enzymatic activity, protein-protein interactions, or molecular machines. The ever-growing database of protein structures, primarily driven by high-throughput structural genomics efforts, further facilitated dramatic improvements in protein-structure prediction based in machine learning (e.g., AlphaFold).

Therefore, an increase in cryo-EM fibril structures and new approaches to analyze these structures will allow engineering efforts to control amyloid assembly in a regulated way of protein sequences that can uniquely adopt specific conformations and also design *de novo* fibril structures. Finally, structural analyses of these fibrillar conformations may offer new avenues to design ligands or biologics that can detect the conformation of these assemblies early in disease and therapeutically prevent pathologic folding of these proteins.

ACKNOWLEDGMENTS

The author thanks Vishruth Mullapudi for helpful comments. L.A.J. is supported by a Chan Zuckerberg Initiative (CZI) Collaborative Science Award (2018-191983) and a grant from the NIH/NIA (1RF1AG065407).

DECLARATION OF INTERESTS

The author declares no competing interests.

REFERENCES

Darwich, N.F., Phan, J.M., Kim, B., Suh, E., Papatriantafyllou, J.D., Changolkar, L., Nguyen, A.T., O'Rourke, C.M., He, Z., Porta, S., et al. (2020). Autosomal dominant VCP hypomorph mutation impairs disaggregation of PHF-tau. Science *370*, eaay8826. https://doi.org/10.1126/science.aay8826.

Eisenberg, D.S., and Sawaya, M.R. (2017). Structural studies of amyloid proteins at the molecular Level. Annu. Rev. Biochem. *86*, 69–95. https://doi.org/10.1146/annurev-biochem-061516-045104.

Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J., and Serrano, L. (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat. Biotechnol. *22*, 1302–1306. https://doi.org/10.1038/nbt1012.

Lövestam, S., Koh, F.A., van Knippenberg, B., Kotecha, A., Murzin, A.G., Goedert, M., and Scheres, S.H.W. (2022). Assembly of recombinant tau into filaments identical to those of Alzheimer's disease and chronic traumatic encephalopathy. Elife *11*, e76494. https://doi.org/10.7554/elife.76494.

Prusiner, S.B. (1998). Prions. Proc. Natl. Acad. Sci. USA *95*, 13363–13383. https://doi.org/10.1073/pnas.95.23.13363.

Shi, Y., Zhang, W., Yang, Y., Murzin, A.G., Falcon, B., Kotecha, A., van Beers, M., Tarutani, A., Kametani, F., Garringer, H.J., et al. (2021). Structure-based classification of tauopathies. Nature *598*, 359–363. https://doi.org/10.1038/s41586-021-03911-7.

Thompson, M.J., Sievers, S.A., Karanicolas, J., Ivanova, M.I., Baker, D., and Eisenberg, D. (2006). The 3D profile method for identifying fibril-forming segments of proteins. Proc. Natl. Acad. Sci. USA *103*, 4074–4078. https://doi.org/10.1073/pnas.0511295103.

Tuite, M.F., and Lindquist, S.L. (1996). Maintenance and inheritance of yeast prions. Trends Genet. *12*, 467–471. https://doi.org/10.1016/0168-9525(96)10045-7.

van der Kant, R., Louros, N., Schymkowitz, J., and Rousseau, F. (2022). Thermodynamic analysis of amyloid fibril structures reveals a common framework for stability in amyloid polymorphs. Structure *30*, 1178–1189. https://doi.org/10.1016/j.str.2022.05.002.

Vaquer-Alicea, J., Diamond, M.I., and Joachimiak, L.A. (2021). Tau strains shape disease. Acta Neuropathol. *142*, 57–71. https://doi.org/10.1007/s00401-021-02301-7.

