生 物 工 程 学 报
Chinese Journal of Biotechnology
http://journals.im.ac.cn/cjbcn
DOI: 10.13345/j.cjb.230675

Jul. 25, 2024, 40(7): 2087-2099

·综 述·

基于生物信息学的蛋白质功能预测方法研究进展

何新媛[1]，刘杨[1]，曾祥荷[2]，高荣凤[2]，田真[3]，樊祥宇[1,2*]

1 济南大学 信息科学与工程学院，山东 济南 250022
2 济南大学 生物科学与技术学院，山东 济南 250022
3 聊城市人民医院 转化医学研究联合实验室，山东 聊城 252000

何新媛，刘杨，曾祥荷，高荣凤，田真，樊祥宇. 基于生物信息学的蛋白质功能预测方法研究进展[J]. 生物工程学报，2024, 40(7): 2087-2099.
HE Xinyuan, LIU Yang, ZENG Xianghe, GAO Rongfeng, TIAN Zhen, FAN Xiangyu. Advances in bioinformatics-based protein function prediction[J]. Chinese Journal of Biotechnology, 2024, 40(7): 2087-2099.

摘 要：随着计算能力的增加和生物数据的快速扩展，利用生物信息学解决一些生物学问题逐渐成为主流的解决方案。蛋白质功能预测是生物医学和药物研究领域的重要任务。利用生物信息学进行蛋白质功能预测成为研究热点。本文将基于生物信息学的蛋白质功能预测方法归纳为 3 类：基于蛋白质序列的方法、基于蛋白质结构的方法和基于蛋白质相互作用网络的方法，并进一步分析和总结了这些方法的具体算法以及最新研究进展，为生物医学和药物研究领域深入探索预测蛋白质功能提供重要参考。

关键词：生物信息学；蛋白质功能预测；基因本体论；机器学习；深度学习

Advances in bioinformatics-based protein function prediction

HE Xinyuan[1], LIU Yang[1], ZENG Xianghe[2], GAO Rongfeng[2], TIAN Zhen[3], FAN Xiangyu[1,2*]

1 School of Information Science and Engineering, University of Jinan, Jinan 250022, Shandong, China
2 School of Biological Science and Technology, University of Jinan, Jinan 250022, Shandong, China
3 Joint Laboratory for Translational Medicine Research, Liaocheng City People's Hospital, Liaocheng 252000, Shandong, China

Abstract: With the increasing of computer power and rapid expansion of biological data, the application of bioinformatics tools has become the mainstream approach to address biological problems. The accurate identification of protein function by bioinformatics tools is crucial for both biomedical research and drug discovery, making it a hot topic of research. In this paper, we categorize bioinformatics-based protein function prediction methods into three categories: protein

资助项目：国家自然科学基金(31600148)；山东省自然科学基金(ZR2021MC018)
This work was supported by the National Natural Science Foundation of China (31600148) and the Natural Science Foundation of Shandong Province (ZR2021MC018).
*Corresponding author. E-mail: bio_fanxy@ujn.edu.cn
Received: 2023-09-30; Accepted: 2023-12-13

sequence-based methods, protein structure-based methods, and protein interaction networks-based methods. We further analyze these specific algorithms, highlighting the latest research advancements and providing valuable references for the application of bioinformatics-based protein function prediction in biomedical research and drug discovery.
Keywords: bioinformatics; protein function prediction; gene ontology; machine learning; deep learning

蛋白质是生命的物质基础，它由氨基酸组成，具有多样的结构和功能。蛋白质在生物体内扮演着重要的角色，包括酶催化、结构支持、信号传递和免疫应答等。因此，了解蛋白质的功能对于揭示生物过程、药物开发和疾病研究都十分关键。对蛋白质进行功能预测不仅有助于理解蛋白质之间的相互作用网络，进一步理解细胞信号传导以及代谢途径，还可以为基因组学研究提供更全的信息，甚至能够用来识别潜在的药物靶或发现新的药物分子。

