Biochemistry
Viewpoint
Cite This: Biochemistry XXXX, XXX, XXX−XXX
pubs.acs.org/biochemistry

Harnessing Human Neural Networks for Protein Design

Po-Ssu Huang* and Kirsten A. Thompson

Department of Bioengineering, Stanford University, Shriram Center for Bioengineering and Chemical Engineering, 443 Via Ortega, Stanford, California 94305, United States

Creating proteins to fulfill a variety of functional needs has long been a goal of biochemists. This requires a thorough understanding of the relationship between the sequence of a polypeptide chain and the resulting protein structure. In recent years, the field of protein design has finally reached a stage where it is possbile to use physical and chemical principles to guild the design of novel protein structures.

The goal of designing a protein structure is to produce an amino acid sequence that can fold into a target shape. To compute the sequence, most current methods explicitly model every atom in the system (with implicit solvent) to find a configuration that satisfies all of the interactions that each residue can make in its environment. While we are not yet capabale of using these methods to design proteins with any arbitrary function, our ability to create structures that significantly differ from those observed in structural databases has reached new heights.

Protein design has become more robust with recent advances in computer processing power, design algorithms, and the decreased cost with DNA synthesis. These breakthroughs have provided the tools to run large-scale simulations, test design hypotheses, and experimentally iterate on and confirm designs. Nonetheless, the word "design" implies the involvement of cognitive activity in determining the outcome. This is arguably the most critical and least tractable element of the approach. Although any new amino acid sequence that can be generated rationally for a protein can be considered a design, in recent years, the meaning of designing a protein "de novo" has referred largely to designs in which both the structure and the sequence are modeled and created from scratch. When both the backbone and sequence are unknown at the onset, a protein designer must creatively choose a topology and construct the proper structural elements to form the backbone. A number of strategies to restrict the local backbone geometries to be native-like have been employed, for example, by borrowing true fragments from actual proteins to initiate the construction or extensively idealizing the peptide chain according to reliable chemical knowledge or parametric equations. While computer algorithms have largely automated specific steps of protein design, the protein designer still controls the process and makes certain that the resulting structures are coherent.

But what decisions do human designers make that today's automated algorithms do not?

This question prompted the development of Foldit, a video game that applies a graphical user interface to the protein modeling suite Rosetta. In addition to serving as an excellent educational tool, Foldit aims to explore the strategies humans use to solve protein structure puzzles in hopes that these operations can be analyzed to improve or automate design algorithms. Foldit began with puzzles that challenged players to predict the folds of natural amino acid sequences (Figure 1A). Recently, it has been extended to allow players to modify previously designed proteins or design novel proteins from scratch (Figure 1B,C).

There are three main components involved in the design of proteins: scoring metrics to guide the moves, strategies to change the structure, and sequence tweaks to improve models (Figure 1D,E). In Foldit, the latter two are controlled by human players. There is little difference between what a player may do compared to what a trained protein designer might because their objectives are the same: to follow the score provided by the force field as it is not possible to mentally follow the entire system of thousands of atoms. In a study published in *Nature*, Koepnick et al. let Foldit players design a folded peptide starting from a linear chain.[3] Players were exceptionally good at exploring the conformational space, as seen in an early iteration of the game, where the players' structures were truly novel and expressive. While many of these creative models would not likely fold to their target structures, the real implication of the crowd sourcing brilliance is that now every aspect of the Rosetta scoring function is being tested, and exploited, in unintended ways by the players to achieve a better score. Fixing scoring deficits identified by players will eventually make the scoring metric more robust. Indeed, in subsequent rounds, Foldit was configured to enforce packing and backbone regularization rules; remarkably, these improvements provided sound guidance, and the citizen scientists were able to design proteins at the same level of accuracy as expert designers who are trained in structural biology.

Perhaps not surprisingly, with the imposition of build rules, the models produced in Foldit are no longer shockingly different from designs that trained experts have long been able to produce. However, for nonscientists to achieve these novel designs by simply maximizing the game score, the Foldit experiment shows that the scoring scheme (i.e., the Rosetta force field) must be remarkably robust. By specifying the secondary structure content required or other more general rules, the scientists behind Foldit also seem to be able to guide players into creating a wide variety of structures within specific folds. The quality of the models seems only as good as the rules set by the scientists.[4] It will be fascinating to see how this interplay between knowledge-derived rules and human creativity can be harnessed to advance science.