Homologues not needed: Structure prediction from a protein language model

Nir Ben-Tal[1,*] and Rachel Kolodny[2,*]
[1]Department of Biochemistry and Molecular Biology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel
[2]Department of Computer Science, University of Haifa, Mount Carmel, Haifa, 3498838, Israel
*Correspondence: bental@tauex.tau.ac.il (N.B.-T.), trachel@cs.haifa.ac.il (R.K.)
https://doi.org/10.1016/j.str.2022.07.002

Accurate protein structure predictors use clusters of homologues, which disregard sequence specific effects. In this issue of *Structure*, Weißenow and colleagues report a deep learning-based tool, EMBER2, that efficiently predicts the distances in a protein structure from its amino acid sequence only. This approach should enable the analysis of mutation effects.

Weißenow and coworkers introduce EMBER2, a multiple sequence alignment (MSA)-free method for predicting internal distances within a protein structure from its amino acid sequence alone. EMBER2 uses a protein language model to predict the internal distances of a structure and relies on external tools to derive the protein coordinates from these distances. MSA-free protein structure prediction methods more directly address the long-standing challenge of understanding the relationship between protein sequence and structure, which has many practical implications. Anfinsen's experiment in the 1960s showed that all the information needed for the three dimensional structure of a protein lies in its amino acid sequence (Kessel and Ben-Tal, 2018) and motivated a worldwide effort to computationally predict structures from sequences (Dill et al., 2008). Originally, the goal was to predict the structure from first principles, i.e., relying on a physicochemical understanding of intra-protein interactions and solvent and other environmental effects (Kessel and Ben-Tal, 2018). However, over time it became clear that this is far too challenging. The biannual rounds of Critical Assessment of Structure Prediction

(CASP) evaluations, where participants predicted protein structures that were publicly released only after each round, demonstrated that even the most sophisticated first-principle approaches failed to consistently deliver accurate predictions (Moult et al., 1995).

On the other hand, the CASP evaluations repeatedly demonstrated the superiority of statistics and knowledge-based approaches. Advances in structure prediction accuracy were because of the enormous power of machine learning methods that harvest the fast-accumulating protein sequence and structural data (AlQuraishi, 2021; Pearce and Zhang, 2021). A significant improvement in accuracy was then made by including co-evolution data, i.e., coupling between residues substitutions in different positions along the sequence across homologous proteins. The couplings could be indicative of direct contacts, and enough of these, especially between sequentially remote amino acids, would allow accurate structure inference. While the idea of using co-evolution has been around for years, it became usable only about a decade ago, when large enough sequence databases and mathematical models to infer direct contacts became available (De Juan et al., 2013; Marks et al., 2011; Morcos et al., 2011).

A major breakthrough happened about two years ago, with the development of AlphaFold—a deep learning structure predictor (Jumper et al., 2021). Starting from a query protein sequence, the MSA of its homologues, and relevant structural templates, AlphaFold consistently provides model structures with an accuracy that closely resembles that of experimental methods such as cryoEM and X-ray crystallography. Key for achieving such high accuracy was an end-to-end learning algorithm, with a loss function based on the difference between the predicted versus true coordinates of the protein. AlphaFold effectively solved the structure prediction problem, albeit only for an ''average'' over clusters of homologous proteins.

Indeed, although AlphaFold was a revolutionary breakthrough for protein structure prediction, it has shortcomings because it depends on multiple sequence alignment (MSA) of homologous proteins. First, many homologues are not always sufficiently available e.g., for orphan proteins [or proteins from small families], for *de novo* designed proteins, and for evolutionary young proteins. Furthermore, an MSA-based prediction that reflects the evolutionary signal of viable proteins in different species is not particularly useful for predicting possibly deleterious effects of mutations because the effect of a single mutation will be ''washed out'' in an MSA. However, these effects are very important in the context of human health, as single mutations may lead to protein instabilities and malfunctions that cause human diseases.

