2024 Volume 2, Issue 1
Big Data Analytics for Healthcare
ISSN: 2630-4767 (Online)

Advances in Bioinformatics-Based Protein Function Prediction

Xinyuan He[1], Yang Liu[1], Xianghe Zeng[2], Rongfeng Gao[2], Zhen Tian[3], Xiangyu Fan[1,2]*

[1]School of Information Science and Engineering, University of Jinan, Jinan 250022, Shandong Province, China

[2]School of Biological Science and Technology, University of Jinan, Jinan 250022, Shandong Province, China

[3]Joint Laboratory for Translational Medicine Research, Liaocheng City People's Hospital, Liaocheng 252000, Shandong Province, China

***Corresponding author:** Xiangyu Fan, bio_fanxy@ujn.edu.cn

Abstract:

With the increasing computational power and the rapid expansion of biological data, the application of bioinformatics tools has emerged as the primary approach to tackling biological challenges. The precise determination of protein function through bioinformatics tools is pivotal for both biomedical research and drug discovery, making it a focal point of investigation. In this article, we classify bioinformatics-based protein function prediction methods into three main categories: methods based on protein sequences, methods based on protein structures, and methods based on protein interaction networks. We delve deeper into these specific algorithms, emphasizing recent research progress and offering invaluable insights for the utilization of bioinformatics-based protein function prediction in biomedical research and drug discovery.

Keywords:

Bioinformatics
Protein function prediction
Gene ontology
Machine learning
Deep learning

Online publication: June 28, 2024

1. Introduction

Proteins are the fundamental building blocks of life, composed of amino acids and exhibiting diverse structures and functions. They play crucial roles in biological systems, including enzymatic catalysis, structural support, signal transduction, and immune responses. Therefore, understanding protein functions is pivotal for unveiling biological processes, drug development, and disease research. Predicting protein functions not only aids in comprehending the interaction networks among proteins, further elucidating cell signaling and metabolic pathways but also provides more comprehensive information for genomics studies. It can even be used to identify potential drug targets or discover novel drug molecules.

Determining protein functions through traditional biochemical experiments is costly and time-consuming [1]. Hence, developing accurate and efficient protein function prediction methods is imperative. With the thriving

development of disciplines like machine learning and mathematical statistics, protein function prediction based on bioinformatics has emerged as a research hotspot [2-4]. These methods can significantly expedite the process of elucidating protein functions. Bioinformatics-based protein function prediction methods primarily utilize information hierarchies such as sequence, structure, evolution, and interactions, constructing models and algorithms to predict protein functions. These methods have evolved from initial sequence similarity searches to modern machine learning and deep learning techniques. In recent years, domestic and international researchers have continuously explored the field of protein function prediction, yielding abundant research outcomes [4]. This article summarizes current protein function prediction methods, analyzes and compares the advantages and limitations of various approaches, and reviews bioinformatics-based protein function prediction methodologies. It aims to acquaint readers with research trends and developments in this field. Additionally, it analyzes existing problems in this domain and delves into future challenges, intending to provide valuable references and insights for researchers in the field of protein function prediction.

2. Characteristics and classification of protein functions

Proteins are among the most crucial biomolecules in organisms, performing various functions such as constituting and repairing cells, transmitting signals, catalyzing chemical reactions, storing energy, and transporting substances. Hence, proteins are considered the fundamental functional units of living organisms. A single protein can participate in different functions within an organism. For instance, transferrin [5] not only serves as an efficient iron transporter, moving iron ions from the liver to the bone marrow in the bloodstream, fulfilling the role of a transport protein, but also binds to cell surface receptors, assisting in the delivery of iron ions to the cell interior, thus acting as a receptor. Additionally, transferrin exhibits catalytic properties, promoting the release and binding of iron ions, providing the driving force for iron transport. Evidently, the study of protein functions is extremely complex and diverse, with any event associated with proteins potentially considered a protein function [6].

To systematically and standardly describe protein functions, Gene Ontology (GO) [7,8] and Functional Categories (FunCat) [7] have been proposed. Currently, GO is one of the most widely accepted and commonly used classification systems for protein functions. The GO protein classification system comprises tens of thousands of terms, covering various functions and locations of proteins within cells and organisms. These terms are organized into a directed acyclic graph (DAG) format, enabling the visualization of hierarchical relationships and associations between protein functions. GO categorizes proteins into three main aspects: molecular function (MF), biological process (BP), and cellular component (CC). Molecular function primarily describes the specific functions exhibited by protein molecules, such as catalyzing reactions or binding to other molecules. Biological process focuses on describing events related to protein functions, including metabolic pathways, cell signaling, and more. Cellular component mainly describes the specific locations and structures where proteins are localized within cells, such as the nucleus, cell membrane, etc. Specifically, each GO term represents a functional label, and the process of predicting protein functions involves determining the labels possessed by proteins [9]. Proteins in databases like UniProt [1], Ensembl [10], and InterPro [11] are annotated with GO functional labels, facilitating the provision of functional annotations for protein sequences.

