Biological molecular function: methods and benchmarks for finding function in biological dark matter

Jason E. McDermott
Computational Biology Group, Pacific Northwest National Laboratory
Richland, Washington 99352, USA
Department of Molecular Microbiology and Immunology, Oregon Health & Science University
Portland, Oregon 97239, USA
Email: Jason.McDemott@pnnl.gov

Yana Bromberg
Departments of Computer Science and Biology, Emory University
Atlanta, Georgia 30322, USA
Email: yana.bromberg@emory.edu

Hannah Carter
Division of Genomics and Precision Medicine, University of California, San Diego
San Diego, California 92093, USA
Email: hkcarter@ucsd.edu

Travis Wheeler
College of Pharmacy, University of Arizona
Tucson, Arizona 85721, USA
Email: twheeler@arizona.edu

The accurate determination of biological molecular function remains one of the most significant challenges in computational biology, with vast areas of biological "dark matter" persisting in microbiomes, viruses, and unexplored sequence space. To meet this challenge, we developed at PSB session to address the limitations of traditional sequence similarity-based functional annotation methods and explores how recent advances in AI/ML and high-throughput data generation are transforming the field. We highlight four innovative contributions presented in this session: a geometric framework using signed distance functions for modeling protein surfaces; a reinforcement learning-based approach for steering protein generative models to design functional sequences; an ensemble framework combining sequence, structural, and network features for subcellular localization prediction; and a scalable factorization method integrating gene-gene interaction data for analyzing high-dimensional genetic perturbation profiles. Together, these methodologies showcase the potential for computational and AI-driven tools to address the complex and multiscale nature of molecular function prediction, paving the way for new discoveries in understanding and engineering biological systems.

Keywords: function, machine learning, artificial intelligence, sequence.

1. Introduction

What is biological molecular function? We doubt you'd be able to get two experts on the subject to agree on what precisely this entails. By relying on textbook definitions, the term can be interpreted to mean the specific biochemical activity performed by any molecule. Within a cell this is typically a protein or RNA. Activity includes enzymatic activities, binding interactions, and structural roles among an infinity of other functions. Here, however, the ease of definitions ends. Can function be described without consideration of molecular, cellular, organismal, and even community/ecosystem context? Are parts of the same pathway considered to be interacting even if they are never in the same location? Does the human ortholog of a yeast enzyme have the same function? In spite of these issues with definition, the field has been developing methods for years aimed at prediction and characterization of gene and protein function (Zhou, Jiang et al. 2019).

Even for well-characterized organisms, the number of genes and proteins with unknown or poorly defined functional annotations is substantial (Rocha, Jayaram et al. 2023), with particular blind spots for non-canonical isoforms. And in microbial communities and viruses, existing methods frequently leave more than half of the proteins with unannotated functions often failing to find even putative homologs. RNA sequences, metabolites and other small molecules are even less well-characterized, with only a fraction of identified species having any kind of known functional role. Furthermore, even the existing functional labels are limited, with many describing only aspects of function or lacking specificity to accurately describe the role of the molecule in the system.

Function depends on the context in which it takes place, as molecules participate in systems and structures at differing levels of scale and abstraction. High-throughput data generation methods, along with automated platforms for experimental investigation, offer ways to provide contextual information by measuring many thousands of molecular species in biological samples under different environmental conditions that cause changes in system interactions. Until now, however, the potential of these data streams to inform learning models for functional annotation has remained largely untapped.

The recent revolution in AI/ML methods represents a significant opportunity to make in-roads into the problem of unknown functions. Especially relevant are advances in protein structure prediction, which can now be accomplished with high fidelity from just protein sequence alone. These approaches have opened new avenues for computational function determination. However, these approaches have not solved the function annotation problem; significant limitations and dark areas of function for many molecules still remain.

This session is focused on bringing to light new thinking about the following questions: How can we define a function of a particular molecule? How can this functional label be propagated to other molecules? What is the relationship of these labels? How can labeling systems describe context at different levels of detail, and how do they link to phenotypic outcomes?

2. Overview of the functional annotation problem

Functional characterization of individual genes and proteins has a long history in science, with years of research leading to specific enzymatic, signaling, and/or systems-level functions being

determined for individual molecules. In the era of high-throughput biological data acquisition, our ability to gather novel genetic material and thus protein sequences has vastly outpaced our ability to determine functions experimentally for those molecules. In part, this is because experimental function determination is not easily possible in unculturable organisms, such as those from environmental sources, or for many viruses where experimental systems do not exist and/or genetic manipulation is challenging (Schloss and Handelsman 2005, Mahler, Costa et al. 2023).

