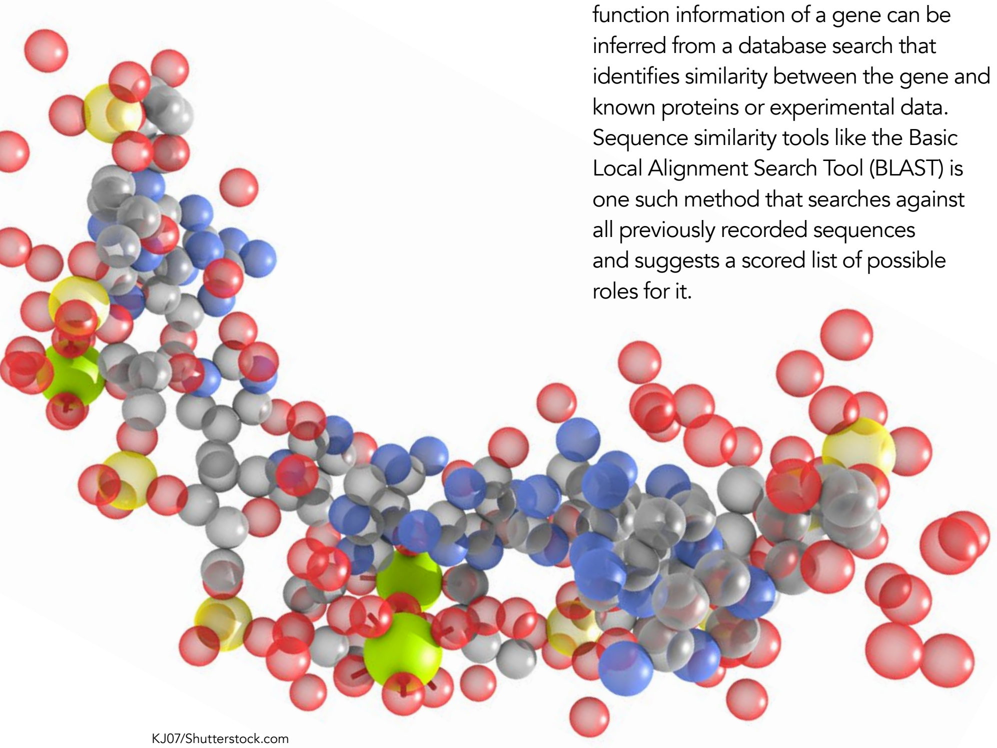
Predicting protein function and annotating complex pathways with machine learning

Dr Daisuke Kihara's team at Purdue University have created novel computational approaches for predicting protein functions. Instead of following a one-protein-one-function approach, their algorithms can predict the functional relationships of entire groups of proteins related to a specific biological process. The team has also expanded into mining oversighted or previously unknown proteins that have multiple, independent functions. The team's methods challenge the logic behind conventional protein function studies and propose tools that may better capture the complicated nature of protein interactions in biological processes.


KJ07/Shutterstock.com

Proteins are the main working units of biology. Identifying and understanding what proteins do is crucial for biologists hoping to solve the complex interactions and systems that drive cellular processes. Although protein function needs to be ultimately validated by hand in the wet lab, researchers first need a hypothesis in order to design assays, which can then define the probable function of a protein.

BIOINFORMATICS FOR PREDICTING PROTEIN FUNCTION

Biologists can build such hypotheses of gene function with computers. As genome sequencing becomes routine in experimental laboratories, computational gene function prediction has also become increasingly important. Computational methods are very suitable for function prediction because function information of a gene can be inferred from a database search that identifies similarity between the gene and known proteins or experimental data. Sequence similarity tools like the Basic Local Alignment Search Tool (BLAST) is one such method that searches against all previously recorded sequences and suggests a scored list of possible roles for it.

