PNAS COMMENTARY

Phylogenetic profiling in eukaryotes comes of age

David Moi[a,b,1] and Christophe Dessimoz[a,b,1]

As sequencing efforts uncover more and more of the staggering diversity of our biomes, we face the challenge of ascribing biological functions to many uncharacterized genes. Phylogenetic profiling exploits the coevolutionary patterns of genes involved in the same biological processes and interactions. In 1999, the pioneering work of Pellegrini et al. (1) introduced this analytical method in PNAS, demonstrating that correlated presence or absence of a gene will often point to metabolic, regulatory, or physical interaction between different proteins. Twenty-four years later, in this issue of PNAS, Dembech et al. (2) show how far phylogenetic profiling has come. The work embodies advances both at the methodological level and in the way phylogenetic profiling drives new hypotheses and directs confirmatory experiments.

Here, we recap some of the key biological and methodological milestones culminating in the contribution by Dembech et al. (2) and share some thoughts on open challenges to drive progress beyond it.

Revealing Increasingly Complex Eukaryotic Networks

Sequencing efforts from projects such as the *Darwin Tree of Life* and the *European Reference Genome Atlas* have resulted in a proliferation of quality eukaryotic genomes. Thanks to these mounting data, we are progressing from the initial success of relatively simple prokaryotic pathways to resolving interactions between individual proteins in large modular complexes with fuzzier network boundaries often found in eukaryotes.

Four studies stand out for demonstrating the potential of large-scale profiling in eukaryotes. In a landmark study published in 2013, Tabach et al. (3) profiled 86 eukaryotes and identified and experimentally validated 80 factors of the RNAi machinery. Scaling up to 177 eukaryotes, Dey et al. (4) inferred a larger set of interacting genes ("modules"), identifying and validating missing components of the actin-nucleating WASH complex and cilia/basal body genes. Next, van Hoeff et al. (5) showed that individual modules of the kinetochore can be resolved using comparative genomics and profile comparisons. Their carefully curated dataset shows the clear signal of coevolution between parts of a known complex. This complex structure underwent modular addition and deletion of components that are peripheral to the core machinery, with each submodule reflecting a clade-specific adaptation. Equally striking is the example of ciliary genes from Nevers et al. (6), where all genes in ciliated and nonciliated lineages followed different evolutionary trajectories, showing interesting neofunctionalization and loss patterns.

In line with these studies, Dembech et al. focused their profiling efforts on characterizing a specific system: the animal purine degradation pathway. The last step of the pathway, which involves the formation of glyoxylate and urea from ureidoglycolate, was known to exist, but the enzyme responsible for catalyzing this step was unknown in animals. Their search led them to interesting candidates, which were then experimentally validated. This illustrates the usefulness of profiling in filling gaps in our knowledge and revealing unknown biology that may have evaded other screening methods.

Improving Methods for the Era of Big Data

At a methodological level, profiling techniques have steadily progressed since their inception (Fig. 1). In Pellegrini et al.'s original approach, the term "phylogenetic profiling" was arguably a misnomer, since there was no phylogeny involved: The profiles of gene presence/absence in each species were constructed without accounting for any phylogenetic relationships among the species. Simplifying assumptions are needed to strike a balance between the level of detail in the representation of protein families' evolutionary histories and computational feasibility of comparing or searching among profiles. Simple presence and absence of each homologous family in extant species remain commonly used today, particularly with prokaryotes, where rampant horizontal gene transfer challenges tree-based models. Presence–absence robustly represents the co-occurence of the different clusters of functionally related genes but is vulnerable to bias when the input set of genomes used for the analysis is highly unbalanced in terms of taxonomic representation.

By contrast, encoding trees as vectors allows for computational tractability in tree comparisons while integrating features that can only be represented on a taxonomic tree. For example, in addition to considering the presence or absence of a family, we could also incorporate information on copy number, duplications, and the gain or loss of paralogous subfamilies. Further refinements can be made to vector representations by projecting them to spaces where comparison or retrieval of profiles is highly efficient (7). Other methods such as using a truncated singular value decomposition of the profiling vector can also project the data to a lower dimensional space and attenuate redundant signals coming from clades that are overrepresented in an input dataset of profiles (8).