To address this gap, databases that capture experimental functional knowledge have been developed and, concomitantly, bioinformatic methods for determining sequence similarity. Sequence similarity is the primary method used to infer evolutionary relationships between sequences and can be used to associate a novel sequence with a sequence or sequence family that has an experimentally determined function allowing transference of annotations computationally to a large number of sequences (Hamp, Kassner et al. 2013). However, it is estimated that in many organisms, 20-40% of the proteins encoded cannot be assigned a specific function (Lobb, Tremblay et al. 2020). This problem is more pronounced in sequences derived from microbiomes and viromes, where the majority of proteins can't be annotated in this way (Wang, Ma et al. 2017).

Traditional methods rely on sequence similarity to infer evolutionary (and therefore functional) relationships between genes and proteins. The BLAST (Altschul, Gish et al. 1990) software suite provides relatively rapid methods for searching large sequence databases, while modern alternatives like MMSeqs2 (Steinegger and Soding 2017) and DIAMOND (Buchfink, Reuter and Drost 2021) introduce greater speed, and profile hidden Markov models (Eddy 2011, Roddy, Rich and Wheeler 2024) can increase sensitivity particularly when provided with organized sequence families (Mistry, Chuguransky et al. 2021).

3. Use of structure for prediction of function

Recent advances in structure prediction have opened new possibilities for determining structure-based similarity from known or predicted structures, which can identify relationships not evident from sequence alone. Tools like FoldSEEK leverage structure predictions to enable rapid searching for structurally similar proteins (Pan, Wu et al. 2024, van Kempen, Kim et al. 2024). Because protein structure determines function these approaches allow transference of annotations using structural similarity. Some methods exist which aim to characterize structural properties or motifs to more directly predict function, for example, predicting binding sites and substrates directly from structure (Krivak and Hoksza 2018, Fang, Jiang et al. 2023). These are particularly exciting as they do not directly depend on the existence of well-characterized orthologs for function determination. Additionally, related methods are being developed that allow prediction of functional context via structure, to predict protein localization and interactions with cellular systems context.

For example, the paper by Scott et al., 'Implicitly and Differentiably Representing Protein Surfaces and Interfaces' presents an innovative method for computationally representing protein shapes,

emphasizing the use of signed distance functions (SDFs) to model solvent-accessible surfaces. By integrating SDFs with constructive solid geometry operations, the authors demonstrate potential advancements in predicting protein conformation with potential impact on estimates of binding potential and structure-aware design predictions. The work sets the stage for leveraging machine learning techniques, such as pretrained protein transformers, to enhance these representations and better understand protein behavior.

The paper by Viggiano et al., 'Steering Protein Generative Models at Test-Time for Guided AAV2 Capsid Design' does not attempt to predict protein function. Instead it explores methods for guiding pretrained generative models to design protein sequences with specific functional properties. Their method, ProVADA+, utilizes reinforcement learning-based adaptive masking and advanced sequence evaluation metrics to efficiently explore sequence space to generate optimized variants. The method demonstrates success in designing viable Adeno-Associated Virus (AAV2) capsids, overcoming challenges posed by rugged fitness landscapes. This approach highlights the potential for computational techniques to accelerate the rational design of proteins with complex, user-defined functionalities.

4. Protein language models for predicting function

New protein language models trained on large sets of sequence data such as ESM (Rives, Meier et al. 2021) and ProteinBERT (Brandes, Ofer et al. 2022) allow searching sequence space for contextual patterns, and may provide insight into underlying biology and thus function of those sequences. These models build on numerous advances in machine learning and deep learning in the past few decades, with many methods being developed to target specific functional prediction problems for classes of proteins, address limitations in standard sequence similarity methods, or predict specific functional properties for proteins like localization or tissue specificity. In our session, the paper by Ahmed, et al. 'HALO: Hybrid Attention Model for Subcellular Localization' integrates structural and sequence-based approaches to predict subcellular localization for proteins, an important component of functional context. It emphasizes the use of large-scale language models like ESM and ProtT5, fine-tuned on subcellular localization datasets, to generate rich protein representations. By combining sequence embeddings, structural features, and graph neural networks within an ensemble framework, the approach achieves enhanced predictive accuracy and robustness. The work underscores the transformative potential of combining structural biology and deep learning to better understand protein function.

