ARCHIVES OF BIOINFORMATICS AND COMPUTATIONAL BIOLOGY
Jan-Jun 2024, VOL. 1, ISSUE 1, pp. 8-9
https://doi.org/10.61577/abcb.2024.100003

RESEAPRO JOURNALS

COMMENTARY

OPEN ACCESS

AlphaFold: the silver lining of the 'protein folding problem'

Prerana Mordina
Department of Zoology, Ravenshaw University, Odisha, India

ARTICLE HISTORY
Received 3 July 2023; Revised 25 September 2023; Accepted 3 October 2023

Introduction

Proteins, the building blocks of life, play vital roles across all domains of life. These versatile biomolecules translate the information encoded in genes to conceive a vast array of discrete structures and functions that regulate life. Proteins translated from the information encoded in an organism's DNA typically emerge as a single thread of amino acids from the manufacturing entities, ribosomes. The amino acid sequence interlinked by peptide bonds determines the protein's native structure, which further dictates its function. Altering the protein's structure *via* its sequence can produce unprecedented implications, from rendering it toxic to completely incapacitating its functional activity. Comprehending the 3-dimensional structure enables a mechanistic understanding of its function and how it can be modified as required. For example, proteins can be engineered to catalyze specific reactions or selected as drug targets to address an epidemic.

With pioneering techniques like next-generation sequencing being employed to unravel the genomes and proteomes of organisms, the wealth of information on gene and protein sequences has been accruing expeditiously. Though the Protein Data Bank (PDB) boasts ~200,000 experimentally derived protein structures, it represents only a fraction of the billions of already deciphered protein sequences. Experimentally determining the 3-dimensional structures of proteins is often long and arduous, creating a bottleneck in the structural coverage of known proteins. Why, one might ask then, do we not try to determine the structure using the sequence since each unique sequence translates to a unique structure? Computational biologists and bioinformaticians have been trying to answer the same for decades, using alternative techniques like homology modeling to determine the structure of unknown proteins with sequences similar to an experimentally determined protein structure. Various computational tools, such as Robetta, Phyre2, SWISS-MODEL, and I-TASSER, use homology modeling, multiple alignments, and iterative simulations to predict the structure of proteins whose structures have not been determined experimentally.

However, these prediction models have several drawbacks, the most significant being their inability to predict the structure of novel proteins with no homology to any known structures. The complexities involved in protein folding prediction can be attributed to two major problems-

The energy function problem: High resource-intensive quantum mechanics simulations are required to calculate the exact electrostatic potential and bond energies of proteins to correctly predict the structure since even a small error in one single conformational energy may eventually lead to a completely different fold prediction.

The sampling problem- The second problem arises from the first one in that each bond and energy potential must be taken into account for an accurate protein structure prediction, creating a colossal sample space. For example, simulating an ideal protein of only 50 amino acids, with constant bond lengths, each amino acid having two rotatable backbone bonds, and considering only 10-degree increments in each, 1072 potential conformations would be obtained, each of which would need to be sampled to find the lowest energy state or the native state of the protein [1].

Though prediction techniques such as molecular dynamic simulations, fragment assembly, and machine learning perpetually upgraded their features of template-based modeling (TBM) or free modeling (FM), they still fell short of predicting structures to experimental accuracy [2]. Integrating Artificial Intelligence (AI) and Deep Learning (DL) to the problem of protein folding has proved a landmark step toward protein structure prediction. DeepMind's AI-driven protein structure prediction model caused huge ripples when they aced the highly coveted and challenging Critical Assessment of Structure Prediction (CASP) in 2020. The 'protein folding problem' that had eluded scientists for more than half a century was scrupulously addressed by the AlphaFold neural network [3].

The biennial blind test CASP allows computational biologists to evaluate their potential methods by predicting the structure of experimentally derived protein structures that had not yet been publicly released. The AlphaFold method produced outstanding results in the 14th CASP assessment, with a huge disparity in the results produced by it and other competing methods. The median accuracy of the protein's backbone predicted by AlphaFold was 0.96 Å r.m.s.d compared to 2.8 Å r.m.s.d achieved by the second-best prediction method. In addition to this excellent feat, the predicted AlphaFold structures also produced highly precise side-chain conformations with better accuracy than template-based prediction methods using strong templates. The AlphaFold method can also be scaled to very large, unique proteins for reliable prediction of domains and domain-packing. The most unique feature of this prediction method is its precise, per-residue reliability prediction.