PROBLEMS WITH PREVIOUS COMPUTATIONAL METHODS

However, existing bioinformatic tools can't always predict protein function accurately, and often end up incorrectly annotating proteins within a biological system. Traditional protein function prediction tools like BLAST are usually reliable when a high sequence similarity is detected, but their accuracy falls quickly for sequences with lower similarities. For example, enzyme functions differ immensely when similarity scores fall below a certain level. Moreover, in many cases traditional methods do not annotate any function if highly similar sequences are not found, leaving many genes unannotated. In addition, other metrics such as similarity in three-dimensional structure, gene expression, or interaction data could be used. However, each of these metrics are often missing for many proteins under investigation, and so have limited applicability in reliable research.

NEW TOOLS FOR BETTER ACCURACY

Recently, several new protein annotation methods have been developed to improve overall prediction accuracy. One such developer is Dr Daisuke Kihara from Purdue University, who develops function prediction methods with new logical frameworks. In 2009, his team created an automated predictive algorithm, called the extended similarity group (ESG) method, which runs a continual comparing system, instead of a single search. From each sequence found from the first inquiry, the ESG algorithm runs a second search through the database. By combining results from this multi-levelled tactic, the ESG method significantly improves functional scoring for query proteins and outperforms previous function prediction algorithms.

Gorodenkoff/Shutterstock.com

Yet the team did not stop here. In a 2019 paper, they combined phylogenetic tree construction tools with traditional sequence-based prediction, called the Phylo-PFP method. They first confirmed that close similarities of protein sequences did not align with the proteins' distances on a phylogenetic tree. By adding these distances into the sequence homology score, the protein query ranks became more reliable, and they could be more accurately linked to their gene source. Unsurprisingly, the study established Phylo-PFP significantly improved the function prediction accuracy over existing methods.

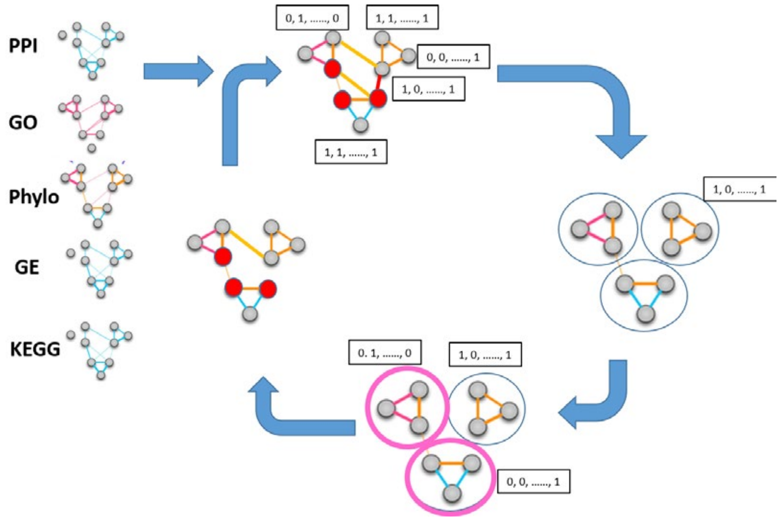
PROTEIN GROUP FUNCTION ANNOTATION

Protein function annotation is typically run on a one-protein-one-function approach, yet this mindset can grossly oversimplify the protein function universe. In fact, most experiments find dozens of interacting proteins related to a single biological event. To understand the role of an entire protein set, their function should be determined from the group as a whole, even if the function of each individual protein is unknown. This is no simple task.

Therefore, Dr Kihara's team focused on a new computational approach for annotating the functions of protein groups. In 2019, they proposed an iterative Group Function Prediction (iGFP) method, which holds a completely new logical framework at its core. The iGFP algorithm considers a set of proteins as input, and predicts the role of the function of the entire group, as well as its individual members. The iGFP algorithm blends sequence data from multiple sources and builds a complementary network. The method then separates the proteins into clusters that have functional relevance and compares them based on functional and interaction relationships.