5. Analysis of high-throughput functional assays for function prediction

The ability to assay thousands of molecules at once provides rich sources of data for a diverse range of systems. These datasets have been used extensively to provide functional information for many genes and proteins at varying levels of scale and for different purposes (McDermott, Arshad et al. 2020, Kustatscher, Collins et al. 2022). They have been also used to determine relationships between molecules, and thus inform about their function through grouping similar functions (Wolfe, Kohane

and Butte 2005) or through associating molecular profiles with phenotypes of disease outcomes or genetic features. Similarly, the ability to genetically manipulate hundreds or thousands of genes in cells provides a powerful way of examining the effects of individual genes on the system function, as well as examining interactions between genes. In our session, the paper by Chang et al., 'PertSpectra: Interpretable Matrix Factorization for Genetic Perturbations' presents a scalable approach for analyzing high-dimensional genetic perturbation datasets to uncover biologically relevant gene interactions. The method combines matrix factorization with biologically informed inductive biases, integrating gene-gene interaction graphs to improve the interpretability of learned embeddings. PertSpectra demonstrates improved computational efficiency and scalability compared to other models, making it well-suited for large-scale genetic screens. The framework offers a powerful tool for predicting cellular responses to complex perturbations and advancing our understanding of gene function and regulation.

6. Large Language Models (LLMs) in protein function prediction

The development of large language models (LLMs) has transformed computational biology by enabling sophisticated analyses of protein sequences and functions. Trained on vast quantities of text from many domains, they provide a great deal of context and can be used to help predict function. However, their ability to generalize and to predict function for proteins and systems that have not been well studied are likely limited. The comprehensiveness and accuracy of training data raises concerns about biases and generalization to understudied proteins, microbiomes, and viromes.

The paper by, Shringapure et al. demonstrate the powerful and scalable application of LLMs for causal gene prioritization in genome-wide association studies (GWAS). By synthesizing complex scientific literature, LLMs helped identify high-probability candidate genes, successfully outperforming current methods across benchmark datasets. The paper highlights the potential of LLMs to fill critical gaps in understanding functional impacts of genetic loci while underscoring the need for refined methodologies to enhance transparency and mitigate biases in their predictions.

The study by Wang et al. introduces Gene-R1, a specialized open-source LLM framework aimed at improving gene set analysis tasks. By incorporating domain-specific knowledge and reinforced fine-tuning methodologies, Gene-R1 narrows the performance gap between open models and proprietary systems. While it exhibits significant advancements in reasoning and task-specific accuracy, this study also acknowledges existing challenges, such as hallucinations outside of training domains and transferability to other ontologies, emphasizing future directions for development.

Zhapa-Camacho et al. explore the emerging potential of agent-based LLM systems, such as LangChain and CAMEL-AI, for protein function prediction. These systems enhance traditional LLM capabilities by introducing multi-step reasoning processes that integrate structured biological datasets and experimental evidence on-the-fly. This approach allows predictions to be iteratively refined, combining computational outputs with curated biological constraints. Such frameworks significantly improve both the interpretability and accuracy of protein function determination,

positioning LLM-agent systems as a powerful tool for addressing the complexity of biological datasets.

These contributions demonstrate both the progress and the persistent challenges for LLMs in protein function prediction. By addressing core issues such as limited interpretability, data bias, and the integration of curated biological knowledge, these studies push the boundaries of applying LLMs to one of computational biology's most enduring challenges.

7. Conclusions

Computational prediction of molecular function is critical to our ability to understand, utilize, and control biological systems. Our session highlights a number of advances in this area, from new approaches to using protein structure to methods for making sense of large-scale genetic perturbations. Advances in AI and systems modeling, coupled with the explosive growth of various forms of biological data from different systems is likely to drive a transformation of this problem in the near future.

8. Acknowledgments

The session organizers would like to thank Dr. Martin Steinegger for giving an invited talk at the session and the anonymous reviewers who reviewed the submitted papers for the session. Funding to support this work was provided to JEM from Pacific Northwest National Laboratory's Predictive Phenomics Initiative conducted under the Laboratory Directed Research and Development Program. PNNL is a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy under Contract No. DE-AC05-76RL01830. Funding to support JEM and TJW was provided under NIH/NIDCR U01DE034176-01.