3. Overview of bioinformatics methods for protein function prediction

With the continuous advancement of genomics and proteomics technologies, researchers have been able to rapidly and efficiently identify protein-coding sequences within the genome. However, gaining a deep understanding of the functions of these proteins and their precise roles in biological processes remains a challenging task. To address this issue, researchers have conducted extensive studies using bioinformatics methods, aiming to predict protein functions and thereby enhance our understanding of biological systems.

The essence of protein function prediction lies in accurately determining the degree of similarity between

Table 1. Bioinformatics tools for protein function prediction based on sequence homology

Year of publication	Method	Model	Model evaluation criteria			Open source situation
			BP	CC	MF	
2020	DeepGOPlus [19]	CNN	F_{max} = 0.390	F_{max} = 0.614	F_{max} = 0.557	https://github.com/bio-ontology-research-group/deepgoplus
2021	NCL+mask BLAST [20]	BLAST+NCL	F-measure = 0.378	F-measure = 0.475	F-measure = 0.496	-

Note: "–" indicates that the author has not yet released the source code, and the following table is the same.

proteins with unknown functions and those with known functions, in terms of their sequences, functions, and other aspects. This involves a complex multi-label classification problem. In this context, this article categorizes protein function prediction methods into three main directions: methods based on protein sequences, methods based on protein structures, and methods based on protein interaction networks.

Firstly, protein sequence-based function prediction methods analyze the amino acid sequences of proteins and utilize similarities and features among sequences to infer possible protein functions. Secondly, protein structure-based function prediction methods focus on the three-dimensional structures of proteins, inferring their functions by simulating and predicting protein structures. Lastly, protein interaction network-based function prediction methods emphasize analyzing the interactions between proteins and other biomolecules to unveil their functions and roles within biological systems.

The integrated application of these methods provides crucial tools for gaining a deeper understanding of protein functions, helping to fill gaps that are difficult to cover through experiments. By comprehensively employing these bioinformatics methods, we can not only fully reveal the multifaceted functions of proteins in biological processes but also provide strong support for research and applications in the field of life sciences.

4. Methods for protein function prediction based on sequences

4.1. Methods based on protein sequence homology

Homology-based methods involve finding proteins with similar sequences to the target protein and assigning their functional annotations (such as protein function, domains, reaction mechanisms, and structural features) to the target. In these methods, the functional annotation of proteins with unknown functions relies on their similarity scores with known protein sequences. Commonly used sequence alignment techniques in this process include FASTA [12], BLAST [13], and PSI-BLAST [14]. These methods typically associate sequence similarity with functional similarity. Additionally, there are databases like the Evolutionary Genealogy of Genes: Non-Supervised Orthologous Groups (EggNOG) [15], which infer GO annotation information for entire gene families based on sequence similarity and homologous relationships. However, researchers have found that the principle of judging functional similarity based on sequence similarity is a relatively weak assumption [16]. While this method is simple to operate, it has several limitations, such as being restricted by the number of known functional sequences and having longer running times. Simultaneously, studies have indicated that 30% of protein functions obtained through this method are incorrect [17,18]. This has motivated researchers to explore alternative methods. Recent approaches for protein function prediction based on sequence homology include DeepGOPlus and NCL+mask BLAST (**Table 1**).

DeepGOPlus [19] is an improved version of DeepGO [21], overcoming the limitations of DeepGO in terms of sequence length, feature relationships, and prediction time. The DeepGOPlus method combines convolutional neural networks (CNNs) with sequence similarity-based predictions. This approach utilizes one-dimensional convolutional neural networks (1-D CNNs) to process multiple variable-size convolution kernels

simultaneously, while the CNN convolution kernels learn patterns similar to structural motifs. The data for the DeepGO model comes from the SwissProt database, filtered to include only experimentally verified proteins with no ambiguous amino acid codes in their sequences, resulting in a final dataset of 60,710 protein sequences with 27,760 GO term categories (19,181 BP, 6,221 MF, and 2,358 CC), covering over 90% of the annotated protein sequences in SwissProt. The dataset is divided randomly into an 80% training set and a 20% test set. Functional predictions are made using a method based on triplet (3mer) sequence feature extraction, and the test set is evaluated using the Computational Assessment of Function Annotation (CAFA)—a common standard for analyzing and evaluating the performance of protein function annotation methods [22]. The results show significant improvement compared to traditional BLAST methods, particularly in protein subcellular localization, but without significant improvements in MF and BP performance. DeepGOPlus innovates on the foundation of DeepGO by increasing the amino acid sequence length to 2,000 (covering over 99% of sequences in UniProt), removing restrictions on sequence length, and making it suitable for genome-scale annotation of protein functions, especially in newly sequenced organisms. Yang *et al.*, from the Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, used the DeepGOPlus model to identify 12 type I and four type II high-affinity potassium ion transporter genes (SaHKTs) from the salt-tolerant plant *Spartina alterniflora*. They further validated the functions of these 16 *HKT* genes through yeast complementation experiments. The experiments demonstrated that type I members SaHKT1;2, SaHKT1;3, and SaHKT1;8, and type II members SaHKT2;1, SaHKT2;3, and SaHKT2;4 have low-affinity K^+ absorption capabilities. Type II members showed a stronger affinity for K^+ than rice and *Arabidopsis*, and most *Spartina alterniflora* HKTs preferentially transport Na^+ [23]. This study applies a deep learning model to the discovery of plant functional genes, demonstrating its ability to identify characteristic sequences of functional genes from vast genomic datasets, accurately locate them, and analyze their properties. It proves the model's practicality in accurately predicting gene families and functional genes [23].