通过传统的生化实验来确定蛋白质的功能成本高且耗时长[1]，所以开发出准确高效的蛋白质功能预测方法是十分必要的。随着机器学习、数理统计等学科的蓬勃发展，基于生物信息学进行蛋白质功能预测已经成为研究热点[2-4]，这些方法可以大大加快解析蛋白质功能的速度。基于生物信息学的蛋白质功能预测方法主要利用序列、结构、进化和相互作用等信息层次，通过构建模型和算法来预测蛋白质的功能。这些方法从最初的序列相似性搜索发展到现在的机器学习和深度学习方法。近些年来，国内外科研人员在蛋白质功能预测领域不断探索，取得了丰富的研究成果[4]。本文对当下的蛋白质功能预测方法进行总结，分析和比较了各种方法的优点和局限性，综述了基于生物信息学的蛋白质功能预测方法研究，以便读者了解该领域的研究动态及发展趋势。与此同时，还分析了当前这一领域所存在的问题，深入探讨了未来面临的挑战，旨在为蛋白质功能预测领域的研究人员提供有价值的参考和启示。

1 蛋白质功能的特点和分类

蛋白质是生物体内最重要的生物大分子之一，具有构成和修复细胞、传递信号、催化化学反应、储存能量和运输物质等多种功能，因此蛋白质被认为是生物体的基本功能单位。同一种蛋白质可以参与生物体的不同功能，例如，转铁蛋白[5]不仅作为高效的铁运输者，在血液中将铁离子从生物体的肝脏转运到骨髓，起到运输蛋白的作用，同时与细胞表面受体结合，协助铁离子传递给细胞内部，起到受体的作用，此外，转铁蛋白还表现出催化剂的特性，促进铁离子的释放和结合过程，为铁的转运提供了动力。可见蛋白质功能的研究极其复杂、多样，任何与蛋白质相关联的事件都可以被认为是蛋白质的功能[6]。

为了对蛋白质功能进行系统和规范的描述，基因本体论(gene ontology, GO)[7-8]和基因分类标准(function categories, FunCat)[7]被提出。目前，GO是最被广泛接受和最为常用的蛋白质功能分类系统之一。GO蛋白质分类体系包括了数万个术语，涵盖了蛋白质在细胞和生物体内的各种功能和位置。这些术语被组织成有向无环图(directed acyclic graph, DAG)的形式，使得蛋白质功能之间的层次关系和关联能够被展示。GO可以将蛋白质分为3个主要方面：分子功能(molecular function, MF)、生物过程(biological process, BP)和细胞组件(cellular component, CC)。分子功能主要是对蛋白质分子所表现出的

特定功能进行描述，比如催化反应、结合其他分子等；生物过程主要是对蛋白质功能相关的事件进行描述，比如代谢途径、细胞信号传导等；细胞组件主要是对蛋白质在细胞中所定位的具体位置和结构进行描述，比如核、细胞膜等。具体来说，每一个 GO 术语是一个功能标签，对蛋白质进行功能预测的过程即为判断蛋白质所拥有标签的过程[9]。Uniprot[1]、Ensembl[10]和 InterPro[11]等数据库中的蛋白质都标注了 GO 功能标签，可以方便地为蛋白质序列提供功能注释。

2 蛋白质功能预测的生物信息学方法总览

随着基因组学和蛋白质组学技术的不断进步，科研人员已经能够迅速而高效地识别基因组中蛋白质编码序列。尽管如此，深入了解这些蛋白质的功能以及它们在生物过程中的确切作用仍然是一个充满挑战的任务。为了解决这一问题，科研人员通过生物信息学方法开展了广泛的研究，旨在预测蛋白质的功能，从而提升人们对生物系统的理解。

蛋白质功能预测的实质在于准确判定未知功能的蛋白和已知功能的蛋白在序列、功能等方面的相似程度，这涉及一个复杂的多标签分类问题。在这一背景下，本文将蛋白质功能预测的方法划分为 3 个主要方向：基于蛋白质序列进行功能预测的方法，基于蛋白质结构进行功能预测的方法，以及基于蛋白质相互作用网络进行功能预测的方法。

首先，基于蛋白质序列的功能预测方法通过分析蛋白质的氨基酸序列，利用序列之间的相似性和特征来推断蛋白质的可能功能。其次，基于蛋白质结构的功能预测方法主要关注蛋白质的三维结构，通过模拟和预测蛋白质结构来推断其功能。最后，基于蛋白质相互作用网络的功能预测方法则侧重于分析蛋白质与其他生物分子之间的相互作用，以揭示其在生物系统中的功能和作用。

