Received: 20 April 2024 | Accepted: 26 April 2024

DOI: 10.1002/ctm2.1689

CLINICAL AND TRANSLATIONAL MEDICINE Open Access

WILEY

COMMENTARY

One step forward towards deep-learning protein complex structure prediction by precise multiple sequence alignment construction

Wei Zheng[1] | **Qiqige Wuyun**[2] | **Yang Zhang**[3,4,5]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA

[2]Department of Computer Science and Engineering, Michigan State University, East Lansing, Michigan, USA

[3]Cancer Science Institute of Singapore, National University of Singapore, Singapore, Singapore

[4]Department of Computer Science, School of Computing, National University of Singapore, Singapore, Singapore

[5]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

Correspondence

Wei Zheng, Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109, USA.
Email: zhengwei@umich.edu

Yang Zhang, Cancer Science Institute of Singapore, National University of Singapore, 117599, Singapore.
Email: zhang@zhanggroup.org

Funding information

National Institute of General Medical Sciences, Grant/Award Numbers: GM136422, S10OD026825; National Institute of Allergy and Infectious Diseases, Grant/Award Number: AI134678; National Science Foundation, Grant/Award Numbers: IIS1901191, DBI2030790, MTM2025426; National University of Singapore Start-up Grant, Grant/Award Number: A-80001129-00-00

KEYWORDS

antibody–antigen interactions, deep-learning, multiple sequence alignments, protein–protein complex structure prediction

Proteins are the 'workhorse' molecules carrying out nearly all biological functions within living organisms. Most of the functions of a protein are performed through its interactions with other proteins, known as protein–protein interactions (PPIs). Common instances include the interactions between antibodies and antigens, which bolster an organism's capacity to identify and combat external pathogens, the binding of ligand and receptor proteins that instigate cellular signalling cascades, and the interplay between enzyme and substrate proteins that facilitate metabolic processes. The important functional roles played by PPIs render them pivotal targets in numerous contemporary drug design initiatives.[1]

Significant efforts have been made to determine the three-dimensional structures of PPI complexes, which could provide a geometric and physical landscape to facilitate biological function annotation and drug discovery efforts targeting PPIs. Although structural biology techniques such as X-ray crystallography and Cryo-Electron Microscopy offer the most precise structural insights into PPIs, they frequently demand significant resources and lack scalability for resolving molecular structures across the entire proteome. The most recent success of artificial intelligence (AI) algorithms, such as AlphaFold2,[2,3] has enabled the computational prediction of protein structures with remarkable accuracy, opening avenues for obtaining high-quality structures of proteins and PPI complexes through computational means.

Most AI methods train models from large-scale experimental structures built on co-evolutionary information obtained from multiple sequence alignments (MSAs), as the latter can offer critical structural information of

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *Clinical and Translational Medicine* published by John Wiley & Sons Australia, Ltd on behalf of Shanghai Institute of Clinical Bioinformatics.