References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Brandes, N., D. Ofer, Y. Peleg, N. Rappoport and M. Linial (2022). "ProteinBERT: a universal deep-learning model of protein sequence and function." Bioinformatics **38**(8): 2102-2110.

Buchfink, B., K. Reuter and H. G. Drost (2021). "Sensitive protein alignments at tree-of-life scale using DIAMOND." Nat Methods **18**(4): 366-368.

Eddy, S. R. (2011). "Accelerated Profile HMM Searches." PLoS Comput Biol **7**(10): e1002195.

Fang, Y., Y. Jiang, L. Wei, Q. Ma, Z. Ren, Q. Yuan and D. Q. Wei (2023). "DeepProSite: structure-aware protein binding site prediction using ESMFold and pretrained language model." Bioinformatics **39**(12).

Hamp, T., R. Kassner, S. Seemayer, E. Vicedo, C. Schaefer, D. Achten, F. Auer, A. Boehm, T. Braun, M. Hecht, M. Heron, P. Honigschmid, T. A. Hopf, S. Kaufmann, M. Kiening, D. Krompass, C. Landerer, Y. Mahlich, M. Roos and B. Rost (2013). "Homology-based inference sets the bar high for protein function prediction." BMC Bioinformatics **14 Suppl 3**(Suppl 3): S7.

Krivak, R. and D. Hoksza (2018). "P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure." J Cheminform **10**(1): 39.

Kustatscher, G., T. Collins, A. C. Gingras, T. Guo, H. Hermjakob, T. Ideker, K. S. Lilley, E. Lundberg, E. M. Marcotte, M. Ralser and J. Rappsilber (2022). "Understudied proteins: opportunities and challenges for functional proteomics." Nat Methods **19**(7): 774-779.

Lobb, B., B. J. Tremblay, G. Moreno-Hagelsieb and A. C. Doxey (2020). "An assessment of genome annotation coverage across the bacterial tree of life." Microb Genom **6**(3).

Mahler, M., A. R. Costa, S. P. B. van Beljouw, P. C. Fineran and S. J. J. Brouns (2023). "Approaches for bacteriophage genome engineering." Trends Biotechnol **41**(5): 669-685.

McDermott, J. E., O. A. Arshad, V. A. Petyuk, Y. Fu, M. A. Gritsenko, T. R. Clauss, R. J. Moore, A. A. Schepmoes, R. Zhao, M. E. Monroe, M. Schnaubelt, C. F. Tsai, S. H. Payne, C. Huang, L. B. Wang, S. Foltz, M. Wyczalkowski, Y. Wu, E. Song, M. A. Brewer, M. Thiagarajan, C. R. Kinsinger, A. I. Robles, E. S. Boja, H. Rodriguez, D. W. Chan, B. Zhang, Z. Zhang, L. Ding, R. D. Smith, T. Liu, K. D. Rodland and C. Clinical Tumor Analysis (2020). "Proteogenomic Characterization of Ovarian HGSC Implicates Mitotic Kinases, Replication Stress in Observed Chromosomal Instability." Cell Rep Med **1**(1).

Mistry, J., S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn and A. Bateman (2021). "Pfam: The protein families database in 2021." Nucleic Acids Res **49**(D1): D412-D419.

Pan, H., Z. Wu, W. Liu and G. Zhang (2024). "AlphaFun: Structural-Alignment-Based Proteome Annotation Reveals why the Functionally Unknown Proteins (uPE1) Are So Understudied." J Proteome Res **23**(5): 1593-1602.

Rives, A., J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma and R. Fergus (2021). "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences." Proc Natl Acad Sci U S A **118**(15).

Rocha, J. J., S. A. Jayaram, T. J. Stevens, N. Muschalik, R. D. Shah, S. Emran, C. Robles, M. Freeman and S. Munro (2023). "Functional unknomics: Systematic screening of conserved genes of unknown function." PLoS Biol **21**(8): e3002222.

Roddy, J. W., D. H. Rich and T. J. Wheeler (2024). "nail: software for high-speed, high-sensitivity protein sequence annotation." bioRxiv.

Schloss, P. D. and J. Handelsman (2005). "Metagenomics for studying unculturable microorganisms: cutting the Gordian knot." Genome Biol **6**(8): 229.

