Biochemistry

pubs.acs.org/biochemistry Viewpoint

Entering an Era of Protein Structuromics

Lunjie Wu, Huan Liu, Yan Xu, and Yao Nie*

Cite This: *Biochemistry* 2023, 62, 3167−3169 Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Sequence determines the structure, and the structure in turn determines the function, are the fundamental principles of protein chemistry. In the genomics era, the paradigm of mining protein functionality and evolutionary insights through sequence analysis has led to remarkable achievements. However, protein sequences often mutate faster than their structural counterparts during evolution. For protein sets characterized by highly divergent sequences, sequence-based analysis is often inadequate, whereas direct extraction of implicit information from the structures appears to be a more effective strategy.

AI-based protein structure prediction tools provide protein structures at the metagenome scale, with the AlphaFold database (AFDB) and ESM Metagenomic Atlas database (ESMDB) containing >200 million and >600 million predicted structures, respectively. This massive number of structures enables travel into the protein structure universe, which is a new impetus for at least three areas of research.

(1) Compared with the PDB, which contains over 200 000 structures, computationally predicted structural databases harbor a substantial number of proteins that lack functional characterization. Analysis of protein structuromics information will greatly facilitate the mining of protein resources.
(2) Protein folds are a reservoir of fundamental building blocks that underlie protein architecture. The extraction of functional structure fragments from diverse protein families could inspire the modular design of proteins, thereby accelerating the exploration of the protein structure universe and de novo design of new functional proteins.
(3) The evolution models of protein functions often hold indicative values for species evolution. As previously mentioned, structures tend to be more conserved, and structure-based phylogenetic trees are beneficial for obtaining evolutionary insights from a new perspective, especially for deciphering the complete functional evolutionary pathways of proteins.

In response to the practical application mentioned above, recently released structural analysis tools, algorithms, and workflows have opened new possibilities for protein structuromics.

The first category focuses primarily on protein sets with structurally similar features. For instance, the newly reported Foldseek[1] can extract proteins with similar folds from databases on the basis of structural queries. Specifically, Foldseek supports rapid searches across various databases, such as AFDB, ESMDB, and PDB, to gather remote homologues, which is challenging for sequence-based searches. Additionally, Structome[2] provides convenient structural clustering options for direct computation of the distance matrix for custom protein sets. When combined, these tools cover a typical structural analysis workflow (Figure 1). The workflow begins with the protein structure of interest and filters structurally similar counterparts to form a protein set. Subsequently, the distance matrix of the structures within the protein set is calculated, and a phylogenetic tree is constructed. An exciting analogous case of this workflow is that Huang et al. discovered novel base editing tools by structure clustering within a protein family.[3] In particular, this structural analysis workflow is nearly identical to the sequence-based workflow, with the primary distinction being the shift in the analytical subject from sequence to structure. In other words, sequence-based analysis theories can be transplanted to structure-based analysis.

The second category of Foldseek-inspired algorithms and workflows directly targets the entire protein structure universe. The currently developed Foldseek Cluster algorithm[4] can conduct a structural alignment for all structures within the AFDB. A primary achievement of this algorithm is the identification of over 2 million non-singleton structural clusters from the AFDB (Figure 2A). Most of these clusters can be traced back to ancient times, whereas approximately one-third lacked discernible knowledge. Furthermore, this algorithm has facilitated diverse explorations within the protein structure universe, such as the discovery of potential novel enzymes (Figure 2B) and prediction of structural domain similarities and relationships between families (Figure 2C). Durairaj and colleagues[5] established a protein network comprising 50 million high-confidence predicted structures within the AFDB, which are accessible at https://uniprot3d.org/atlas/AFDB90v4. The authors also presented an integrated analytical workflow encompassing structural, sequential, and semantic

Received: October 10, 2023
Published: November 11, 2023

https://doi.org/10.1021/acs.biochem.3c00547
Biochemistry 2023, 62, 3167−3169

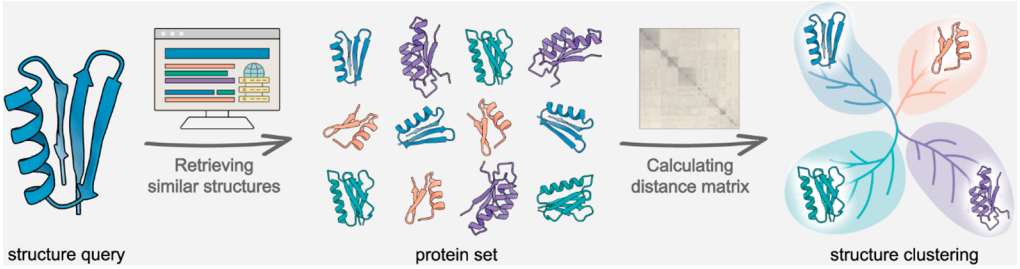

Figure 1. A basic structural analysis workflow.

