Proceedings of the 3rd International Conference on Modern Medicine and Global Health
DOI: 10.54254/2753-8818/89/2025.21447

Recent Advances in Bioinformatics for Protein Function Prediction

Shang Ma[1,a,*]

[1]*Nanjing Agricultural University, Agriculture, Nanjing, 300451, China*
a. 1527379675@qq.com
**corresponding author*

Abstract: With the continuous enhancement of computing power and the rapid expansion of biological data, bioinformatics has gradually emerged as a mainstream approach to solving biological problems. In biological research, protein function prediction is a crucial task that can significantly reduce research costs and improve efficiency. Therefore, the use of bioinformatics for protein function prediction has become a cutting-edge research focus. This article categorizes the application methods of bioinformatics in protein function prediction into three types: prediction methods based on protein sequences, prediction methods based on protein structures, and prediction methods based on protein interaction networks. It also provides a detailed analysis of key algorithms such as DeepGOPlus, MultiPredGO, and DeepGraphGO, highlighting their strengths, limitations, and recent progress. By offering an overview of state-of-the-art methods, this paper aims to serve as a comprehensive reference for advancing protein function prediction and exploring its implications in life sciences.

Keywords: bioinformatics, protein function prediction, gene ontology

1. Introduction

Proteins are indispensable molecules in living organisms, playing pivotal roles in almost every biological process, from cellular signaling and metabolism to structural support and immune response. Their functional diversity is vast and complex, contributing to the intricate interplay of biological systems that sustain life. Consequently, understanding protein function is a cornerstone of biological research, attracting a multitude of scientists who strive to unravel the mysteries encoded within these versatile molecules.

In recent years, the field of artificial intelligence (AI) has undergone rapid advancements, finding widespread applications across numerous domains. AI's ability to process vast amounts of data and uncover hidden patterns has revolutionized fields such as healthcare, finance, and engineering. Similarly, the intersection of AI with biology, particularly through bioinformatics, holds immense promise for transforming biological research.

Despite the significance of protein function research, the task remains formidable. Experimental methods for determining protein function are often time-consuming, labor-intensive, and costly. As a result, biologists frequently grapple with the challenge of understanding the functions of newly discovered proteins, which can hinder the progress of scientific discovery. This limitation underscores the urgent need for alternative approaches that can streamline the process of protein function prediction.

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/89/2025.21447

Bioinformatics, particularly when augmented by AI techniques, offers a promising solution to this challenge. By leveraging large datasets and sophisticated algorithms, bioinformatics methods can predict protein function with significantly lower costs and higher efficiency than traditional experimental methods. These methods have the potential to provide insights into the functions of unknown proteins, thereby accelerating scientific research and expanding our understanding of biological processes.

2. Definition and categorization of protein function

Protein is one of the most important biological macromolecules in living organisms, which has many functions such as building and repairing cells, transmitting signals, catalyzing chemical reactions, storing energy, and transporting substances. So protein is considered the basic functional unit of living organisms. [1,2] Anything associated with a protein can be viewed as a function of the protein.

Gene ontology (GO) [3,4] and function categories (FUN CAT) came up. GO is one of the most widely accepted and commonly used classification systems for protein energy supply. The GO protein classification system includes tens of thousands of terms covering the various functions and locations of proteins in cells and organisms. GO can divide proteins into three main aspects: molecular function (MF), biological process (BP), and cellular component (CC). Molecular function mainly describes the specific functions of protein molecules, such as catalytic reaction, binding to other molecules, etc. Biological processes primarily describe events related to protein function, such as metabolic pathways, cell signaling, etc. Cell components, such as nuclei and cell membranes, are mainly used to describe the specific location and structure of proteins in cells. Each GO term is a functional label, and the process of making a functional prediction of a protein is the process of determining that the protein has a label. Proteins in databases such as Uniprot, Ensembl, and InterPro are labeled with GO function tags, making it easy to provide functional annotations for protein sequences.

3. Protein Sequence-Based Prediction

3.1. Implications of Protein Sequence

With the continuous advancement of genomics and proteomics technologies, researchers have been able to rapidly and accurately identify protein-coding sequences within the genome. Nevertheless, fully understanding the functions of these proteins and their exact roles in biological processes remains a significant challenge yet to be tackled. To this end, scientists have extensively utilized bioinformatics approaches to conduct extensive research aimed at predicting protein functions, thereby deepening their understanding of biological systems.