The enhanced accuracy of the AlphaFold network is majorly owing to the incorporation of novel neural network architectures and training methods derived from the physical, geometric, and evolutionary constraints of the protein structures. The novel architecture employs several distinct protein sequence databases to construct multiple sequence alignments (MSAs) and a pair representation that forms the initial representation of the targeted structure. The evoformer

***Correspondence:** Prerana Mordina, Department of Zoology, Ravenshaw University, Odisha, India, e-mail: preranamordina1993@gmail.com

neural network module extracts and assesses the MSA and templates through back-and-forth information exchange throughout the network. The structure neural network module prioritizes the protein backbone's orientation, taking into account the rotations and translations of the residues and localizing each side chain of each residue into highly constrained frames, and finally, enforces local refinement and energy minimization through gradient descent.

The pioneering machine-learning method has been implemented in various ambitious projects, including the structure prediction of the human proteome. Following decades of effort through experimental structure determination, only 17% of the total known human proteome could be deciphered. Utilizing the AlphaFold network for structure prediction allowed the coverage of almost the entire human proteome and resulted in confident prediction for 58% of residues, with high confidence in 36% [4].

Despite the breakthroughs in the precision of protein structure prediction, AlphaFold presents certain limitations. The algorithm has been shown to perform poorly in cases of intrinsically disordered regions or proteins, which often overlap the regions of low accuracy in AlphaFold predictions [5]. The algorithm also tends to present loops as secondary protein conformations, mostly as alpha helices, with only short loops of<20 amino acids being predicted with high accuracy [6]. The algorithm has also been shown to predict either the apo-form or the holo-form of a protein. Increased conformational changes between the apo- and holo-forms of a protein decreased the efficiency of AlphaFold in correctly predicting the protein's structure [7]. A similar trend was also observed in AlphaFold's predictions of mutated and native protein structures, with only very slight differences of less than 1Å r.m.s.d. between the backbones of the two [8].

The success of AlphaFold has laid the cornerstone for embedding DL in addressing other challenges across biosciences, such as protein function prediction, phylogenetic inference, genome engineering, systems biology, and data integration [9]. Though the application of AI and DL to biosciences is still in its inception, it has made an enduring impact on one of the most poignant questions of biology- 'the protein folding problem.' Further advances are still required to provide the final sheen to this algorithm, allowing it to predict proteins to atomic accuracy.

Disclosure statement

No potential conflict of interest was reported by the author.

References

1. Marcu ŞB, Tăbîrcă S, Tangney M. An Overview of Alphafold's Breakthrough. Front Artif Intell. 2022;5:875587. https://doi.org/10.3389/frai.2022.875587
2. Bertoline LMF, Lima AN, Krieger JE, Teixeira SK. Before and after AlphaFold2: An overview of protein structure prediction. Front Bioinform. 2023;3:1120370. https://doi.org/10.3389/fbinf.2023.1120370
3. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-589. https://doi.org/10.1038/s41586-021-03819-2
4. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature. 2021;596(7873):590-596. https://doi.org/10.1038/s41586-021-03828-1
5. Ruff KM, Pappu RV. AlphaFold and Implications for Intrinsically Disordered Proteins. J Mol Biol. 2021;433(20):167208. https://doi.org/10.1016/j.jmb.2021.167208
6. Stevens AO, He Y. Benchmarking the Accuracy of AlphaFold 2 in Loop Structure Prediction. Biomolecules. 2022;12(7):985. https://doi.org/10.3390/biom12070985
7. Saldaño T, Escobedo N, Marchetti J, Zea DJ, Mac Donagh J, Velez Rueda AJ, et al. Impact of protein conformational diversity on AlphaFold predictions. Bioinformatics. 2022;38(10):2742-2748. https://doi.org/10.1093/bioinformatics/btac202
8. Buel GR, Walters KJ. Can AlphaFold2 predict the impact of missense mutations on structure? Nat Struct Mol Biol. 2022;29(1):1-2. https://doi.org/10.1038/s41594-021-00714-2
9. Sapoval N, Aghazadeh A, Nute MG, Antunes DA, Balaji A, Baraniuk R, et al. Current progress and open challenges for applying deep learning across the biosciences. Nat Commun. 2022;13(1):1728. https://doi.org/10.1038/s41467-022-29268-7

https://doi.org/10.61577/abcb.2024.100003