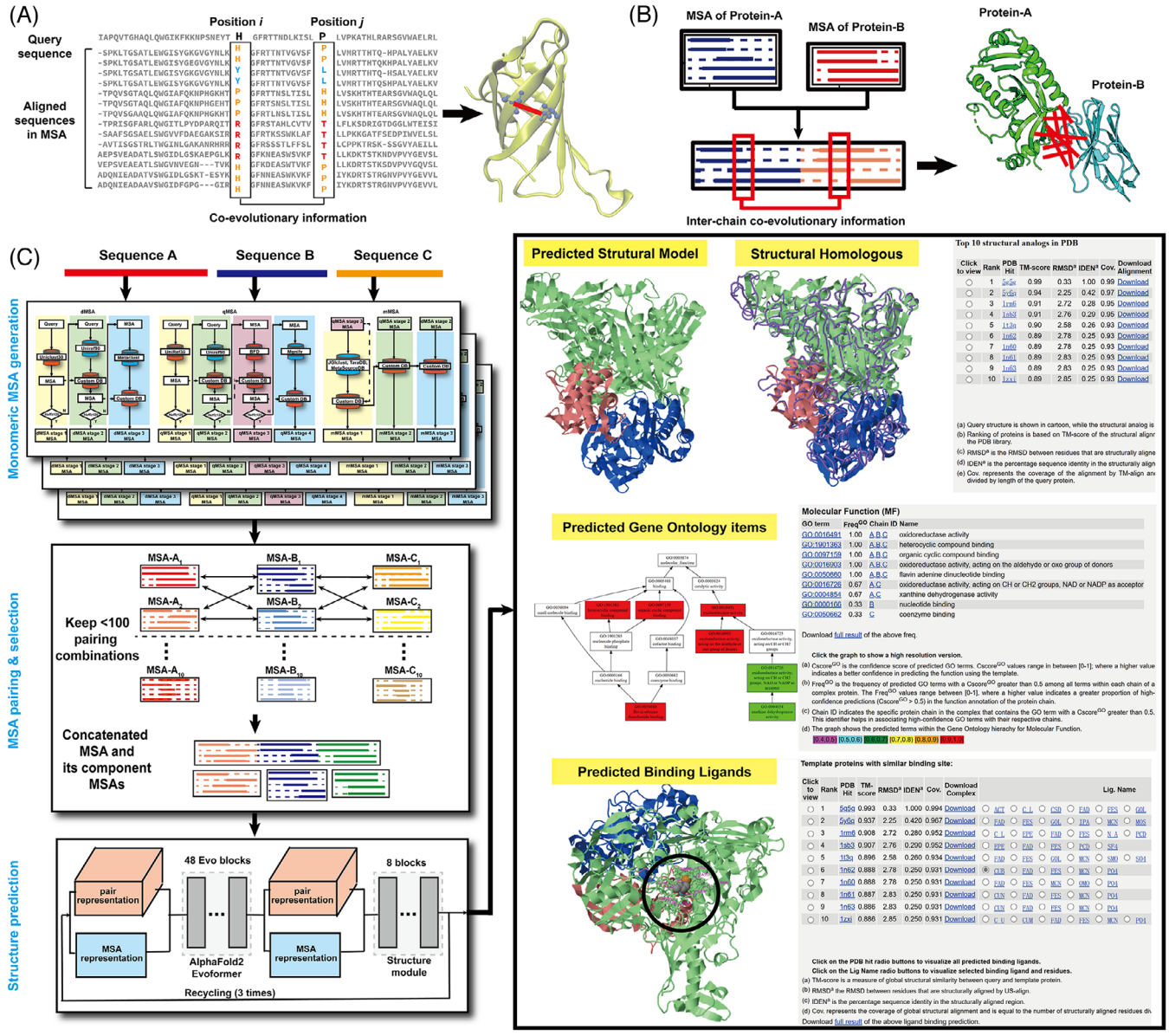


FIGURE 1 The DMFold pipeline for protein complex structure prediction. (A) Co-evolutionary information from monomer MSAs contains 3D structural insights for protein structure prediction. (B) Co-evolution from appropriately paired protein complex MSAs can be used for protein complex structure prediction. (C) DMFold pipeline for protein complex structure prediction followed by structure-based function annotations.

the target protein. Specifically, during evolution when a mutation occurs at one residue site and disrupts its interaction with other residues, the protein may become unstable, making it difficult for species with such mutation to survive. However, if the interacting residues mutate at the same time and stabilize the protein structure, the species can continue to survive. This phenomenon refers to 'co-evolution'. Since proteins in current organisms have all undergone the rigours of co-evolution over hundreds of millions of years, aligning a vast array of protein sequences in MSAs can effectively deduce information about protein spatial distances between residues (see Figure 1A). Although MSA and the co-evolutionary information have been successfully utilized by various AI-based 3D structure prediction approaches,[2,4] due to the lack of large-scale PPI sequence databases, however, constructing comprehensive MSAs and deducing reliable quaternary co-evolutionary structure insights remain a major bottle-neck problem for PPI complex structure predictions.

To tackle the challenge, we developed the DMFold[5,6] algorithm, in which we utilized a simple protocol to construct PPI MSAs and deduce inter-chain co-evolutionary information by pairing monomer-chain MSAs that can benefit from abundant sequence resources from