NCL+mask BLAST is a novel protein function prediction method proposed by Pathak *et al.* [20]. The basic logic combines a new amino acid classification method based on chemical measurements with BLAST. It uses the New Chemical Logic (NCL) [24,25] to filter out functionally unrelated protein sequences before aligning them with sequences in the database to obtain functional predictions. NCL treats amino acids as a complete set of chemical templates and proposes a new method based on their stereochemical properties. This method is over three times more accurate than BLAST in detecting biological, molecular, and cellular functions of all 69,306 known protein sequences in SwissProt. For example, while traditional BLAST methods cannot predict the molecular function of 1-deoxy-11-beta-hydroxypentanoate dehydrogenase, the NCL method can predict its molecular function as an oxidoreductase. A notable feature of this method is the use of NCL to filter out functionally unrelated protein sequences, thereby improving performance.

4.2. Methods based on protein feature extraction

Protein sequences are composed of permutations and combinations of 20 different amino acids. To be recognized by computers, these string sequences need to be converted into numerical forms. The numerical representation of protein sequences serves as features for machine learning models, and the process of converting protein sequences into their corresponding numerical forms is called feature extraction. Some scientists have designed tools for predicting protein function based on feature extraction methods (**Table 2**).

In recent years, learning to rank (LTR) has been effectively applied in bioinformatics, such as in predicting drug-target interactions. Learning to rank is a machine learning paradigm suitable for handling multi-label classification problems. You *et al.* [26] developed GOLabeler based on LTR, which is a framework that integrates different sequence information for automated function prediction (AFP). It integrates five component classifiers trained with different feature information such as GO term frequency, sequence alignment, amino acid triplets (3-mers), domains and motifs, and biophysical properties, within the LTR framework to achieve protein

Table 2. Bioinformatics tools utilizing feature extraction methods for protein function prediction

Year of publication	Method	Model evaluation criteria			Open source situation
		BP	CC	MF	
2018	GOLabeler [26]	F_{max} = 0.372 AUPR = 0.236	F_{max} = 0.586 AUPR = 0.697	F_{max} = 0.691 AUPR = 0.549	https://github.com/ddofer/ProFET
2019	GRNN [27]	Macro F1 for the brain cells dataset = 0.86 Macro F1 for the circulation cells dataset = 0.77 Macro F1 for the generic cells dataset = 0.70			-
2020	DeepAdd [28]	F_{max} = 0.345 AUC = 0.896	F_{max} = 0.547 AUC = 0.958	F_{max} = 0.516 AUC = 0.912	-
2020	FFPred-GAN [29]	F_{max} = 0.567	F_{max} = 0.755	F_{max} = 0.750	https://github.com/psipred/FFPredGAN
2022	FUTUSA [30]	F1 = 0.532	-	-	https://github.com/snuhl-crain/FUTUSA
2022	PFmulDL [31]	F_{max} = 0.459 AUPR = 0.452	F_{max} = 0.677 AUPR = 0.729	F_{max} = 0.508 AUPR = 0.509	https://github.com/idrblab/PFmulDL
2022	PFP-Autoencoders [32]	F_{max} = 0.422 AUPR = 0.400	-	F_{max} = 0.475 AUPR = 0.430	https://github.com/richadhanuka/PFP-Autoencoders/tree/main
2022	GAT-GO [33]	F_{max} = 0.492 AUPR = 0.381	F_{max} = 0.547 AUPR = 0.479	F_{max} = 0.633 AUPR = 0.660	-
2023	Global-ProtEnc [34]	F_{max} = 0.523	F_{max} = 0.636	F_{max} = 0.515	https://github.com/ashi24cc/Global-ProtEnc-Plus/tree/main

function prediction. Experimental results show that when using two large-scale datasets, CAFA1 [1] and CAFA2 [35], divided into training and testing sets according to certain criteria, GOLabeler has significant advantages compared to other models in predicting GO terms for proteins with unknown functions.

Additionally, due to the complexity of protein sequence data, ordinary neural network models cannot effectively extract information from protein sequences. To address this issue, Ko *et al.* [30] developed FUTUSA. This model applies Convolutional Neural Networks (CNNs) to segment sequences and extract features, dividing protein sequences into fragments of different sizes for model training. The advantage of this method is its ability to detect functional motifs and predict mutation sites. However, it focuses on specific target functions and binary classification, which poses certain limitations. A similar model was proposed by Wan *et al.* [29], introducing a new method for predicting protein functions based on sequences called DeepAdd. This method utilizes a training set derived from 558,590 protein sequences in SwissProt (as of April 24, 2018) and a test set from 130,787 protein sequences in CAFA3. It treats protein sequences as natural language and employs the CBOW model from Word2Vec to extract word vectors. Then, it learns features through two CNN models: one based on the protein-protein interaction (PPI) network and the other on sequence similarity profiles (SSP).