Moreover, the system automatically assumes that some proteins are unknown and uses a range of other comparative features to make an accurate prediction. During this scan, the algorithm considers protein-protein interactions, phylogenetic profile similarity, gene co-expression, large-scale pathway similarity, and gene ontology similarity. This type of comprehensive group function prediction could be an altogether improved

Dr Daisuke Kihara from Purdue University develops function prediction methods with new logical frameworks.



The iGFP algorithm iteratively assigns functions to protein groups and to individual proteins in the groups.

Aashish Jain and Dr Kihara discuss functions assigned to a metabolic pathway.

reflection of the real mechanisms at work in, for example, developmental or disease-causing pathways.

IDENTIFYING PROTEINS WITH MULTIPLE FUNCTIONS

In addition to analysing protein groups, the Kihara team has taken another step away from the one-protein-one-function scheme by studying multi-functional proteins. Most bioinformatic tools do not take into account that proteins, enzymes in particular, can be multi-functional. The Kihara lab has thus aimed to predict whether a query protein is a moonlighting protein – one that has multiple autonomous and often unrelated functions. These proteins are difficult to annotate, since their functions are not genome or protein family specific, nor linked to other indicators, such as a shared switching mechanism. Yet these proteins play key roles in cellular disease states such as cancers, and so identifying them is important.

To solve the problem, Dr Kihara's team has developed a new systematic approach to study moonlighting proteins. In 2016, the team proposed an automated prediction framework which uses several non-sequence-based data to identify moonlighting proteins. They used machine learning classifiers to predict multi-functional proteins, after which they cross validated the results using existing databases. Dr Kihara's team could predict moonlighting proteins that had previous gene sequence data with 98% accuracy. Even if no sequence data was available, the system showed an impressive 75% accuracy.

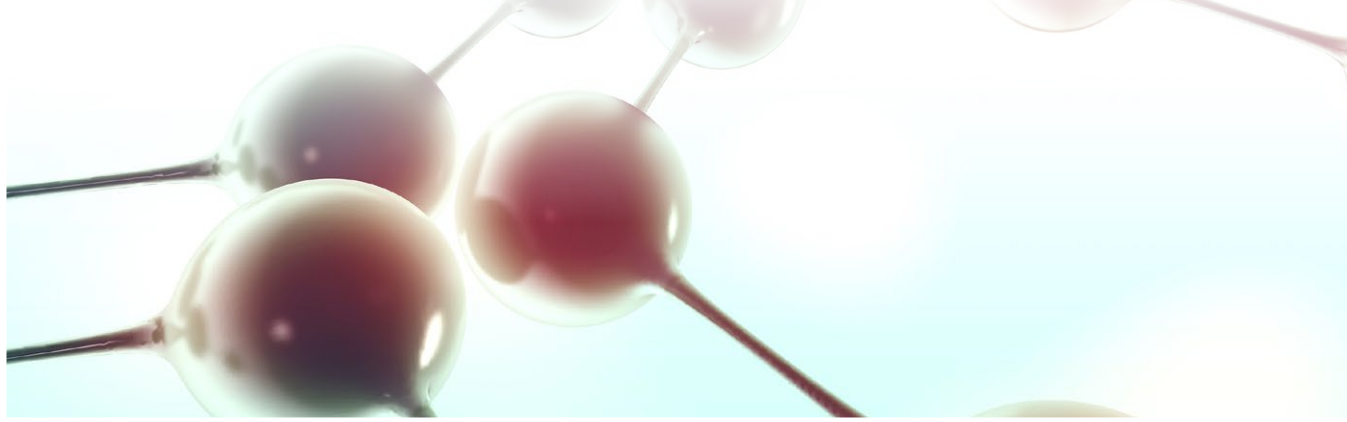
Furthermore, in a 2018 paper the team used deep learning to sniff out moonlighting proteins from previously published literature. Their text mining tool DextMP could find out whether a protein had multiple functions or not based on information from journal publications and functional descriptions from protein databases. Using systematic literature processing tools, the researchers could significantly reduce time to annotate moonlighting proteins and move closer to clarifying the complex interplay of proteins within the cell.