Author affiliations: [a]Department of Computational Biology, University of Lausanne, 1015 Lausanne, Switzerland; and [b]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

Author contributions: D.M. and C.D. wrote the paper.

The authors declare no competing interest.

See companion article, "Identification of hidden associations among eukaryotic genes through statistical analysis of coevolutionary transitions," 10.1073/pnas.2218329120.

[1]To whom correspondence may be addressed. Email: David.Moi@unil.ch or Christophe.Dessimoz@unil.ch.

Published May 1, 2023.

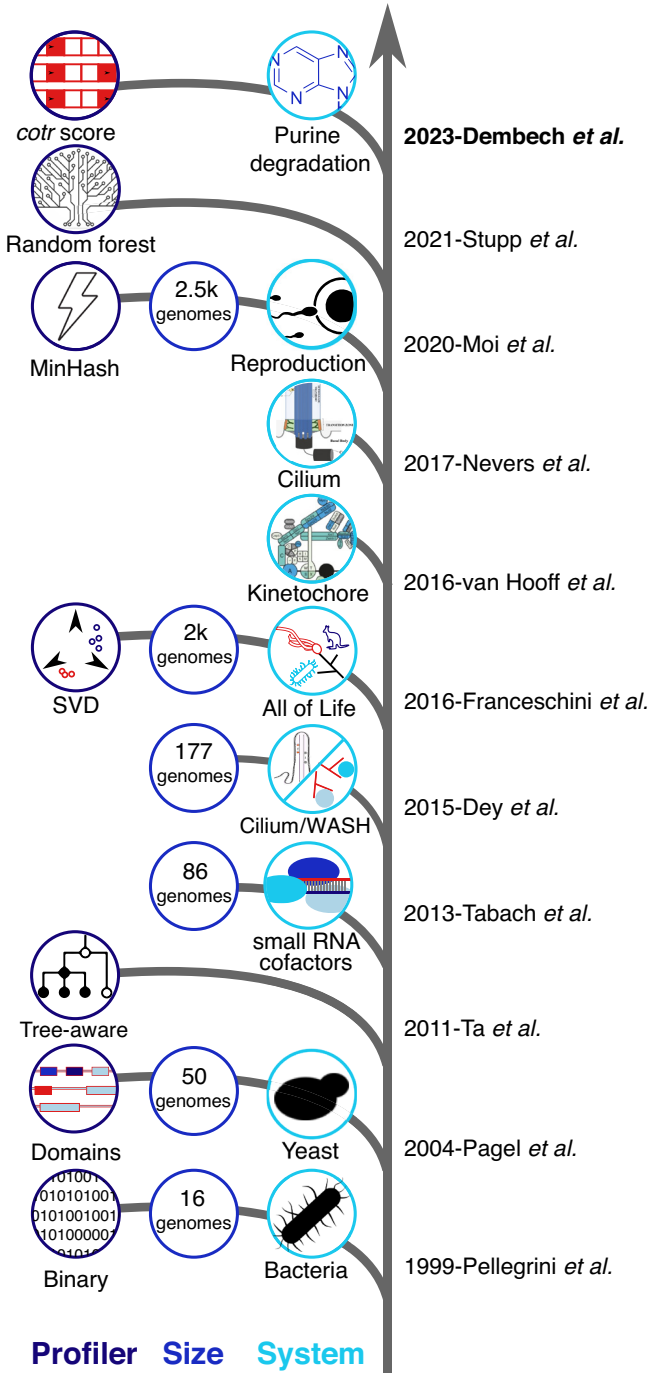

Fig. 1. Timeline of selected phylogenetic profiling advances with a focus on eukaryotes. Since Pellegrini et al.'s pioneering paper, advances in phylogenetic profiling have been made in terms of method development (and in particular profile representation), size of the datasets processed, and the kinds of systems studied—from genome-wide enrichment of known associations to reconstructing complex systems such as protein complexes and interaction networks. Novel milestones in terms of profiler, size, and system are depicted where applicable. Owing to space limitations, this list is necessarily incomplete and subjective; we apologize to authors whose valuable work is not included.