Overall, CNNs have certain advantages in feature extraction compared to ordinary neural networks, but they perform poorly when processing sequential data. To overcome this limitation, Xia *et al.* [31] proposed PFmulDL. This method uses a dataset of 67,888 protein sequences from the UniPort database, representing protein sequences with one-hot encoding. It combines CNN models with recurrent neural networks (RNNs) and introduces transfer learning, successfully annotating the functions of 5,825 protein families, making it one of the models with the broadest coverage of GO families. Simultaneously, researchers found that it achieved functional predictions for superfamily proteins without sacrificing the prediction performance of "major category" proteins.

Compared to traditional sequence analysis and feature extraction methods, graph neural networks

(GNNs) have unique characteristics in protein function prediction. By modeling the structure and interaction graphs of proteins, GNNs can capture relationships and structural information between proteins, enabling the inference and prediction of protein functions. Ioannidis *et al.* [27] proposed a general regression neural network (GRNN) that utilizes multi-relational graphs for semi-supervised learning, weighting different relationships with learnable parameters to predict protein functions. Furthermore, to comprehensively utilize both local and global information on proteins, Lai *et al.* [33] introduced GAT-GO. This method employs a combination of sequenced residue-residue contact maps and protein sequence embeddings based on graph neural networks for functional prediction, aiming to improve the accuracy and efficiency of protein function prediction. GAT-GO incorporates various features such as one-hot protein sequences, PSSM, HMM, and ESM-1b [36] embedding information. It utilizes protein structure information predicted by RaptorX [37] and generates embeddings using Facebook's ESM-1b. Experimental results [33] demonstrate that, in the PDB-mmseqs test set with sequence similarity below 15% between training and test proteins, the GAT-GO model achieved F_{max} scores of 0.508, 0.416, and 0.501 in the MFO, BPO, and CCO domains, respectively, and area under the recall curve (AUPRC) scores of 0.427, 0.253, and 0.411. In contrast, the traditional BLAST method obtained F_{max} scores of 0.117, 0.121, and 0.207, and AUPRC scores of 0.120, 0.120, and 0.163, indicating the superior performance of the GAT-GO model over traditional methods.

In existing protein databases, a significant number of proteins still lack functional labels, and traditional supervised learning methods are prone to limitations caused by this phenomenon, resulting in low prediction accuracy. Generative adversarial networks (GANs) offer a solution to address the issue of insufficient data. Wan *et al.* [29] proposed FFPred-GAN, which learns the distribution of protein features through an FFPred feature extractor. It employs a Wasserstein generative adversarial network with gradient penalty and successfully enhances the original training samples by generating synthetic samples, achieving higher accuracy in predicting all three domains of GO terms. Additionally, Ranjan *et al.* [34] introduced a multi-faceted network model called Global-ProtEnc. This model utilizes a multi-attention mechanism to correlate subsequences across different functional domains. It features a bidirectional attention mechanism layer to capture highly relevant protein segments and resolve semantic ambiguities, enabling subsequence classification and protein function prediction. Global-ProtEnc and its enhanced version, Global-ProtEnc-Plus, exhibit excellent performance on the benchmark CAFA3 dataset. Compared to DeepGOPlus, Global-ProtEnc-Plus achieved a 6.50% improvement in F_{max} for biological processes and a 1.90% improvement in cellular components [34].

5. Methods for protein function prediction based on structure

Methods for protein function prediction based on structure involve utilizing structural information of proteins, including structural similarity comparison, protein structure modeling, and model prediction. Some researchers have developed bioinformatics tools for protein function prediction based on protein structure (**Table 3**).

MultiPredGO [38] is a multi-modal method based on deep learning. Its basic idea is to utilize two different types of information: protein sequence and protein secondary structure, designing two CNN models for

Table 3. Bioinformatics tools utilizing structure for protein function prediction

Year of publication	Method	Model	Model evaluation criteria			Open source situation
			BP	CC	MF	
2020	MultiPredGO [38]	ResNet-50	F_{max} = 0.328 AUC = 0.817	F_{max} = 0.537 AUC = 0.851	F_{max} = 0.367 AUC = 0.910	https://github.com/SwagarikaGiri/Multi-PredGO
2023	CNN model [39]	CNN	-	-	-	-

feature extraction. To accelerate prediction efficiency, protein-protein interaction information is integrated to generate 256-dimensional knowledge graph embeddings. These extracted features are then used to train a hierarchical classification model for predicting protein functions. The novelty of this method lies in its use of a multi-modal approach to fuse multiple types of information and the utilization of ResNet-50 to extract 3D structures from the Protein Data Bank (PDB) for use as 2D voxels. The method was trained on a dataset of 11,536,998,210,741 proteins created by combining protein sequences and 3D protein structures. Comparison of the results with various single-modal and multi-modal protein function prediction methods, including INGA [40] and DeepGO, showed that MultiPredGO achieved better overall performance in terms of accuracy, F-measure, precision, and recall. Specifically, it improved by an average of 13.05% and 30.87% in CC and MF, respectively, compared to DeepGO. However, a limitation of this method is its relatively low accuracy in predicting protein biological processes [38].

