PROTEIN FUNCTION PREDICTION

Combining views for newly sequenced organisms

Newly sequenced organisms present a challenge for protein function prediction, as they lack experimental characterisation. A network-propagation approach that integrates functional network relationships with protein annotations, transferred from well-studied organisms, produces a more complete picture of the possible protein functions.

Yingying Zhang, Shayne D. Wierbowski and Haiyuan Yu

How does 3D vision work? By projecting an object from two different angles and combining disparate views, extra information in a third dimension is gained that represents a more comprehensive characterisation of the object than a 2D picture taken from one angle. By analogy, if we regard proteins of a newly sequenced organism as an object, could we gain more insight into their functions by considering two different viewpoints? In this issue of *Nature Machine Intelligence*, Torres et al. present a Sequence to Function (S2F) approach that introduces a similar idea to facilitate highly accurate protein function prediction for newly sequenced organisms[1] (Fig. 1).

Since the advent of next-generation sequencing techniques, the expansion of protein sequence databases has exponentially surpassed the rate at which these proteins can be experimentally characterised; today, fewer than 1% of available sequences are reliably annotated[1]. As a result, functional characterisation of newly sequenced proteins increasingly relies on annotation transfer from well-studied homologous proteins in other organisms[2]. Homologous proteins originate from common ancestral protein sequences and usually fulfil similar functional roles in different organisms. However, just like projecting an object from one angle, homology-based transfer of functional annotations alone is not enough to comprehensively characterise protein functions. In addition to relying on the intrinsic properties encoded by its sequence, a protein's function also relies on its relationship with other proteins. For instance, proteins encoded by genes that are co-expressed or proteins that interact to form a single complex are likely to share molecular functions. Therefore, transferring experimentally characterised functional network relationships from well-studied model organisms provides a

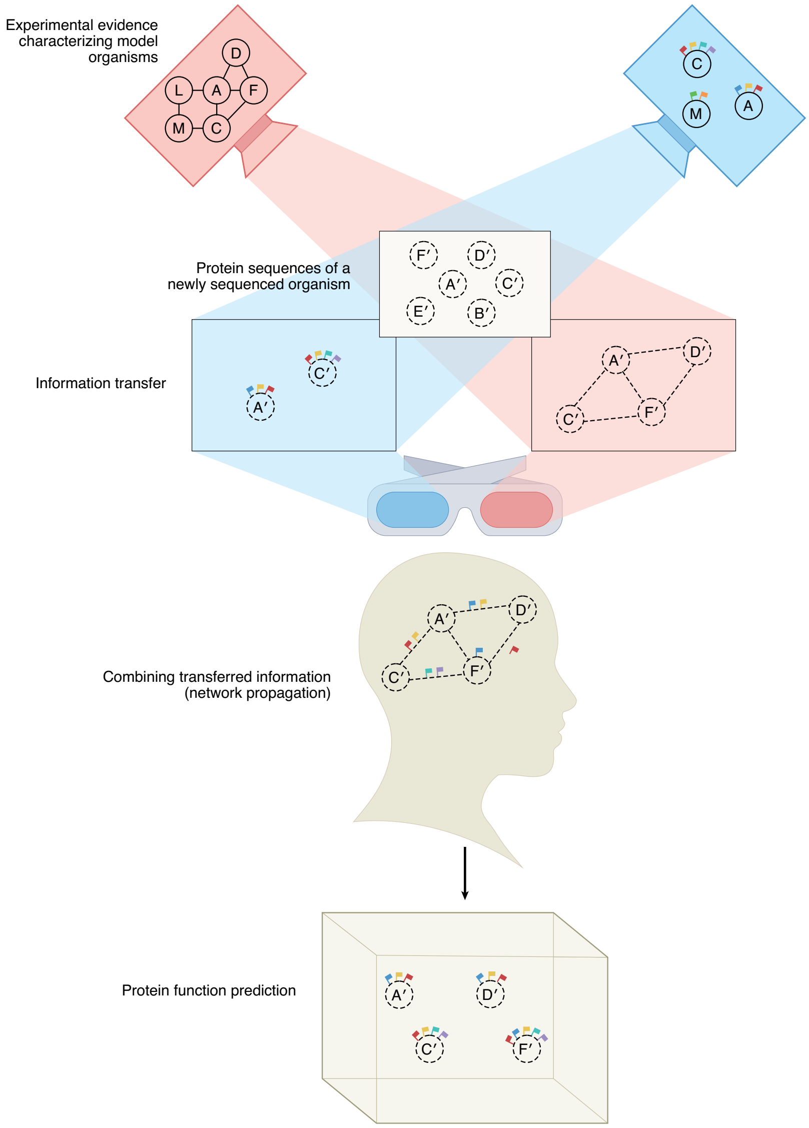
Experimental evidence characterizing model organisms
D
L
A
F
M
C
C
M
A
Protein sequences of a newly sequenced organism
F′
D′
A′
C′
E′
B′
Information transfer
C′
A′
A′
D′
C′
F′
Combining transferred information (network propagation)
A′
D′
C′
F′
Protein function prediction
A′
D′
C′
F′