Automated computer algorithms today cannot carry out the design tasks like the human Foldit players; the calculations would take far too long to sample to produce a viable structure

Received: September 5, 2019

DOI: 10.1021/acs.biochem.9b00820
Biochemistry XXXX, XXX, XXX−XXX

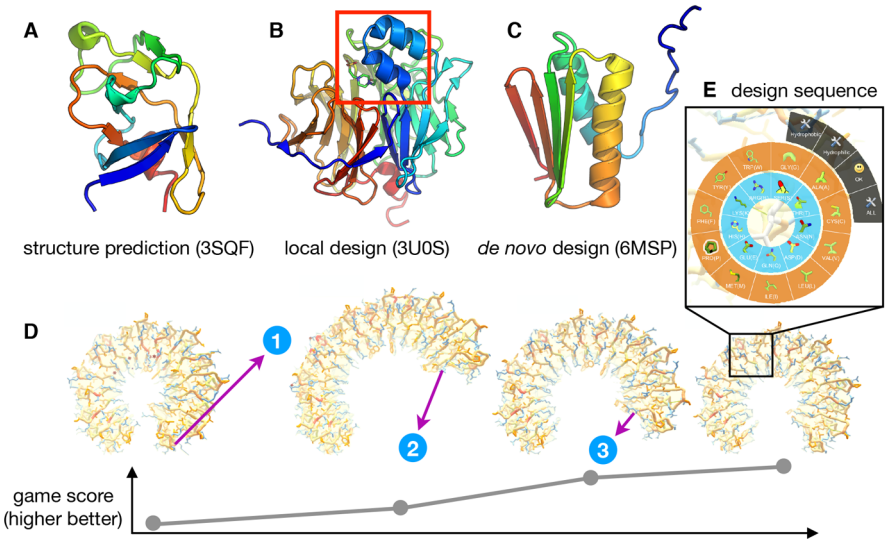

Figure 1. Foldit examples and the design steps. Foldit players have successfully completed various protein structure prediction and design problems, including (A) aiding structure prediction of Mason-Pfizer monkey virus (M-PMV),[1] (B) increasing Diehls-alderase activity by redesigning the loops around the active site of a previously designed enzyme,[2] and (C) designing a new structure from scratch. (D) The Foldit interface enables players to visualize the "score" of their designed structures as they perturb the protein backbone and (E) modify the amino acid sequence of the design.

without human guidance. How do we leverage these impressive results from citizen scientists to improve design algorithms? Crowdsourcing the "human neural network" by Foldit games can efficiently sample the protein fold space and expose flaws in the simulation; a highly refined yet robust scoring scheme might also help in improving artificial neural networks to the task, as the field has started to pay attention to these new approaches.[5] Foldit will continue to push the field toward solving complex modeling problems with creative human solutions, creating paths where new algorithms may follow. However, with machine learning agents beating humans in Go, chess, and video games, it would not come as a surprise if someday computers also beat humans at protein design.

AUTHOR INFORMATION

Corresponding Author

*E-mail: possu@stanford.edu.

Funding

P.-S.H. and K.A.T. are supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Scientific Discovery Through Advanced Computing (SciDAC) program, and the Schools of Medicine and Engineering of Stanford University.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Brian Koepnick, Ian Haydon, Paul Nuyujukian, and Ross Venook for suggestions and feedback.

REFERENCES

(1) Khatib, F., DiMaio, F., Cooper, S., Kazmierczyk, M., Gilski, M., Krzywda, S., Zabranska, H., Pichova, I., Thompson, J., Popovic, Z., Jaskolski, M., and Baker, D. (2011) Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nat. Struct. Mol. Biol. 18* (10), 1175−1177.

(2) Eiben, C. B., Siegel, J. B., Bale, J. B., Cooper, S., Khatib, F., Shen, B. W., Stoddard, B. L., Popovic, Z., and Baker, D. (2012) Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nat. Biotechnol. 30* (2), 190−192.

(3) Koepnick, B., Flatten, J., Husain, T., Ford, A., Silva, D. A., Bick, M. J., Bauer, A., Liu, G., Ishida, Y., Boykov, A., Estep, R. D., Kleinfelter, S., Nørgård-Solano, T., Wei, L., Foldit Players, Montelione, G. T., DiMaio, F., Popovic, Z., Khatib, F., Cooper, S., and Baker, D. (2019) De novo protein design by citizen scientists. *Nature 570*, 390−394.

(4) Koepnick, B., and Haydon, I. (2019) Protein Design by Citizen Scientists. Nature research bioengineering community behind the paper. https://bioengineeringcommunity.nature.com/users/264137-brian-koepnick/posts/50212-protein-design-by-citizen-scientists.

(5) Anand, N., and Huang, P. (2018) Generative modeling for protein structures. *Advances in Neural Information Processing Systems 31*, 7504−7515.

DOI: 10.1021/acs.biochem.9b00820
Biochemistry XXXX, XXX, XXX−XXX