The CNN model [39] is a structure-function prediction method based on convolutional neural networks. It is used to predict protein functions from the tertiary structure of active sites in heme proteins, studying the relationship between structure and function. This method converts the tertiary structure of heme-binding sites into the xy plane and divides the space into small cubic regions (voxels). Researchers collected 6,866 heme molecules from 3,206 different protein structure entries in the PDB, a public database that stores structural information on proteins, nucleic acids, and other biomolecules. The CNN model was used to learn the association between heme protein structure and function, with the output serving as category labels for protein functions [39]. By training the model on a large dataset of heme proteins, researchers were able to accurately predict heme protein functions. However, further improvement and development are needed to apply this method to the prediction of a large number of proteins with unknown functions. AlphaFold2 is a deep learning algorithm that predicts protein tertiary structure from amino acid sequences, accurately predicting the structure of heme-binding sites in heme proteins. If the challenge of predicting heme-binding sites from amino acid sequences can be overcome, it may be possible to directly predict protein functions using amino acid sequences of heme proteins through deep learning methods.

6. Methods for protein function prediction based on interaction networks

Methods for protein function prediction based on interaction networks involve utilizing the interactions between proteins and other biomolecules to infer protein functions. Some researchers have developed bioinformatics tools for protein function prediction based on these interaction networks (**Table 4**).

DeepFunc [41] is a deep learning framework capable of accurately predicting protein functions from protein sequences and network information. Specifically, DeepFunc transforms relevant feature information

Table 4. Bioinformatics tools using protein-protein interaction networks for protein function prediction

Year of publication	Method	Model evaluation criteria			Open source situation
		BP	CC	MF	
2017	DeepGO [21]	F_{max} = 0.395	F_{max} = 0.633	F_{max} = 0.470	https://github.com/bio-ontology-research-group/deepgo
2019	DeepFunc [41]	-	-	F_{max} = 0.540 AUC = 0.940	-
2020	Graph2GO [42]	F_{max} =0.490	F_{max} =0.686	F_{max} = 0.718	https://github.com/yanzhanglab/Graph2GO
2020	SDN2GO [43]	F_{max} = 0.361 AUPR = 0.203	F_{max} = 0.432 AUPR = 0.290	F_{max} = 0.561 AUPR = 0.471	https://github.com/Charrick/SDN2GO
2021	DeepGraphGO [44]	F_{max} = 0.327 AUPR = 0.194	F_{max} = 0.692 AUPR = 0.695	F_{max} = 0.623 AUPR = 0.543	https://github.com/yourh/DeepGraphGO

collected from InterPro tools, such as domains, families, and motifs associated with the input protein sequence, into a high-dimensional binary vector of 35,000 dimensions. It then uses two fully connected layers to reduce the dimensionality of this vector, obtaining a low-dimensional vector. Meanwhile, EggNOG [15] is employed to acquire functional connections, which are combined with interactions from the STRING [45] tool to construct a PPI network. Deepwalk [46] algorithm is then used to extract a comprehensive set of topological features from the underlying PPI network. This vector is further fused with the topological features to form a fully connected network, culminating in functional classification. In summary, DeepFunc utilizes neural networks to make accurate predictions from protein sequences and network-derived information. This method integrates topological features of the PPI network with subsequence-based features, primarily using deep learning techniques to efficiently simplify high-dimensional vectors extracted from InterProScan and combine them with topological features extracted from the PPI network for functional prediction. When tested on the CAFA3 dataset, DeepFunc achieved the highest area under the curve (AUC) (a value closer to 1.0 indicates higher authenticity of the detection method) compared to methods like DeepGO and FFPred3, with an AUC of 0.94.

Graph2GO [42] is a feed-forward neural network architecture based on a multimodal graph for predicting protein functions. It integrates multiple data types such as protein structure, sequence, subcellular location, and interaction networks. It employs variational graph auto-encoders (VGAE) and graph convolutional networks (GCN) to infer functions based on gene ontology. The model consists of two parts: an unsupervised graph representation model and a deep-learning neural network (DNN) classifier. Researchers converted and concatenated information on 15,133 human proteins, subcellular locations, and protein domain information screened from SwissProt into vectors. These vectors were then fed into 1,713,652 protein-protein interactions (PPIs) obtained from STRING and 843,212 protein interaction edges in the sequence similarity network (SSN). The embeddings generated from these two networks were concatenated and used as input for the DNN framework to predict protein functions [42]. Experimental results showed that Graph2GO achieved higher precision, recall, and F1 scores than the sequence-based BLAST method in CC, MF, and BP categories. Additionally, the method demonstrated good performance and robustness when tested on other species such as *Drosophila melanogaster* and mice. However, there is still room for improvement, such as considering the hierarchical relationships of GO terms.

