercial company, was made openly available to the entire community, including its source code. This promoted the quick development of different AI-based bioinformatic methods for various goals, such as the Humphreys et al. study. AI-based methods are clearly promoting a shift in the way life sciences research will be carried out in the future, where 3D computational models will routinely inspire new experimentally testable hypotheses.

Acknowledgments

We thank G. Studer, J. Durairaj, and X. Robin for helpful discussions.

References

1 I. R. Humphreys et al., Science 374, eabm4805 (2021).

2 A. N. Lupas et al., Biochem. J. 478, 1885 (2021).

3 S. M. Kandathil, J. G. Greener, D. T. Jones, Proteins 87, 1179 (2019).

4 T. Nakane et al., Nature 587, 152 (2020).

5 A. Sali, J. Biol. Chem. 296, 100743 (2021).

6 J. Jumper et al., Nature 596, 583 (2021).

7 A. Kryshtafovych et al., Proteins 89, 1607 (2021).

8 Q. Cong et al., Science 365, 185 (2019).

9 A. G. Green et al., Nat. Commun. 12, 1396 (2021).

10 J. Pereira et al., Proteins 89, 1687 (2021).

11 M. Baek et al., Science 373, 871 (2021).

12 R. Evans et al., bioRxiv2021).

13 J. Pei et al., bioRxiv2021).

14 M. Turk, W. Baumeister, FEBS Lett. 594, 3243 (2020).