The iGFP algorithm considers a set of proteins as input and predicts the function of the entire group, as well as its individual proteins.

IMPROVEMENTS AND FUTURE PREDICTIONS

Computational biology desperately needs new ways to accurately reflect the true nature of biological processes. Dr Kihara's team has made innovative strides to step away from a traditional one-protein-one-function effort and identified functions for entire protein groups. Their algorithms outperform previous sequence-based methods by layering multiple protein characteristics and taking into account evolutionary relationships, which can be better indicators of shared functions than the simple amino acid backbone. Further, the team's machine learning methods can predict whether a protein serves a double role, and whether such proteins have unknowingly been described in previous literature.

Despite these promising developments, bioinformatic prediction tools are only as intelligent as their design, and there is still a way to go towards fully automated, AI-driven research in protein function annotation. Overall, Dr Kihara's team suggests that combining previous methods with emerging ones from omics experiments and evolution distance analysis will further solidify functional prediction accuracies in the future.



SmirkDingo/Shutterstock.com

Behind the Research

Dr Daisuke Kihara

E: dkihara@purdue.edu **T:** +1 765 496 2284 **W:** http://kiharalab.org

Research Objectives

Dr Kihara's work focuses on developing new techniques for computational protein function prediction.

Detail

Department of Biological Sciences
Department of Computer Science
Purdue University
249 S. Martin Jischke Dr
West Lafayette, IN 47907, USA

Bio

Dr Kihara is a full professor in the Department of Biological Sciences and the Department of Computer Science at Purdue University, West Lafayette, Indiana. He received a BS degree from the University of Tokyo, Japan in 1994, and a PhD degree from Kyoto University, Japan in 1999. After studying as a postdoctoral researcher with Prof Jeffrey Skolnick he joined Purdue University in 2003. He was promoted to full professor in 2014. Since 2018, he has held an adjunct professor position at Department of Pediatrics, University of Cincinnati. He has been working in various topics in protein bioinformatics. His current research projects include the developments of algorithms for protein-protein docking, protein tertiary structure prediction, structure modelling from low-resolution image data, structure- and sequence-based protein function prediction, and computational drug design. He has published over 150 research papers and book chapters. His research projects have been supported by funding from the National Institutes of Health, the National Science Foundation, the Office of the Director of National Intelligence, and industry. He has served on the program committee of various bioinformatics conferences including the Intelligent Systems for Molecular Biology (ISMB) where he is a track chair in 2019. In 2013, he was named a University Faculty Scholar by Purdue University.

Funding

- National Science Foundation (NSF)
- National Institutes of Health (NIH)

Collaborators

- Aashish Jain
- Ishita Khan

References

Chitale, M., Hawkins, T., Park, C., & Kihara, D. (2009). ESG: extended similarity group method for automated protein function prediction. *Bioinformatics*, 25(14), 1739–1745.

Jain, A., & Kihara, D. (2019). Phylo-PFP: improved automated protein function prediction using phylogenetic distance of distantly related sequences. *Bioinformatics*. 35(5):753-759.

Jain, A., Gali, H., & Kihara, D. (2018). Identification of Moonlighting Proteins in Genomes Using Text Mining Techniques. *Proteomics*, 18(21–22), 1800083.

Khan, I. K., & Kihara, D. (2016). Genome-scale prediction of moonlighting proteins using diverse protein association information. *Bioinformatics*, 32(15), 2281–2288.

Khan, I. K., Jain, A., Rawi, R., Bensmail, H., & Kihara, D. (2019). Prediction of protein group function by iterative classification on functional relevance network. *Bioinformatics*, 8, 1388-1394.

Personal Response

What kind of role will machine learning play in protein function prediction and understanding biological processes?

" Machine learning has already been playing a big role in protein function prediction, and more widely, in bioinformatics. It is particularly effective in identifying subtle signatures that are easily overlooked by humans in input data including protein sequences that are relevant to particular functions. It is also very suitable for integrating many different types of data together to make predictions. "

PURDUE
UNIVERSITY®