The SDN2GO [43] model employs a Convolutional Neural Network (CNN) to learn, extract, and integrate features from protein sequences, protein domains, and PPI networks for protein function prediction. Specifically, it utilizes 13,704 human proteins and 6,623 yeast proteins annotated with 13,882 and 4,796 GO terms, respectively, obtained from the GOA database. PPI network information for humans and yeast, acquired through STRING, and protein domain information from Interpro serve as training data. Sub-models based on CNNs are used to extract features from protein sequences, PPI networks, and protein domains. A weighted classifier then receives output vectors from these three sub-models and, through training, learns and optimizes the weights of the features received by each GO classifier, enabling multi-label classification. Testing on CAFA showed AUCs of 0.917, 0.964, and 0.948 for BP, MF, and CC, respectively, surpassing BLAST, DeepGO, and NetGO [43]. A notable advantage of this method is its integration of general features of protein domains, such as type, quantity, and location information. However, it has limitations, including the encoding of protein sequences based only on 3-tuples, which may overlook longer-range sequence-related features. Future improvements could consider more comprehensive sequence encoding strategies to capture broader sequence information.

DeepGraphGO [44] is an end-to-end model based on Graph Neural Networks (GNNs). It integrates protein sequence and PPI network information and adopts a multi-species strategy for training, adapting to different species. Integrated as part of NetGO [47], it further enhances performance. The model's structure comprises an input layer, graph convolutional layers, and an output layer. The input layer receives a protein network graph G or a weighted adjacency matrix A, along with N binary feature vectors for proteins. The graph convolutional layers capture high-level information from graph

edges by updating node representation vectors. Node updates can be achieved through weighted averaging of neighbor node information or iterative updates via a multi-layer GCN network. The output layer uses a fully connected layer to predict GO term scores for each protein, mapping node representation vectors to GO term scores, indicating the probability of those functions being present in the protein. Experimental results showed that DeepGraphGO achieved the best performance in three aspects, particularly in BPO and CCO. For instance, in BPO, DeepGraphGO achieved the highest F_{max} (0.327), improving by 7.2% and 12.8% compared to Net-KNN [47] (0.305) and DeepGOPlus (0.290) [44], respectively. This result indicates that DeepGraphGO effectively integrates protein sequence and network information through graph neural networks. Overall, DeepGraphGO can fully utilize protein network information and protein feature vectors, capture high-level relationships between proteins through graph convolutional layers, and predict GO terms, namely protein functions, through the output layer.

7. Summary and outlook

In recent years, with the advancement of computing power and the rapid expansion of biological data, the use of bioinformatics and deep learning algorithms to address biological problems has garnered widespread interest in the scientific community. This is especially true in the field of protein function prediction, where bioinformatics is considered an indispensable tool. This article summarized and analyzed methods for predicting protein functions using bioinformatics, exploring their characteristics and limitations.

Despite progress, the field of protein function prediction still faces challenges and limitations. Firstly, current methods have limitations in predicting functions and modeling the complex relationships between proteins. Complex proteins often consist of multiple domains with intricate interactions and functional relationships, requiring more precise and accurate methods to decipher this complexity. Secondly, there are difficulties in processing multi-relational protein interaction network data. Existing methods often struggle to effectively handle and integrate different types of relationships, limiting comprehensive protein function prediction and understanding. Finally, there is a lack of targeted protein function prediction tools for different types of biological data. Our research team focuses on predicting diverse protein functions in the virome. In a recent study on predicting the functions of soil virus-assisted metabolism-related proteins [48], we could only compare identified proteins of unknown function to corresponding metabolic gene databases one by one, making the process time-consuming and labor-intensive. Therefore, there is a need to develop new methods that better utilize information from different data sources to address these issues.

To further enhance the accuracy and reliability of protein function prediction, improvements can be made in the following three aspects. Firstly, researchers can explore the integration of information from multiple data sources, including protein sequences, structures, and interactions. By comprehensively analyzing these data, the performance of prediction models can be improved. Secondly, researchers can delve deeper into understanding the decision-making process of deep learning models in protein function prediction, thereby enhancing the interpretability of the models. This will assist scientists in better comprehending prediction results and provide guidance for subsequent experimental design and molecular engineering. Additionally, prediction performance and accuracy can be enhanced by refining existing algorithms, seeking targeted training data, or developing novel models. This may involve the introduction of more advanced machine learning algorithms, optimization of feature selection and model training processes, as well as the incorporation of additional high-quality training data. Specifically, for protein function prediction in virome data, efforts can be focused on training, algorithm improvement, and model development using high-quality viral protein function data. Our team is currently engaged in this line of work. Lastly, protein function can be further explored through protein structure analysis. Although the limited availability of protein structures hinders the establishment of data-driven protein function prediction models based on the principle of structure determining function, the structures predicted by AlphaFold2 [49] offer a potential solution to this challenge. The core components of AlphaFold2, Evoformer and Structure Module, essentially rely on the attention mechanism. The underlying logic is

to identify the most relevant nodes for refinement within a complex graph structure, thereby reducing sample complexity. Researchers can attempt to incorporate the attention mechanism or combine it with convolutional and recurrent operations to further enhance function prediction through practical training. By obtaining more information about protein functions concurrently with predicting protein structures, the study of protein sequence-structure-function relationships can be further advanced.

In summary, significant progress has been made in the use of bioinformatics for protein function prediction. In the future, scientists can achieve more accurate and reliable protein function prediction results by comprehensively analyzing multiple data sources, enhancing model interpretability, and exploring novel algorithms and models.