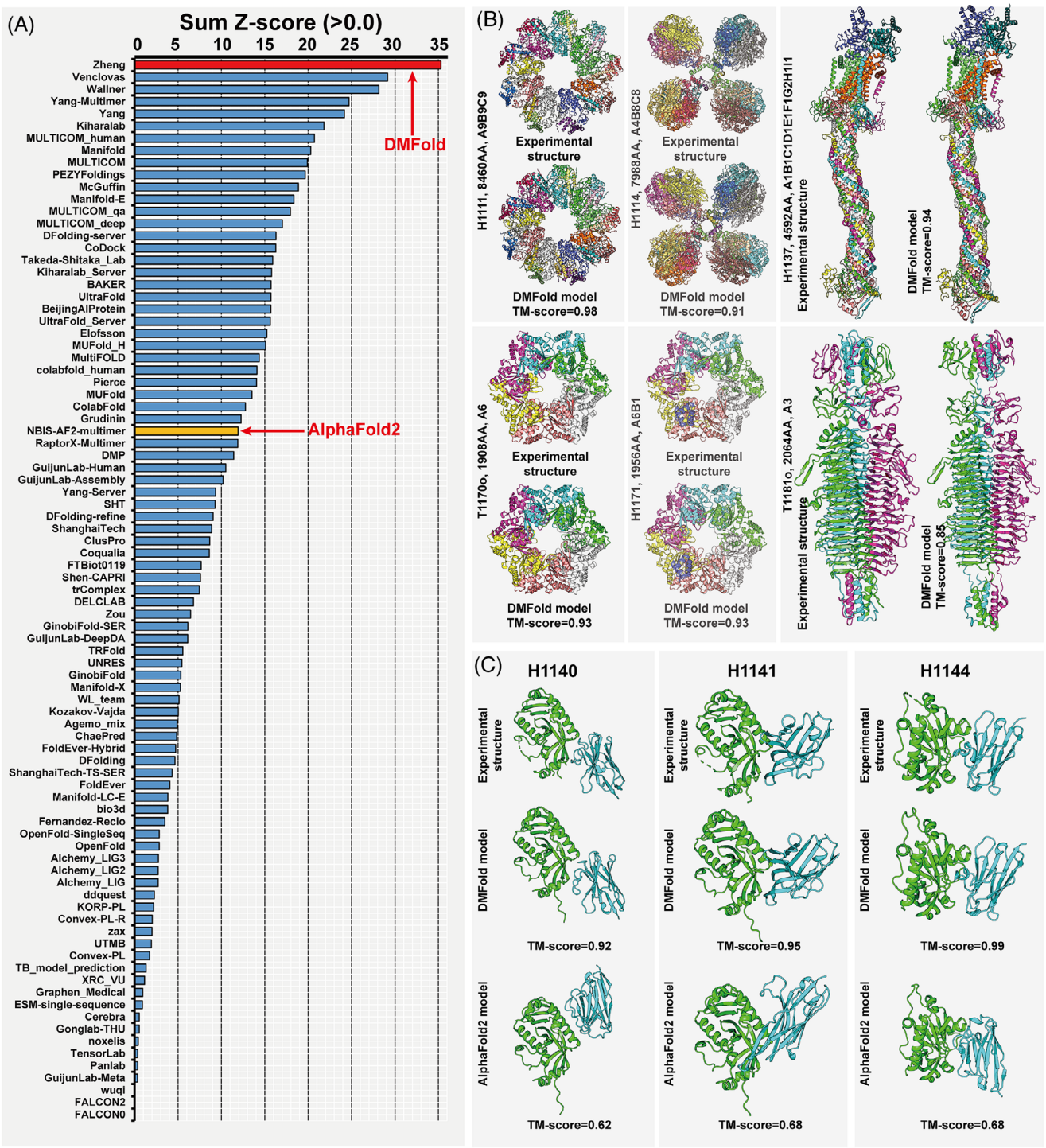


FIGURE 2 Blind test results of DMFold in CASP15 experiment. (A) Overall ranking of protein complex structure prediction methods based on 41 targets for the 87 registered CASP15 assembly groups, with data taken from the CASP15 official webpage (https://predictioncenter.org/casp15/zscores_multimer.cgi). DMFold (registered as 'Zheng') and the public March 2022 version of the AlphaFold2 server (registered as 'NBIS-AF2-multimer') are marked in red and orange, respectively. (B) DMFold models associated with experimentally solved structures for six large-size CASP15 complex targets (> 1500 residues) successfully modelled by DMFold. (C) The experimental structures and PPI models predicted by AlphaFold2 and DMFold for three CASP15 nanobody-antigen complexes (H1140, H1141 and H1144).

metagenome repositories[7] (see Figure 1B). As outlined in Figure 1C, DMFold first constructs monomer MSAs for each component chain through iterative dynamic programming and Hidden Markov Model searches against multiple metagenome sequence libraries.[8] To ensure orthologous sequence alignments, which is critical for inter-chain co-evolutionary map deduction, DMFold selectively links the homologous sequences of different chains coming from the same species, where a new AI-driven MSA scoring strategy has been introduced to rank the MSAs prior to sequence pairing. The assembled PPI MSAs are subsequently fed into an end-to-end deep-learning neural network module to generate PPI complex structure predictions. To enhance the biomedical usefulness, the online DMFold server (https://zhanggroup.org/DMFold/) creates multiple outputs from primary sequences in addition to the PPI complex structure models, including the top 10 experimental structures closest to the target and structure-based functional annotations on gene ontology, enzyme commission and binding ligands, respectively.

The Critical Assessment of protein Structure Prediction (CASP) is a community-wide blind experiment which holds biennially and aims at providing an objective benchmark of the state-of-the-art structure prediction technologies. DMFold participated (as 'Zheng') in the most recent CASP15 and ranked as the best protein complex structure prediction method by the official CASP assessors,[9] where the overall Z-score of DMFold (35.4) was 18% higher than the second best method (29.9) and 2.9 times higher than that of the standard version of AlphaFold2 (12.3) (Figure 2A).

