Preview

Advancing remote homology detection: A step toward understanding and accurately predicting protein function

Predrag Radivojac[1],*
[1]Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115, USA
*Correspondence: predrag@northeastern.edu

https://doi.org/10.1016/j.cels.2022.05.006

Identifying homologous proteins with divergent amino acid sequences can add to our understanding of protein evolution, structure, and function. A new study reports the development of a deep-network-based method to identify 6.8 million new Pfam members, a dramatic singular increase that exceeds a decade of accumulation using traditional approaches.

Finding groups of proteins with similar amino acid sequences has been of interest since the early days of computational biology. Russell Doolittle (1986) discussed "surprising similarities" between proteins found almost half a century ago (Table V in his book titled "Of URFs and ORFs"), as the new field of database searching and protein family modeling was about to take shape. With the subsequent rise of genome and metagenome sequencing contributing an ever-increasing number of sequences, the identification of protein families—and especially their divergent members, called remote homologs—created the framework for asking fundamental questions about evolutionary, structural, and functional properties of these families. What were the evolutionary processes that led to the "surprising" sequence similarities within and between species? Do proteins from the same family share structure and conformational dynamics? And to what extent do they carry out the same function?

There have been two principal threads in the field of identifying remote homologs: methods whose core engine is sequence alignments (alignment-based methods) and methods based on protein sequence signatures powered by supervised machine learning (alignment-free methods), though there is not always a clear distinction between the two. Alignment-based methodologies emerged first, predominantly involving pairwise similarity searches, multiple sequence alignments, and sequence-to-profile alignments (Gribskov et al., 1987; Altschul et al., 1997). The most formal of those techniques is the probabilistic modeling of protein families using hidden Markov models (HMMs), a generative approach based on the Markov-chain assumption underlying observed protein primary structure (Eddy, 1998). The turn of the 21st century introduced alignment-free methods in the form of string and profile kernels, typically based on sparse encoding (embedding) of proteins into high-dimensional spaces, from which supervised methods could be trained to identify new family members (Leslie and Kuang, 2004). These methods depart from sequence alignments, except when the seed families are constructed, and instead rely on counting *k*-mers to identify signatures of each family—that is, groups of *k*-mers that are enriched in individual families but not others (Leslie and Kuang, 2004). However, the state of the art has not substantially changed, in part because alignment-free methods, unlike HMM-based approaches (Eddy, 2011), have not readily been translated into software tools.

The work by Bileschi et al. (2022), recently published in *Nature Biotechnology*, presents a conceptual and practical advance in the field of remote homology detection, with implications for protein function prediction. The authors have developed a new deep-learning alignment-free approach by combining a series of convolutional neural network layers that create an embedding of an input sequence, on top of which is a multiclass logistic regression model that scores Pfam (Mistry et al., 2021) categories and picks the most likely one for the given embedding. The embedding machinery is the core of this paper, resulting in a real-valued vector representation of each sequence. It starts with accepting a one-hot encoding of an amino acid sequence as input, basically a 20 × n matrix per sequence of length n, where only a single element in each column is a 1 (denoting the observed amino acid in the sequence) and all other elements are 0. It then applies a series of filtering and pooling operations to generate a flat 1,100-dimensional vector; the filtering coefficients as well as other parameters (kernel size, number of channels, etc.) are learned from the data during optimization and model selection.

There are two extensions to the base model (ProtCNN). The first one (ProtREP) is created by averaging the embeddings of sequences from the same family. This step effectively creates a cluster center in the embedded space, allowing for easy assignment of new sequences to the nearest cluster centers. It also allows for visualization of learned embeddings after further projection to two-dimensional spaces. The second extension is created by training multiple ProtCNN models on the same data through differential network initialization, leading to an ensemble of deep networks (ProtENN) that is more stable and more accurate than a single network. These models are trained on a large set of labeled (hand-curated) protein domain sequences from Pfam. Although these models are effectively classification machines, they are easily transformed from domain

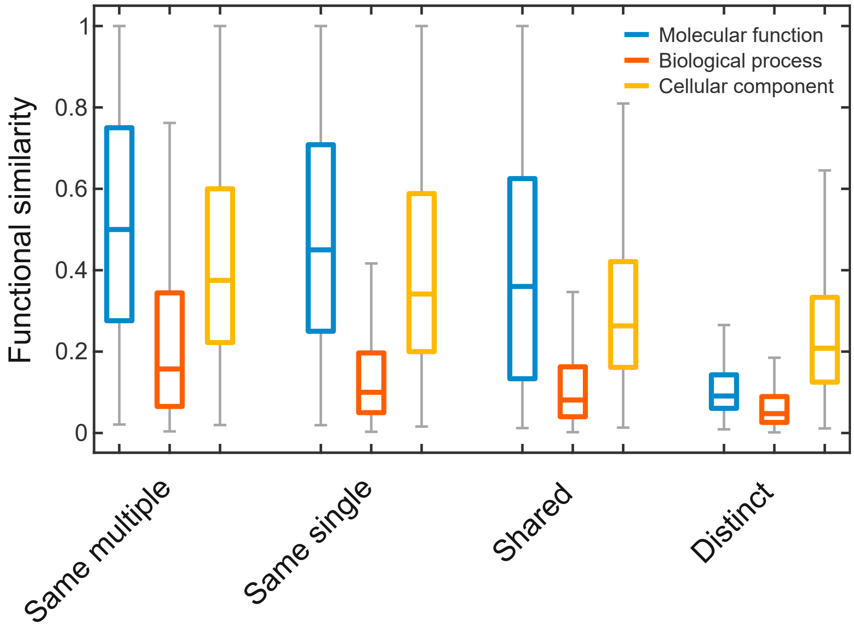

Figure 1. Similarity of functional annotations between pairs of proteins with different Pfam domain structure