Funding

National Natural Science Foundation of China (31600148); Natural Science Foundation of Shandong Province (ZR2021MC018)

Disclosure statement

The authors declare no conflict of interest.

References

[1] Boadu F, Cao H, Cheng J, 2023, Combining Protein Sequences and Structures with Transformers and Equivariant Graph Neural Networks to Predict Protein Function. Bioinformatics, 39(39 supplment 1): i318–i325.

[2] Yuan QM, Chen S, Rao JH, et al., 2022, AlphaFold2-Aware Protein-DNA Binding Site Prediction Using Graph Transformer. Briefings in Bioinformatics, 23(2): bbab564.

[3] Xia Y, Xia CQ, Pan XY, et al., 2021, GraphBind: Protein Structural Context Embedded Rules Learned by Hierarchical Graph Neural Networks for Recognizing Nucleic-Acid-Binding Residues. Nucleic Acids Research, 49(9): e51.

[4] Yuan QM, Chen JW, Zhao HY, et al., 2021, Structure-Aware Protein-Protein Interaction Site Prediction Using Deep Graph Convolutional Network. Bioinformatics, 38(1): 125–132.

[5] Guan WQ, 2019, Research Progress of Human Serum Transferrin Glycosylation. Laboratory Medicine, 34(6): 563–566.

[6] Rost B, Liu J, Nair R, et al., 2003, Automatic Prediction of Protein Function. Cellular and Molecular Life Sciences (CMLS), 60(12): 2637–2650.

[7] Ashburner M, Ball CA, Blake JA, et al., 2000, Gene Ontology: Tool for the Unification of Biology. Nature Genetics, 25(1): 25–29.

[8] Tetko IV, Rodchenkov IV, Walter MC, et al., 2008, Beyond the 'Best' Match: Machine Learning Annotation of Protein Sequences by Integration of Different Sources of Information. Bioinformatics, 24(5): 621–628.

[9] Teng ZX, Guo MZ, 2016, Research Progress on Protein Function Prediction Methods. Intelligent Computers and Applications, 6(4): 1–4, 8.

[10] Tiwari AK, Srivastava R, 2014, A Survey of Computational Intelligence Techniques in Protein Function Prediction. International Journal of Proteomics, 2014: 845479.

[11] Zhou NH, Jiang YX, Bergquist TR, et al., 2019, The CAFA Challenge Reports Improved Protein Function Prediction and New Functional Annotations for Hundreds of Genes Through Experimental Screens. Genome Biology, 20(1): 244.

[12] Lipman DJ, Pearson WR, 1985, Rapid and Sensitive Protein Similarity Searches. Science, 227(4693): 1435–1441.

[13] Altschul SF, Gish W, Miller W, et al., 1990, Basic Local Alignment Search Tool. Journal of Molecular Biology, 215(3): 403–410.

[14] Altschul SF, Madden TL, Schäffer AA, et al., 1997, Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. Nucleic Acids Research, 25(17): 3389–3402.

[15] Hernández-Plaza A, Szklarczyk D, Botas J, et al., 2023, eggNOG 6.0: Enabling Comparative Genomics Across 12,535 Organisms. Nucleic Acids Research, 51(D1): D389–D394.

[16] Ranjan A, Fahad MS, Fernandez-Baca D, et al., 2019, Deep Robust Framework for Protein Function Prediction Using Variable-Length Protein Sequences. ACM Transactions on Computational Biology and Bioinformatics, 2019: 1.

[17] Devos D, Valencia A, 2000, Practical Limits of Function Prediction. Proteins: Structure, Function, and Genetics, 41(1): 98–107.

[18] Devos D, Valencia A, 2001, Intrinsic Errors in Genome Annotation. Trends in Genetics, 17(8): 429–431.

[19] Kulmanov M, Hoehndorf R, 2020, DeepGOPlus: Improved Protein Function Prediction from Sequence. Bioinformatics, 36(2): 422–429.

[20] Pathak A, Roy T, Edubilli A, et al., 2021, Mask Blast with a New Chemical Logic of Amino Acids for Improved Protein Function Prediction. Proteins: Structure, Function, and Bioinformatics, 89(8): 922–924.

[21] Kulmanov M, Khan MA, Hoehndorf R, 2018, DeepGO: Predicting Protein Functions from Sequence and Interactions Using a Deep Ontology-Aware Classifier. Bioinformatics, 34(4): 660–668.

[22] Radivojac P, Clark WT, Oron TR, et al., 2013, A Large-Scale Evaluation of Computational Protein Function Prediction. Nature Methods, 10(3): 221–227.

[23] Yang MG, Chen SK, Huang ZP, et al., 2023, Deep Learning-Enabled Discovery and Characterization of *HKT* Genes in *Spartina alterniflora*. The Plant Journal: for Cell and Molecular Biology, 116(3): 690–705.

[24] Jayaram B, 2008, Decoding the Design Principles of Amino Acids and the Chemical Logic of Protein Sequences. Nature Precedings, 3: 1.

[25] Kaushik R, Singh A, Jayaram B, 2018, Where Informatics Lags Chemistry Leads. Biochemistry, 57(5): 503–506.