Dembech et al.'s method of profiling is centered around a cotransition ("*cotr*") score and focuses on the signal of loss and gain throughout evolutionary history. Unlike other profile representations that will categorize all "housekeeping" genes together based solely on their ubiquitous presence, the *cotr* score considers only changes in the presence and absence signal over the tree of life, revealing potentially sparse patterns of tandem losses or gains that would otherwise be insignificant. The approach is also effective for identifying proteins that neofunctionalize during the evolution of a clade and enter new subnetworks. Such emerging proteins often have largely uncorrelated patterns throughout much of the tree but show informative coevolutionary signals after the neofunctionalization event when compared using this approach.

Other recent papers have proposed more computationally intensive models that consider annotated taxonomies with the events represented on different branches (e.g., ref. 9). Tree topology can also provide valuable insights, with correlated branch lengths or sequence dissimilarity indicating selective pressures acting on multiple genes simultaneously (10). Finally, using the vector or tree representations as input for machine learning approaches is starting to take off. Stupp et al. (11) reported that when considering the interactions within a specific clade of the species tree, gradient-boosted decision tree–based regression is able to outperform explicit vector distances and vector distances projected using singular value decomposition.

Each of these profile representations and the associated comparison techniques presents a tradeoff between speed and precision. Each approach has varying suitability for answering specific biological questions and describing network evolution at particular time scales, and each has different limits on the number of genomes that can be included in the analysis.

Profiling across Time and Space to Reconstruct Ancestral and Environmental Networks

Looking ahead, open questions regarding eukaryotic biology range widely in their evolutionary time scales. The debate on what biological processes the last eukaryotic common ancestor was endowed with is still largely open (12). The same can be said of ancestral taxa such as fungi and metazoa. Finally, we can consider more recent phenomena such as the evolution of diverse clades such as birds or beetles, retrace the evolution of their phenotypes, and attempt to match transitions at the phenotypic level to a particular evolutionary signature at the genetic level. Profiling allows us to consider all of these questions that would be prohibitive to carry out considering the combinatorial possibilities of interactions of thousands of protein families within thousands of nonmodel organisms.

Furthermore, metagenomics has opened up new avenues for studying microbial diversity and understanding their metabolic capabilities. However, a significant proportion of the genomes of microbes in the environment remain uncharacterized, making it even more difficult to understand the interactions and relationships between them. These "dark networks" represent a treasure trove of novel biology waiting to be described.

By studying the relative abundances of organisms in metagenomic samples, it is possible to reconstruct ecological networks in much the same way phylogenetic profiling detects proteomic interactions (13). These ecological relationships can be explored further at the proteome level, by finding the evolutionary signatures of the metabolically

complementary roles that organisms assume in their ecological niche. With improvements in profiling methods' scalability, we may soon be able to deal with large-scale interaction networks at this scale. We may begin to understand the metabolic flows throughout an entire ecosystem and adapt conservation and bioremediation approaches in concert with these dynamics. A deeper understanding of dark networks also has great potential in industrial applications such as new enzymes or the repurposing of natural products as drugs. However, all of these potential applications are dependent on knowing what the constituents of the networks are and who they interact with.

The Need for Usability, Standards, and Community

Methodological advances should result in new biological discoveries. To see widespread adoption of cutting-edge techniques, we must also focus on the usability of our methods. So far, profiling efforts have largely been one-shot affairs, with little in the way of command-line tools, software packages, and web servers. The availability of profiling databases and easy-to-use and interoperable tools will greatly accelerate the pace of adoption of these methods. Serving predictions online will also increase their visibility but presents its own set of challenges. Notably, the STRING database systematically includes co-occurrence scores produced by the Singular Value Decomposition-phy (SVD-phy) method as part of the potential evidence channels supporting the interactions in their database (8). OrthoInspector (14) and the OMA orthology database from our lab (7) also offer phylogenetic profiling search functionality.