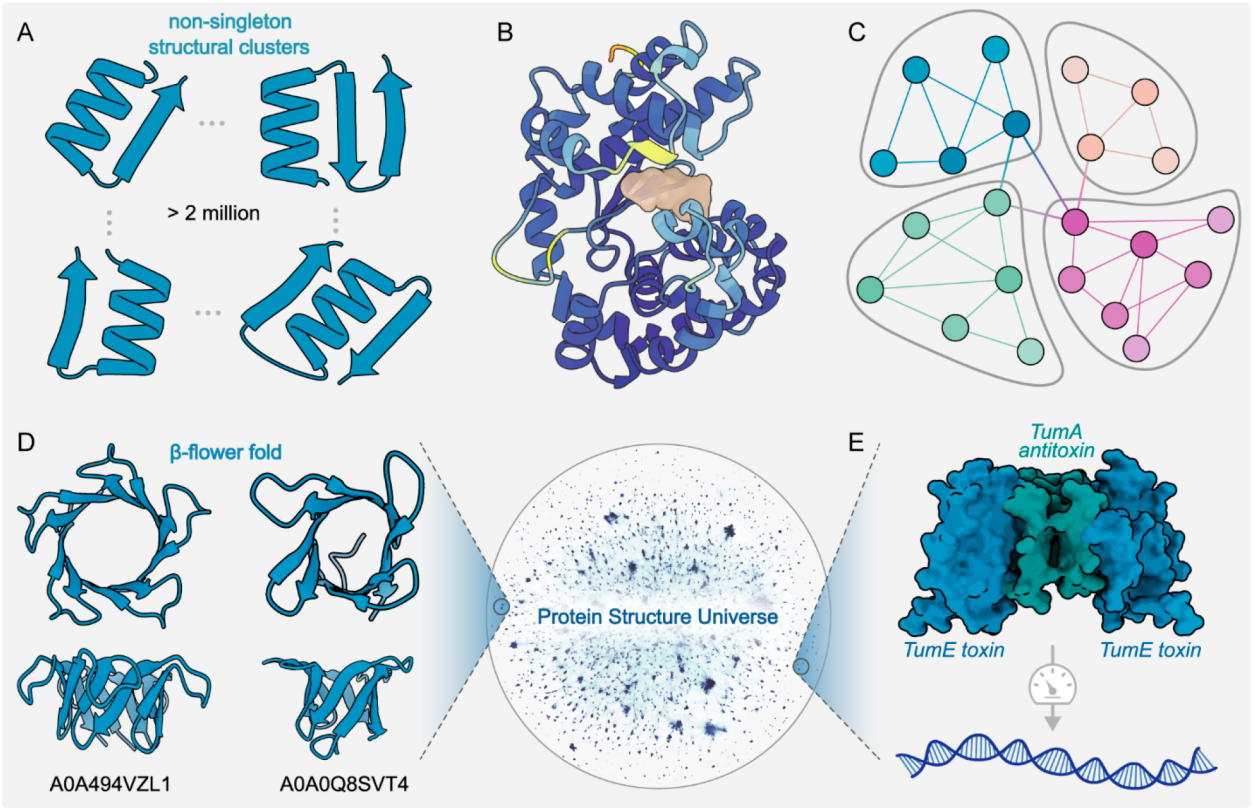

Figure 2. Diverse explorations targeting the protein universe. (A) Schematic representation of non-singleton structural clusters. (B) Structure of a potential new enzyme colored by pLDDT. (C) Diagram of structural domain similarity prediction and relationship analysis. (D) Representative structures of β-flower fold. (E) A new superfamily of toxin−antitoxin systems TumE-TumA.

aspects to dissect the dark regions within this protein network. Employing this workflow, they dug up an unprecedented structural type called β-flower fold (Figure 2D) and unearthed a novel superfamily of the translation-targeting toxin−antitoxin system, TumE-TumA (Figure 2E).