[26] You RH, Zhang ZH, Xiong Y, et al., 2018, GOLabeler: Improving Sequence-Based Large-Scale Protein Function Prediction by Learning to Rank. Bioinformatics, 34(14): 2465–2473.

[27] Ioannidis VN, Marques AG, Giannakis GB, 2020, Graph Neural Networks for Predicting Protein Functions, 2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), Guadeloupe, Le Gosier, 221–225.

[28] Du ZH, He YF, Li JQ, et al., 2020, DeepAdd: Protein Function Prediction from k-mer Embedding and Additional Features. Computational Biology and Chemistry, 89: 107379.

[29] Wan C, Jones DT, 2020, Protein Function Prediction is Improved by Creating Synthetic Feature Samples with Generative Adversarial Networks. Nature Machine Intelligence, 2(9): 540–550.

[30] Ko CW, Huh J, Park JW, 2022, Deep Learning Program to Predict Protein Functions Based on Sequence Information. MethodsX, 9: 101622.

[31] Xia WQ, Zheng LY, Fang JB, et al., 2022, PFmulDL: A Novel Strategy Enabling Multi-Class and Multi-Label Protein Function Annotation by Integrating Diverse Deep Learning Methods. Computers in Biology and Medicine, 145: 105465.

[32] Dhanuka R, Tripathi A, Singh JP, 2022, A Semi-Supervised Autoencoder-Based Approach for Protein Function Prediction. IEEE Journal of Biomedical and Health Informatics, 26(10): 4957–4965.

[33] Lai BQ, Xu JB, 2022, Accurate Protein Function Prediction Via Graph Attention Networks with Predicted Structure Information. Briefings in Bioinformatics, 23(1): bbab502.

[34] Ranjan A, Tiwari A, Deepak A, 2023, A Sub-Sequence Based Approach to Protein Function Prediction Via Multi-Attention Based Multi-Aspect Network. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 20(1): 94–105.

[35] Jiang YX, Oron TR, Clark WT, et al., 2016, An Expanded Evaluation of Protein Function Prediction Methods Shows an Improvement in Accuracy. Genome Biology, 17(1): 184.

[36] Rives A, Meier J, Sercu T, et al., 2021, Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. Proceedings of the National Academy of Sciences of the United States of America, 118(15): e2016239118.

[37] Xu JB, McPartlon M, Li J, 2021, Improved Protein Structure Prediction by Deep Learning Irrespective of Co-Evolution Information. Nature Machine Intelligence, 3(7): 601–609.

[38] Giri SJ, Dutta P, Halani P, et al., 2021, MultiPredGO: Deep Multi-Modal Protein Function Prediction by Amalgamating Protein Structure, Sequence, and Interaction Information. IEEE Journal of Biomedical and Health Informatics, 25(5): 1832–1838.

[39] Kondohx, Iizuka H, Masumoto G, et al., 2023, Prediction of Protein Function from Tertiary Structure of the Active Site in Heme Proteins by Convolutional Neural Network. Biomolecules, 13(1): 137.

[40] Piovesan D, Giollo M, Leonardi E, et al., 2015, INGA: Protein Function Prediction Combining Interaction Networks, Domain Assignments, and Sequence Similarity. Nucleic Acids Research, 43(W1): W134–W140.

[41] Zhang FH, Song H, Zeng M, et al., 2019, DeepFunc: A Deep Learning Framework for Accurate Prediction of Protein Functions from Protein Sequences and Interactions. Proteomics, 19(12): e1900019.

[42] Fan KJ, Guan YF, Zhang Y, 2020, Graph2GO: A Multi-Modal Attributed Network Embedding Method for Inferring Protein Functions. GigaScience, 9(8): giaa081.

[43] Cai YD, Wang JC, Deng L, 2020, SDN2GO: An Integrated Deep Learning Model for Protein Function Prediction. Frontiers in Bioengineering and Biotechnology, 8: 391.

[44] You RH, Yao SW, Mamitsuka H, et al., 2021, DeepGraphGO: Graph Neural Network for Large-Scale, Multispecies Protein Function Prediction. Bioinformatics, 37(supplement_1): i262–i271.

[45] Szklarczyk D, Franceschini A, Wyder S, et al., 2015, STRING v10: Protein-Protein Interaction Networks, Integrated Over the Tree of Life. Nucleic Acids Research, 43(D1): D447–D452.

[46] Perozzi B, Al-Rfou R, Skiena S, 2014, DeepWalk: Online Learning of Social Representations, Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, 701–710.

[47] You RH, Yao SW, Xiong Y, et al., 2019, NetGO: Improving Large-Scale Protein Function Prediction with Massive Network Information. Nucleic Acids Research, 47(W1): W379–W387.

[48] Ji MZ, Fan XY, Cornell CR, et al., 2023, Tundra Soil Viruses Mediate Responses of Microbial Communities to Climate Warming. mBio, 14(2): e0300922.

[49] Jumper J, Evans R, Pritzel A, et al., 2021, Highly Accurate Protein Structure Prediction with AlphaFold. Nature, 596(7873): 583–589.