Furthermore, there is a noticeable absence of a cohesive community centered on phylogenetic profiling research. Other communities often rely on benchmarking datasets and competitions to compare performance across different tasks but these standards are not yet established in phylogenetic profiling. As this technique gains recognition and proves its value, it is essential to establish clear standards, practices, and community competitions similar to the Critical Assessment of protein Structure Prediction (CASP) (15), the Critical Assessment of Functional Annotation (CAFA) (16), or the Quest for Orthologs benchmarking efforts (17). This will facilitate the comparison of performance across different methods and tasks as well as helping a community to form around the methods, ultimately advancing the field.

ACKNOWLEDGMENTS. We thank Natasha Glover and Yannis Nevers for valuable feedback on this commentary as well as the Swiss NSF for funding (grant 205085).

1. M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, T. O. Yeates, Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 4285–4288 (1999).
2. E. Dembech *et al.*, Identification of hidden associations among eukaryotic genes through statistical analysis of coevolutionary transitions. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2218329120 (2023).
3. Y. Tabach *et al.*, Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence. *Nature* **493**, 694–698 (2013).
4. G. Dey, A. Jaimovich, S. R. Collins, A. Seki, T. Meyer, Systematic discovery of human gene function and principles of modular organization through phylogenetic profiling. *Cell Rep.* **10**, 993–1006 (2015), 10.1016/j.celrep.2015.01.025.
5. J. J. van Hooff, E. Tromer, L. M. van Wijk, B. Snel, G. J. Kops, Evolutionary dynamics of the kinetochore network in eukaryotes as revealed by comparative genomics. *EMBO Rep.* **18**, 1559–1571 (2017).
6. Y. Nevers *et al.*, Insights into ciliary genes and evolution from multi-level phylogenetic profiling. *Mol. Biol. Evol.* **34**, 2016–2034 (2017).
7. D. Moi, L. Kilchoer, P. S. Aguilar, C. Dessimoz, Scalable phylogenetic profiling using MinHash uncovers likely eukaryotic sexual reproduction genes. *PLoS Comput. Biol.* **16**, e1007553 (2020).
8. A. Franceschini, J. Lin, C. von Mering, L. J. Jensen, SVD-phy: Improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles. *Bioinformatics* **32**, 1085–1087 (2016).
9. H. X. Ta, P. Koskinen, L. Holm, A novel method for assigning functional linkages to proteins using enhanced phylogenetic trees. *Bioinformatics* **27**, 700–706 (2011).
10. I. R. Sadreyev, F. Ji, E. Cohen, G. Ruvkun, Y. Tabach, PhyloGene server for identification and visualization of co-evolving proteins using normalized phylogenetic profiles. *Nucleic Acids Res.* **43**, W154–W159 (2015).
11. D. Stupp *et al.*, Co-evolution based machine-learning for predicting functional interactions between human genes. *Nat. Commun.* **12**, 6454 (2021).
12. A. J. Crapitto, A. Campbell, A. J. Harris, A. D. Goldman, A consensus view of the proteome of the last universal common ancestor. *Ecol. Evol.* **12**, e8930 (2022).
13. Y. Deng *et al.*, Molecular ecological network analyses. *BMC Bioinformatics* **13**, 113 (2012).
14. Y. Nevers *et al.*, OrthoInspector 3.0: Open portal for comparative genomics. *Nucleic Acids Res.* **47**, D411–D418 (2019).
15. A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, J. Moult, Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins* **89**, 1607–1617 (2021).
16. N. Zhou *et al.*, The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.* **20**, 244 (2019).
17. Y. Nevers *et al.*, The quest for orthologs orthology benchmark service in 2022. *Nucleic Acids Res.* **50**, W623–W632 (2022).