Functional similarity is visualized using standard box plots (with removed outliers) over all pairs of sequences in a category. It was measured using Jaccard coefficient between propagated Gene Ontology terms separately for "Molecular Function," "Biological Process," and "Cellular Component" subontologies. "Same multiple" refers to proteins with multiple Pfam domains, identical for both proteins; "same single" refers to single Pfam domain proteins with an identical domain; "shared" refers to proteins that have at least one common Pfam domain but are not in the previous two categories; and "distinct" refers to pairs of proteins that do not share any Pfam domains. A sample of UniProtKB proteins was considered, and proteins without any Pfam annotations were omitted. It was required that both proteins in each pair have experimentally determined Gene Ontology terms in the relevant subontology. Note, however, that incomplete annotations may differentially affect average annotation similarity as well as relative comparisons.

classification into domain calling by considering all contiguous subsequences of an input protein, thus allowing the new system to find regions that belong to different Pfam domains. The embedding engine further allows the identification of new domains by clustering in the embedded space without ever performing sequence alignments.

The authors first evaluate the accuracy of new methods by comparing them to a traditional sequence alignment (based on the Basic Local Alignment Search Tool), profile HMMs, and supervised *k*-mer approaches (with dimensionality reduction and logistic regression) to demonstrate superior performance in identifying family members of existing manually curated Pfam families. While the new methods outperformed all other individual methods, the best performance was achieved by a combination of HMMs and deep networks, suggesting complementarity. The authors then go on to explore how *in silico* mutations impact sequence-to-family classification. They similarly derive a new data-driven scoring matrix, showing it to be similar to BLOSUM62 and providing evidence that the networks have learned meaningful information. Finally, and importantly, the top-performing method was applied to unannotated sequences to increase the confident annotations in the full Pfam database by 9.5%—a staggering 6.8 million previously unclassified Pfam members. This is a significant advance compared to only a 5% cumulative increase in the past several years.

There are three additional takeaways from this work. First, the deep networks outperformed profile HMM-based models. This is a surprising outcome that questions the long-standing underpinnings of protein sequence modeling—i.e., the Markov-chain assumption as an all-encompassing model. Though researchers in the field are certainly not surprised by the limitations of Markov modeling, this challenge comes from a set of obscure operations of filtering, pooling, and gradient-descent-driven parameter optimization that will take some time to understand. Second, the authors find that a combination of deep-learning-based models and profile HMMs in fact gives the best performance, suggesting complementarity of the two approaches—and a new standard for identifying remote homologs! Third, it will be worth understanding whether the new approaches emulate advances in 3D structure prediction or are in fact capturing long-range residue interactions beyond what was previously possible from solved 3D structures (Jumper et al., 2021). Since the Pfam training data is considerably larger than that used for structure prediction, additional information such as underlying evolutionary processes, conformational dynamics, or protein function could be modeled.

Though a tangible advance in the field, the task of accurately predicting protein function at all levels of abstraction remains challenging (Figure 1). About 50% of human and yeast proteins and 60% of *E. coli* proteins have only a single Pfam domain, and even for those, the large within-family sequence divergence may hamper fine-grained function prediction. For those proteins with multiple domains (40% in human, 25% in yeast, 30% in *E. coli*), the combinatorial effects of divergent Pfam sequences present even more difficult challenges. Some answers to these questions may emerge soon. The upcoming rounds of the Critical Assessment of Functional Annotation could be an opportunity to deploy the new paradigm to directly infer protein function (Radivojac et al., 2013). The infusion of new ideas will benefit protein function prediction and with it the biomedical sciences.

ACKNOWLEDGMENTS

The author is grateful to Profs. Yana Bromberg, Iddo Friedberg, and Matthew Hahn for discussion and feedback.

DECLARATION OF INTERESTS

The author declares no competing interests.

REFERENCES

Altschul, S.F., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402. https://doi.org/10.1093/nar/25.17.3389.

Bileschi, M.L., Belanger, D., Bryant, D.H., Sanderson, T., Carter, B., Sculley, D., Bateman, A., DePristo, M.A., and Colwell, L.J. (2022). Using deep learning to annotate the protein universe. Nat. Biotechnol. https://doi.org/10.1038/s41587-021-01179-w.

Doolittle, R.F. (1986). Of URFs and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences (University Science Books). Mill Valley, California, USA.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics *14*, 755–763. https://doi.org/10.1093/bioinformatics/14.9.755.

Eddy, S.R. (2011). Accelerated profile HMM searches. PLoS Comput. Biol. *7*, e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

Gribskov, M., McLachlan, A.D., and Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. Proc. Natl. Acad. Sci. U. S. A. *84*, 4355–4358. https://doi.org/10.1073/pnas.84.13.4355.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Leslie, C., and Kuang, R. (2004). Fast string kernels using inexact matching for protein sequences. J. Mach. Learn. Res. *5*, 1435–1455.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: the protein families database in 2021. Nucleic Acids Res. *49*, D412–D419. https://doi.org/10.1093/nar/gkaa913.

Radivojac, P., Clark, W.T., Oron, T.R., Schnoes, A.M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., et al. (2013). A large-scale evaluation of computational protein function prediction. Nat. Methods *10*, 221–227. https://doi.org/10.1038/nmeth.2340.