The main challenge lies in precisely assessing the similarity between proteins of unknown function and those of known function in terms of sequence, function, and other aspects, posing a complex multi-dimensional classification problem. In this context, this article categorizes the strategies for protein function prediction into three main approaches: prediction based on protein sequence, prediction based on protein structure, and prediction based on protein interaction networks.

Initially, the approach for function prediction based on protein sequences relies on a thorough analysis of amino acid sequences, utilizing comparisons of sequence similarities and characteristic patterns to deduce potential protein functions. Secondly, the prediction method focusing on protein three-dimensional structures reveals functions by simulating and predicting the spatial configurations of proteins. Lastly, by analyzing the interaction networks between proteins and other biomolecules, the network-based prediction method provides insights into the functions and roles of proteins within biological systems.

The integrated application of these methods offers invaluable tools for deepening the exploration of protein functions, effectively filling knowledge gaps that are difficult to reach through

DOI: 10.54254/2753-8818/89/2025.21447

experimental means. By integrating these bioinformatics methods, we can not only comprehensively unveil the multifaceted functions of proteins in biological processes but also provide solid support for research advancements and practical applications in the field of life sciences.

3.2. Prediction Based on Protein Sequence Homology

A protein whose sequence is like that of the predicted protein is found, and the functional annotation (function, domain, reaction mechanism, structural characteristics, etc.) of this protein is assigned to the predicted protein.[5]

Employing the homology strategy involves searching for known functional proteins with sequences closely related to the target protein and then transferring the functional information (including function, structural domains, mechanisms of action, and structural characteristics) of these known proteins to the protein under prediction. In this strategy, the annotation of unknown functional proteins depends on the similarity assessment between their sequences and those of known proteins. The sequence alignment tools commonly employed in this process include FASTA, BLAST, and PSI-BLAST. However, researchers have discovered that the theory of relying on sequence similarity to infer functional similarity is a relatively weak hypothesis.

3.3. Limitations of Sequence-Based Prediction

Although this method is simple to operate for prediction, it has many limitations, such as being constrained by the number of known functional sequences and requiring long computation times. In addition, studies have shown that up to 30% of the protein function results obtained through this method are incorrect. This discovery has prompted researchers to start looking for other more effective prediction methods. In the field of protein function prediction based on sequence homology, the latest research achievements include methods such as DeepGOPlus and NCL+mask BLAST. DeepGOPlus represents an enhanced version of DeepGO, effectively addressing its shortcomings in terms of sequence length accommodation, feature relationship exploration, and prediction time consumption. By integrating convolutional neural networks (CNN) with sequence similarity-based prediction methods, DeepGOPlus offers a sophisticated solution. Pathak et al. introduced a novel protein function prediction method called NCL+mask BLAST [7]. The core of this method lies in its integration of a new amino acid classification logic based on chemical measurements with the BLAST algorithm. Specifically, it employs a new chemical logic (NCL) to filter out protein sequences irrelevant to function prediction before aligning them with sequences in the database to obtain functional predictions. The NCL method views amino acids as templates with complete chemical characteristics and proposes an innovative classification based on their stereochemical properties. During testing, this method achieved more than three times the accuracy of BLAST in predicting the biological, molecular, and cellular functions of 69,306 protein sequences with known functions in the SwissProt database. Notably, NCL+mask BLAST significantly improves prediction performance by using NCL to filter out functionally irrelevant protein sequences.

4. Protein Structure-Based Prediction

4.1. 3D structure of Proteins

This method is used to predict the function of proteins by structural information, including structural similarity comparison, protein structure simulation, and model prediction. Utilizing protein structure information for functional prediction has become a key area of research in bioinformatics. [6,7] The structure prediction method may be more accurate in predicting the function of proteins with specific structural characteristics because it directly utilizes the three-dimensional structure information of the

DOI: 10.54254/2753-8818/89/2025.21447

protein. However, due to the relatively difficult and time-consuming acquisition of high-quality 3D structural data, the application scope of structural prediction methods is limited. However, the sequence prediction method has a wider applicability because it can process many amino acid sequence data. With the development of high-throughput sequencing technology, more and more protein sequences have been discovered, which provides rich data resources for sequence prediction methods. In addition, sequence prediction methods have advantages when dealing with proteins of unknown structure. Sequence prediction methods are also more suitable for large-scale protein function prediction tasks, especially when dealing with a large number of proteins with unknown structures. In addition, sequence prediction methods can be combined with other bioinformatics methods, such as gene expression analysis, protein interaction network analysis, etc., to provide a more comprehensive prediction of protein function. The core of this method lies in recognizing the close relationship between protein structure and function, namely, that protein function is largely determined by its structure.