Figure 2B lists the results of DMFold on six large-size protein complexes containing > 1500 residues, where TM-scores of the DMFold models are .98, .91, .94, .93, .93 and .85, respectively. Here, TM-score is a standard measure of structure prediction accuracy with larger values indicating higher prediction accuracy, where a TM-score > .9 generally corresponds to models with medium-to-high experimental structure accuracy.[10] Notably, the two largest targets (H1111 and H1114) contain > 7900 residues both being heteromeric complexes with stoichiometry variable of 'A9B9C9' and 'A4B8C8', respectively. These results underscore the remarkable capacity of DMFold to model large-size PPI complexes, towards the solution of a long-standing challenge faced by traditional quaternary structure modelling approaches.

A particularly exciting application of DMFold is on modelling antibody- or nanobody-antigen complexes, a type of PPI that plays pivotal roles in the defence of our body against the invasion of external pathogens. Figure 2C presents a comparison of structural models by DMFold and AlphaFold2 on three nanobody-antigen targets in CASP15, which represent three typical interaction modes of nanobodies with mouse CNPase. While AlphaFold2 failed to model the PPI orientations with TM-score below .7 for all three complexes, DMFold demonstrates exceptional prediction power with three models achieving TM-scores of .92, .95 and .99, respectively. The capability to accurately model functional PPIs such as antibody–antigen complex structures can help significantly expedite the development process of antibody-based therapeutics.

In summary, we developed DMFold, an AI-based pipeline for atomic-level PPI structure prediction. The major advancement of DMFold lies in the construction of precise multi-chain MSAs by iteratively collecting and pairing high-quality monomeric MSAs from huge metagenome sequence databases, which allows for the derivation of reliable cross-chain co-evolutionary information and, therefore, enables AI network models to derive accurate quaternary structural patterns and 3D PPI conformations. We expect that the improved power of AI-based complex structure predictions will significantly enhance the accuracy of large-scale protein function annotations and the effectiveness of structure-based drug discovery processes targeting various PPI-related human diseases.

AUTHOR CONTRIBUTIONS

Wei Zheng and Qiqige Wuyun wrote the initial version of the manuscript and prepared the figures. Yang Zhang revised and submitted the manuscript.

ACKNOWLEDGEMENTS

This work is supported in part by the National Institute of General Medical Sciences (GM136422 and S10OD026825 to Y.Z.), the National Institute of Allergy and Infectious Diseases (AI134678 to Y.Z.), the National Science Foundation (IIS1901191, DBI2030790 and MTM2025426 to Y.Z.) and the National University of Singapore Start-up Grant (A-80001129-00-00 to Y.Z.).

CONFLICT OF INTEREST STATEMENT

The authors declare no competing interests.

ETHICS STATEMENT

Not applicable.

ORCID

Yang Zhang https://orcid.org/0000-0002-2739-1916

REFERENCES

1. Wells JA, McClendon CL. Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature*. 2007;450(7172):1001-1009.
2. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583-589.

3. Evans R, O'Neill M, Pritzel A, et al. Protein complex prediction with AlphaFold-Multimer. *Biorxiv*. 2021. https://www.biorxiv.org/content/10.1101/2021.10.04.463034v2
4. Zheng W, Zhang C, Li Y, Pearce R, Bell EW, Zhang Y. Folding non-homology proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. *Cell Rep Methods*. 2021;1:100014.
5. Zheng W, Wuyun Q, Freddolino PL, Zhang Y. Integrating deep learning, threading alignments, and a multi-MSA strategy for high-quality protein monomer and complex structure prediction in CASP15. *Proteins Struct Funct Bioinf*. 2023;91(12):1684-1703.
6. Zheng W, Wuyun Q, Li Y, Zhang C, Freddolino PL, Zhang Y. Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data. *Nat Methods*. 2024;21:279-289.
7. Yang P, Zheng W, Ning K, Zhang Y. Decoding the link of microbiome niches with homologous sequences enables accurately targeted protein structure prediction. *Proc Natl Acad Sci USA*. 2021;118(49):e2110828118.
8. Zhang C, Zheng W, Mortuza SM, Li Y, Zhang Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics*. 2020;36(7):2105-2112.
9. Ozden B, Kryshtafovych A, Karaca E. The impact of AI-based modeling on the accuracy of protein assembly prediction: insights from CASP15. *Proteins Struct Funct Bioinf*. 2023;91(12):1636-1657.
10. Pearce R, Zhang Y. Deep learning techniques have significantly impacted protein structure prediction and protein design. *Curr Opin Struct Biol*. 2021;68:194-207.

How to cite this article: Zheng W, Wuyun Q, Zhang Y. One step forward towards deep-learning protein complex structure prediction by precise multiple sequence alignment construction. *Clin Transl Med*. 2024;14:e1689.
https://doi.org/10.1002/ctm2.1689