Rather than using an MSA, predictors can use protein language models (pLMs)—computational models inspired by machine learning tools developed in natural language processing (NLP). Embeddings of natural languages are calculated from large text corpora and model semantic relationships between words, so that words with similar meaning are embedded as vectors that are near each other. Similarly, pLMs calculate embeddings for amino acids within the context of the protein chain from large datasets of protein sequences. These embeddings can be the exclusive input to a second step of a supervised training, which, in EMBER2, is aimed at predicting internal distances. Weiβenow et al. (2022) further show that using merely 120 attention heads from the pre-trained pLM leads to a performance comparable with far larger models (Weiβenow et al., 2022). However, using only 120 attention heads substantially reduces the computational burden. Indeed, the authors predicted the structures of the whole human proteome within a week on a single machine, and for a protein set, they predicted all point mutations and showed that their predictions correlate with deep mutational scans data.

MSA-free structure prediction has many advantages beyond computational efficiency and brings structure prediction back to its single sequence roots (AlQuraishi, 2021). There is a clear advantage in terms of efficiency once the pLM is trained. However, as Sergey Ovchinnikov remarked in a recent twitter discussion (https://twitter.com/sokrypton/status/1538570849401425929) over another soon-to-be published MSA-free structure prediction method called OmegaFold, from the Peng Lab, the computational costs should include those for refining and retraining pLMs as sequence databases grow. Additionally, AlphaFold showed that there is much to be gained from an end-to-end learning model, so future improvements of EMBER2 may be coming. Once we have a model that can accurately predict the structures of not only a protein sequence, but also all its variants with site specific amino acid substitutions, we could differentiate between disease-causing versus benign mutations. This has major implications for personalized medicine, where minor variations in protein sequence between individuals appear to determine the outcome of treatments. Such a model should also be useful for protein design.

Protein function critically depends on conformational changes. Thus, for example, enzymes and receptors alternate between active and inactive conformations, membrane transporters shift between outward- and inward-facing conformations, and channels between gate-open and closed conformations. To really understand function, we need to predict the structures of all physiologically relevant conformations of the protein, rather than only one, and also interactions with ligands, nucleic acids, and other proteins. The group of Burkhard Rost and other top labs are working on these problems, so let's stay tuned.

ACKNOWLEDGMENTS

We acknowledge the support of grants 450/16 and 1764/21 of the Israeli Science Foundation (ISF). R.K is supported in part by the DSRC in the University of Haifa. N.B.-T.'s research is supported in part by the Abraham E. Kazan Chair in Structural Biology, Tel Aviv University.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

AlQuraishi, M. (2021). Machine learning in protein structure prediction. Curr. Opin. Chem. Biol. *65*, 1–8. https://doi.org/10.1016/j.cbpa.2021.04.005.

De Juan, D., Pazos, F., and Valencia, A. (2013). Emerging methods in protein co-evolution. Nat. Rev. Genet. *14*, 249–261. https://doi.org/10.1038/nrg3414.

Dill, K.A., Ozkan, S.B., Shell, M.S., and Weikl, T.R. (2008). The protein Folding problem *37*, 289–316. https://doi.org/10.1146/annurev.biophys.37.092707.153558.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Kessel, A., and Ben-Tal, N. (2018). Introduction to Proteins: Structure, Function, and Motion, Second Edition (CRC Press).

Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. PLoS One *6*, e28766. https://doi.org/10.1371/journal.pone.0028766.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families *108*, E1293–E1301. https://doi.org/10.1073/pnas.1111471108.

Moult, J., Pedersen, J.T., Judson, R., and Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. Proteins *23*. ii–iv. https://doi.org/10.1002/prot.340230303.

Pearce, R., and Zhang, Y. (2021). Deep learning techniques have significantly impacted protein structure prediction and protein design. Curr. Opin. Struct. Biol. *68*, 194–207. https://doi.org/10.1016/j.sbi.2021.01.007.

Weißenow, K., Heinzinger, M., and Rost, B. (2022). Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction. Structure *30*, 1169–1177.