这些方法的综合应用为深入了解蛋白质功能提供了重要工具，有助于填补实验难以覆盖的空白。通过综合运用这些生物信息学方法，不仅全面揭示了蛋白质在生物学过程中的多方面功能，并且为生命科学领域的研究和应用提供了强有力的支持。

3 基于序列进行蛋白质功能预测的方法

3.1 基于蛋白质序列同源性的方法

基于同源性的方法是指找到与待预测蛋白质序列相似的蛋白质，并将这一蛋白质的功能注释(蛋白质的功能、结构域、反应机制和结构特征等)赋予待预测蛋白质。在基于同源性的方法中，对未知功能蛋白质的功能注释依赖于其与已知蛋白质序列的相似性评分。这一过程经常使用的序列比对技术有 FASTA[12]、BLAST[13]和 PSI-BLAST[14]。这些方法通常将序列相似性与功能相似性联系起来。此外，还有基因的进化谱系：非监督直系同源群体[15](evolutionary genealogy of genes: non-supervised orthologous groups, EggNOG)等数据库，基于序列相似性和同源关系推测对整个基因家族的 GO 注释信息。但是，研究人员发现这种基于序列相似性判断功能相似性的原理是一个较弱的假设[16]。使用该种方法进行预测时，操作简单，但是存在较多的局限性，例如受已知功能序列数目的限制、运行时间较长等。同时，有研究指出，由这种方法得到的结果中有 30%的蛋白质功能是错误的[17-18]。这就促使研究人员开始探究别的方法。基于序列同源性进行蛋白质功能预测的最新方法有 DeepGOPlus 和 NCL+mask BLAST (表 1)。

表 1 基于序列同源性的蛋白质功能预测的生物信息学工具

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	–

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

DeepGOPlus[19]是 DeepGO[21]的改进版，其克服了 DeepGO 在序列长度、特征关系和预测时间方面的缺陷，DeepGOPlus 方法将卷积神经网络(convolutional neural networks, CNN)与基于序列相似性的预测相结合。该方法使用一维卷积神经网络(one-dimensional convolutional neural networks, 1-D CNN)同时处理具有多个大小可变的卷积核，同时 CNN 卷积核学习了类似于结构域基序的模式。DeepGO 模型的数据来自 SwissProt 数据库，筛选出经过实验验证且忽略序列中具有不明确氨基酸编码的数据，最终得到 60 710 种蛋白质序列，其中 GO 术语 27 760 个类别(BP 19 181 个，MF 6 221 个，CC 2 358 个)，涵盖了 SwissProt 数据库中 90%以上已注释的蛋白质序列。通过随机划分的方法将数据集的 80%作为训练集、20%作为测试集，使用基于三联体(3mer)序列特征提取的方法完成功能预测，并且将测试集通过函数标注的计算评估[22](computational assessment of function annotation, CAFA) (分析和评估蛋白质功能注释方法的性能的常用标准)进行评估，结果表明该方法与传统的 BLAST 方法相比有显著的改进，同时在蛋白质细胞定位方面拥有良好性能，但是没有改善在 MF 和 BP 方面的性能。DeepGOPlus 在 DeepGO 的基础上进行了创新，该模型在序列长度上将氨基酸序列增加到 2 000 个(覆盖了 UniProt 中 99%以上的序列)，对氨基酸序列的长度解除了限制，因此可以用于蛋白质功能的基因组规模注释，特别体现在新测序的生物体中。中国农业科学院作物科学研究所 Yang 等使用 DeepGOPlus 模型从盐生植物互花米草中挖掘出 12 个I型互花米草高亲和钾离子转运蛋白基因(spartina alterniflora high-affinity K^+ transporters, SaHKTs)和 4 个II型 SaHKTs[23]；并且他们进一步通过酵母互补实验验证了通过深度学习挖掘出的 16 个 *HKT* 基因的功能。实验表明，I 型成员 SaHKT1;2、SaHKT1;3 和 SaHKT1;8 以及II型成员 SaHKT2;1、SaHKT2;3 和 SaHKT2;4 具有低亲和力 K^+吸收能力，II型成员对 K^+的亲和力较水稻和拟南芥强，大多数互花米草 HKT 优先转运 Na^+[23]。该研究将深度学习模型应用到植物功能基因的挖掘，证明它能够从庞大的基因组数据中识别出功能基因的特征序列，并对其进行准确定位和特性分析，证明了该模型在准确预测基因家族和功能基因方面的实用性[23]。