4.2. Multimodal Prediction Model Based on CNN

MultiPredGO[8] is a multimodal technique incorporating deep learning, with its core concept lying in combining two different data sources—protein sequences and secondary structures—to design two specialized CNN models for feature extraction. To enhance prediction efficiency, this technique also integrates protein interaction information to generate 256-dimensional knowledge graph embeddings. Ultimately, using these extracted features, a hierarchical classification model for predicting protein functions is trained.

The CNN model is a structure-function prediction method based on convolutional neural network principles aimed at predicting protein function from the tertiary structure of heme protein active sites, thereby deeply exploring the internal relationship between structure and function. [9]The core steps of this method include: first, mapping the tertiary structure of heme binding sites onto the xy plane and dividing it into multiple small cubic regions (voxels); then, selecting 3,206 different protein structure entries from the PDB (Protein Data Bank, a public database storing structural information of biomolecules) and extracting 6,866 heme molecules from them; subsequently, using the CNN model to learn the correspondence between heme protein structure and function; and finally, using the model's output as the classification label for protein function.

Emerging 3D structure prediction models, such as AlphaFold2, have significantly improved the accuracy of the field of protein function prediction. Their ability to predict the three-dimensional structure of proteins from amino acid sequences with high accuracy has not only expanded the scope of structural prediction but also included complex protein structures that were previously difficult to obtain experimentally. It also promotes the deep integration of sequence prediction and structure prediction. By using the structural data generated by these models to train the sequence prediction model and combining the prediction results of the two, researchers can more comprehensively reveal the intrinsic relationship between protein function and structure, opening a new path for protein function research, drug development, disease diagnosis and treatment and other fields, and promoting the rapid development and progress in this field.

5. Interaction Network-Based Prediction

5.1. Deep Learning Framework

Method of function prediction based on interaction networks focuses on utilizing protein structural information to predict its function, covering various aspects such as structural similarity comparison, protein structure simulation, and model prediction [5,6].Researchers have already developed bioinformatics tools based on protein structure for predicting protein function.

DOI: 10.54254/2753-8818/89/2025.21447

DeepFunc is a deep learning framework designed to predict protein function from protein sequences and network information accurately. Specifically, DeepFunc first transforms the feature information related to the input protein sequence, such as structural domains, families, and motifs collected by the InterPro tool, into a high-dimensional binary vector with 35,000 dimensions.[10] Then, two fully connected layers are used to reduce the dimensionality of this high-dimensional vector, obtaining a low-dimensional vector. Meanwhile, functional connections are obtained with EggNOG and combined with interaction information from the STRING tool to construct a Protein-Protein Interaction (PPI) network. Subsequently, the Deepwalk algorithm is employed to comprehensively extract the underlying topological features of the PPI network, which are then fused with the previous low-dimensional vector to form a fully connected network. Finally, functional classification is performed based on this fully connected network.

5.2. Graph Neural Network

DeepGraphGO is an end-to-end model based on graph neural networks. It integrates protein sequence and protein network information and employs a multi-species training strategy to accommodate different species. As an integral part of NetGO, DeepGraphGO further enhances overall performance. The model's structure comprises an input layer, a graph convolutional layer, and an output layer.

Understanding the function of a target protein is a crucial step in the drug development process. Models such as DeepFunc and DeepGraphGO can accurately predict protein function through in-depth analysis of protein sequence and interaction network information, thus providing key information for drug design. For example, when researchers are trying to develop a drug for a specific disease, they can use these models to identify proteins associated with the disease and further analyze the function of those proteins to design drug molecules that can bind specifically to the target protein.In the study of physiological mechanism, the function of protein often determines the various physiological activities of the organism. Through models such as DeepFunc and DeepGraphGO, researchers can systematically analyze the function of proteins, revealing their interactions and regulatory mechanisms in living organisms. This will help us to understand the physiological process of the organism more deeply and provide new perspectives and ideas for the analysis of disease mechanisms.

Combining interaction networks with AI for protein function prediction has significant advantages. First, the interaction network provides a wealth of protein interaction information that is critical to understanding the function of proteins. Second, AI technologies, especially deep learning, have powerful data processing and pattern recognition capabilities to efficiently process and analyze these complex interactive network data. By combining the two, we can more accurately predict the function of proteins and reveal their interactions and regulatory mechanisms in living organisms. In addition, AI technology is also able to process large amounts of data and extract useful information from it. In protein function prediction, this means that we can use AI techniques to analyze large amounts of protein sequence and interaction network information to discover new functional patterns and regulatory mechanisms. These findings will help us to better understand the physiological and pathological processes of organisms and provide new ideas and methods for disease treatment and drug development.