In summary, AI-based structural prediction tools have provided a rich source of structural data and recent advancements in structural analysis tools, algorithms, and workflows that have significantly enhanced our capacity to explore computationally predicted structural databases. These developments provide a promising solution for unlocking the vast knowledge concealed within the protein structure universe and suggest that we have entered a new era of protein structuromics.

AUTHOR INFORMATION

Corresponding Author

Yao Nie − *Laboratory of Brewing Microbiology and Applied Enzymology, School of Biotechnology and Key Laboratory of Industrial Biotechnology of Ministry of Education, Jiangnan University, Wuxi 214122, China; Suqian Industrial Technology Research Institute of Jiangnan University, Suqian 223814, China;* orcid.org/0000-0001-8065-7640; Email: ynie@jiangnan.edu.cn

Authors

Lunjie Wu − *Laboratory of Brewing Microbiology and Applied Enzymology, School of Biotechnology and Key Laboratory of Industrial Biotechnology of Ministry of Education, Jiangnan University, Wuxi 214122, China;* orcid.org/0000-0001-7802-6372

Huan Liu − *Laboratory of Brewing Microbiology and Applied Enzymology, School of Biotechnology and Key Laboratory of Industrial Biotechnology of Ministry of Education, Jiangnan University, Wuxi 214122, China;* orcid.org/0000-0002-2042-368X

Yan Xu − *Laboratory of Brewing Microbiology and Applied Enzymology, School of Biotechnology and Key Laboratory of*

https://doi.org/10.1021/acs.biochem.3c00547
Biochemistry 2023, 62, 3167−3169

Industrial Biotechnology of Ministry of Education, Jiangnan University, Wuxi 214122, China; orcid.org/0000-0002-7919-4762

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.biochem.3c00547

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (2021YFC2102000), the National Natural Science Foundation of China (NSFC) (22178147, 22378168), the 111 Project (111-2-06), the High-end Foreign Experts Recruitment Program (G2021144005L), the National Program for Support of Top-notch Young Professionals, Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX21-2024, KYCX22-2358), the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions, Top-notch Academic Programs Project of Jiangsu Higher Education Institutions, and the Jiangsu province "Collaborative Innovation Center for Advanced Industrial Fermentation" industry development program.

■ REFERENCES

(1) van Kempen, M.; Kim, S. S.; Tumescheit, C.; Mirdita, M.; Lee, J.; Gilchrist, C. L. M.; Söding, J.; Steinegger, M. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **2023**, DOI: 10.1038/s41587-023-01773-0.

(2) Ashar, J. M.; Chandra, S. V.; Anthony, M. P.; Jane, R. A. Structome: Exploring the structural neighbourhood of proteins. *bioRxiv*, February 21, 2023, 529083. DOI: 10.1101/2023.02.18.529083.

(3) Huang, J.; Lin, Q.; Fei, H.; He, Z.; Xu, H.; Li, Y.; Qu, K.; Han, P.; Gao, Q.; Li, B.; Liu, G.; Zhang, L.; Hu, J.; Zhang, R.; Zuo, E.; Luo, Y.; Ran, Y.; Qiu, J.-L.; Zhao, K. T.; Gao, C. Discovery of deaminase functions by structure-based protein clustering. *Cell* **2023**, *186* (15), 3182−3195.

(4) Barrio-Hernandez, I.; Yeo, J.; Jänes, J.; Mirdita, M.; Gilchrist, C. L. M.; Wein, T.; Varadi, M.; Velankar, S.; Beltrao, P.; Steinegger, M. Clustering-predicted structures at the scale of the known protein universe. *Nature* **2023**, *622*, 637.

(5) Durairaj, J.; Waterhouse, A. M.; Mets, T.; Brodiazhenko, T.; Abdullah, M.; Studer, G.; Tauriello, G.; Akdel, M.; Andreeva, A.; Bateman, A.; Tenson, T.; Hauryliuk, V.; Schwede, T.; Pereira, J. Uncovering new families and folds in the natural protein universe. *Nature* **2023**, *622*, 646.