Fig. 1 | Combining different views of a newly sequenced organism produces a more complete picture of possible protein functions. Functional annotations and pairwise relationships characterise proteins from different perspectives. Combining the two types of information transferred from well-studied organisms produces a more complete picture of the possible protein functions for a newly sequenced organism.

distinct perspective for studying the protein functions of a newly sequenced organism.

Following that idea, the authors develop S2F, which performs two types of homology-based information transfer from model organisms with rich experimental evidence to a newly sequenced organism: one is the transfer of functional annotations[3,4], and the other is the projection of pairwise relationships. In order to combine the transferred information that characterises protein functions from two different perspectives, the authors introduce a network-propagation approach, whereby the proteins are represented by nodes with labels (annotations), and the pairwise relationships are represented by edges. By diffusing the initial seed labels assigned to nodes on the basis of direct transfer of functional annotations, the labels are propagated to other connected nodes.

The authors validate their method by reproducing functional annotations from ten bacterial genomes. S2F consistently outperforms existing sequence-based functional annotation tools. Importantly, the authors demonstrate that the seed-label propagation over the transferred functional networks is the critical factor in successful annotation. Most strikingly, this approach yields considerably higher precision, indicative of highly confident gene annotation among these bacterial proteins.

By combining two different perspectives of functional characterisation transferred from well-studied organisms, Torres et al. introduce an innovative concept of protein function prediction under the circumstance in which only sequences are available for an organism. The work opens up the possibility of extending sequence-based protein function prediction in two distinct directions. First, by switching the angles of projection, more experimental evidence could be transferred to newly sequenced organisms, including but not limited to protein structural information[5], post-translational modifications[6] and other experimental characterisations. Second, the network propagation could be substituted by deep learning–based approaches such as graph convolutional networks[7]. Such approaches could learn the weights of neighbouring proteins in different networks, which would allow a more adaptable propagation process.

Although sequence-based protein function prediction is a promising research direction, new proteins that lack well-characterised homologues continue to pose a challenge for the field. This emphasises the importance of continued experimental efforts to characterise the function of less-studied proteins. At the same time, methods such as the one presented by Torres et al. can help bridge the gap between unannotated sequences and annotated sequences. ❐

Yingying Zhang[1,2,3],
Shayne D. Wierbowski[1,2] and Haiyuan Yu[1,2] ✉

[1]Department of Computational Biology, Cornell University, Ithaca, NY, USA. [2]Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY, USA. [3]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA.
✉*e-mail: haiyuan.yu@cornell.edu*

Published online: 9 December 2021
https://doi.org/10.1038/s42256-021-00426-8

References

1. Torres, M., Yang, H., Romero, A. E. & Paccanaro, A. *Nat. Mach. Intell.* https://doi.org/10.1038/s42256-021-00419-7 (2021).
2. Zhao, Y. et al. *Front. Genet.* **11**, 400 (2020).
3. Mitchell, A. L. et al. *Nucleic Acids Res.* **47**, D351–D360 (2019).
4. Wheeler, T. J. & Eddy, S. R. *Bioinformatics* **29**, 2487–2489 (2013).
5. Gligorijević, V. et al. *Nat. Commun.* **12**, 3168 (2021).
6. Jensen, L. J. et al. *J. Mol. Biol.* **319**, 1257–1265 (2002).
7. Zhang, S., Tong, H., Xu, J. & Maciejewski, R. *Computat. Soc. Netw.* **6**, 11 (2019).

Competing interests

The authors declare no competing interests.