6. Conclusion

In recent years, with significant advancements in computing power and a dramatic increase in the volume of biological data, there has been a growing interest in using bioinformatics and deep learning techniques to address biological issues. Bioinformatics has emerged as a crucial tool. This article

DOI: 10.54254/2753-8818/89/2025.21447

reviews various methods of utilizing bioinformatics for protein function prediction and delves into the strengths and limitations of these methods.

Despite some notable progress in this field, protein function prediction still faces numerous challenges and limitations. Firstly, current methods have limitations in predicting the functions of complex proteins and modeling the intricate relationships among them. Since complex proteins often contain multiple structural domains, the interactions and functional relationships among these domains are highly complex, necessitating more refined and accurate methods to decipher this complexity. Secondly, there are certain difficulties in handling protein-protein interaction network data that encompass multiple types of relationships. Existing methods often struggle to effectively process and integrate different types of relationships, which limits the comprehensive prediction and deep understanding of protein functions. Finally, there is a lack of targeted protein function prediction tools for different types of biological data. Therefore, there is a need to develop new methods that better leverage information from different data sources to address these issues.

Research on protein function prediction using bioinformatics has made significant progress. In the future, scientists can achieve more accurate and reliable protein function prediction results by comprehensively analyzing multiple data sources, improving the interpretability of models, and exploring new algorithms and models.

References

[1] *Yuan QM, Chen S, RAO JH, Zheng SJ, Zhao HY, Yang YD. (2022). AlphaFold2-aware protein-DNA binding site prediction using graph transformer. Briefings in Bioinformatics. 2, bbab564.*

[2] *Yuan QM, Chen S, RAO JH, Zheng SJ, Zhao HY, Yang YD. (2021). Structure-aware protein-protein interaction site prediction using deep graph convolutional network. Bioinformatics. 38(1), 125-132.*

[3] *ASHBURNER M, BALL CA, BLAKE JA, BOTSTEIN D, BUTLER H, CHERRY JM, DAVIS AP, DOLINSKI K, DWIGHT SS, EPPIG JT, HARRIS MA, HILL DP, ISSEL-TARVER L, KASARSKIS A, LEWIS S, MATESE JC, RICHARDSON JE, RINGWALD M, RUBIN GM, SHERLOCK G. (2000,). Gene ontology: tool for the unification of biology. Nature Genetics. 1, 25-29.*

[4] *TETKOIV, RODCHENKOV IV, WALTER MC, RATTEI T, MEWES HW. (2008). Beyond the 'best' match: machine learning annotation of protein sequences by integration of different sources of information. Bioinformatics. 5, 621-628.*

[5] *PATHAK A, ROY T, EDUBILLI A, JAYARAM B. Mask blast with a new chemical logic of amino acids for improved protein function prediction[J]. Proteins: Structure, Function, and Bioinformatics, 2021, 89(8): 922-924.*

[6] *He Xinyuan, LIU Yang, Zeng Xianghe, GAO Rongfeng, Tian Zhen, Fan Xiangyu. (2024). Advances in bioinformatics-based protein function prediction. Chinese Journal of Biotechnology. 40(7), 2087-2099.(In Chinese)*

[7] *LiXinhui, QIANYurong, YUEHaitao, HUYue, CHENJiaying, LENG Hongyong, MAMengnan. (2023). Survey of Bioinformatics-Based Protein Function Prediction. Computer Engineering and Applications. 59(16), 50-60.(In Chinese)*

[8] *GIRI SJ, DUTTA P, HALANI P, SAHA S. MultiPredGO: deep multi-modal protein function prediction by amalgamating protein structure, sequence, and interaction information[J]. IEEE Journal of Biomedical and Health Informatics, 2021, 25(5): 1832-1838.*

[9] *KONDOHX, IIZUKA H, MASUMOTO G, KABAYA Y, KANEMATSU Y, TAKANO Y. Prediction of protein function from tertiary structure of the active site in heme proteins by convolutional neural network[J]. Biomolecules, 2023, 13(1): 137.*

[10] *Zhang FH, Song H, Zeng M, Li YH, KURGAN L, Li M. DeepFunc: a deep learning framework for accurate prediction of protein functions from protein sequences and interactions[J]. Proteomics, 2019, 19(12): e1900019.*

[11] *You RH, Yao SW, MAMITSUKA H, Zhu SF. DeepGraphGO: graph neural network for large-scale, multispecies protein function prediction[J]. Bioinformatics, 2021, 37(supplement_1): i262-i271.*