Steinegger, M. and J. Soding (2017). "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets." Nat Biotechnol **35**(11): 1026-1028.

van Kempen, M., S. S. Kim, C. Tumescheit, M. Mirdita, J. Lee, C. L. M. Gilchrist, J. Soding and M. Steinegger (2024). "Fast and accurate protein structure search with Foldseek." Nat Biotechnol **42**(2): 243-246.

Wang, J., Z. Ma, S. A. Carr, P. Mertins, H. Zhang, Z. Zhang, D. W. Chan, M. J. Ellis, R. R. Townsend, R. D. Smith, J. E. McDermott, X. Chen, A. G. Paulovich, E. S. Boja, M. Mesri, C. R. Kinsinger, H. Rodriguez, K. D. Rodland, D. C. Liebler and B. Zhang (2017). "Proteome Profiling Outperforms Transcriptome Profiling for Coexpression Based Gene Function Prediction." Mol Cell Proteomics **16**(1): 121-134.

Wolfe, C. J., I. S. Kohane and A. J. Butte (2005). "Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks." BMC Bioinformatics **6**: 227.

Zhou, N., Y. Jiang, T. R. Bergquist, A. J. Lee, B. Z. Kacsoh, A. W. Crocker, K. A. Lewis, G. Georghiou, H. N. Nguyen, M. N. Hamid, L. Davis, T. Dogan, V. Atalay, A. S. Rifaioglu, A. Dalkıran, R. Cetin Atalay, C. Zhang, R. L. Hurto, P. L. Freddolino, Y. Zhang, P. Bhat, F. Supek, J. M. Fernández, B. Gemovic, V. R. Perovic, R. S. Davidović, N. Sumonja, N. Veljkovic, E. Asgari, M. R. K. Mofrad, G. Profiti, C. Savojardo, P. L. Martelli, R. Casadio, F. Boecker, H. Schoof, I. Kahanda, N. Thurlby, A. C. McHardy, A. Renaux, R. Saidi, J. Gough, A. A. Freitas, M. Antczak, F. Fabris, M. N. Wass, J. Hou, J. Cheng, Z. Wang, A. E. Romero, A. Paccanaro, H. Yang, T. Goldberg, C. Zhao, L. Holm, P. Törönen, A. J. Medlar, E. Zosa, I. Borukhov, I. Novikov, A. Wilkins, O. Lichtarge, P. H. Chi, W. C. Tseng, M. Linial, P. W. Rose, C. Dessimoz, V. Vidulin, S. Dzeroski, I. Sillitoe, S. Das, J. G. Lees, D. T. Jones, C. Wan, D. Cozzetto, R. Fa, M. Torres, A. Warwick Vesztrocy, J. M. Rodriguez, M. L. Tress, M. Frasca, M. Notaro, G. Grossi, A. Petrini, M. Re, G. Valentini, M. Mesiti, D. B. Roche, J. Reeb, D. W. Ritchie, S. Aridhi, S. Z. Alborzi, M. D. Devignes, D. C. E. Koo, R. Bonneau, V. Gligorijević, M. Barot, H. Fang, S. Toppo, E. Lavezzo, M. Falda, M. Berselli, S. C. E. Tosatto, M. Carraro, D. Piovesan, H. Ur Rehman, Q. Mao, S. Zhang, S. Vucetic, G. S. Black, D. Jo, E. Suh, J. B. Dayton, D. J. Larsen, A. R. Omdahl, L. J. McGuffin, D. A. Brackenridge, P. C. Babbitt, J. M. Yunes, P. Fontana, F. Zhang, S. Zhu, R. You, Z. Zhang, S. Dai, S. Yao, W. Tian, R. Cao, C. Chandler, M. Amezola, D. Johnson, J. M. Chang, W. H. Liao, Y. W. Liu, S. Pascarelli, Y. Frank, R. Hoehndorf, M. Kulmanov, I. Boudellioua, G. Politano, S. Di Carlo, A. Benso, K. Hakala, F. Ginter, F. Mehryary, S. Kaewphan, J. Björne, H. Moen, M. E. E. Tolvanen, T. Salakoski, D. Kihara, A. Jain, T. Šmuc, A. Altenhoff, A. Ben-Hur, B. Rost, S. E. Brenner, C. A. Orengo, C. J. Jeffery, G. Bosco, D. A. Hogan, M. J. Martin, C. O'Donovan, S. D. Mooney, C. S. Greene, P. Radivojac and I. Friedberg (2019). "The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens." Genome Biol **20**(1): 244.