NCL+mask BLAST 是 Pathak 等[20]提出的一种新的蛋白质功能预测方法。基本逻辑是，基于化学测量的新氨基酸分类方法，把这种新分类方法与 BLAST 进行结合，使用新化学逻辑(new chemical logic, NCL)[24-25]过滤功能不相关的蛋白质序列，之后与数据库中的序列进行比对，得到功能预测的结果。NCL 是把氨基酸看作是

一套完整的化学模板，根据其立体化学性质，提出的一种新方法。该方法在检测 SwissProt 中所有已知功能的 69 306 个蛋白序列的生物、分子和细胞功能方面的精度是 BLAST 的 3 倍以上，例如对于预测 1-去氧-11-β-羟基戊二酸脱氢酶的分子功能而言，传统的 BLAST 方法无法进行预测，而 NCL 方法能够预测其作为氧化还原酶的分子功能。该方法的一个显著特点是使用了 NCL 过滤掉功能上不相关的蛋白质序列，从而改善性能。

3.2 基于蛋白质特征提取的方法

蛋白质序列由 20 个不同的氨基酸排列组合而成，将这些字符串转换成数值形式才能被计算机识别。蛋白质序列的数值表示是机器学习模型的特征，将蛋白质序列转换为其对应的数值形式的过程被称为特征提取。有部分科学家基于特征提取的方法设计了蛋白质功能预测的工具(表 2)。

近年来，排序学习(learning to rank, LTR)已经有效地应用于生物信息学，例如药物-靶点相互作用的预测。排序学习是一种机器学习范式，适用于处理多标签分类问题。You 等[26]基于 LTR 开发出 GOLabeler，该方法是一个集成自动功能预测(automated function prediction, AFP)的不同序列信息的框架，它是在排序学习框架下集成 GO 项频率、序列比对、氨基酸三联体(3 mer)、结构域与基序以及生物物理属性等不同特征信息训练的 5 个成分分类器，实现了蛋白质功能的预测。实验结果表明将 CAFA1[1]和 CAFA2[35]两个大规模数据集按照一定标准得到训练集和测试集中，GOLabeler 与其他模型相比在对未知功能蛋白的 GO 术语项预测方面存在显著优势。

表 2 基于特征提取方法的蛋白质功能预测的生物信息学工具

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

此外，由于蛋白质序列数据的复杂性，普通的神经网络模型无法很好地提取蛋白质序列中的信息。针对该问题，Ko 等[30]开发出 FUTUSA。该模型应用 CNN 进行序列分割来提取特征，将蛋白质序列分割成大小不同的片段，并对模型进行训练。该方法的优点是可以检测功能基序和预测突变位点，但是该方法集中在特定的目标功能和二元分类上，存在一定的限制。与其使用相同模型的还有 Wan 等[29]提出的一种基于序列预测蛋白质功能的新方法——DeepAdd：该方法采用的训练集来自 SwissProt 中的 558 590 个蛋白序列(截至 2018 年 4 月 24 日)，测试集来自 CAFA3 的 130 787 个蛋白序列，它将蛋白质序列视为自然语言，并使用 Word2Vec 中的 CBOW 模型提取词向量；然后，通过两个 CNN 模型学习特征，其中一个 CNN 模型基于 PPI 模型的蛋白质-蛋白质相互作用(protein-protein interaction, PPI)网络，另一个模型基于序列相似性曲线(sequence similarity profiles, SSP)模型的序列相似性概要。总体而言，CNN 在普通神经网络的特征提取方面具有一定优势，但是在处理序列型的数据方面却表现欠佳。针对该问题，Xia 等[31]提出了 PFmulDL，该方法以来自 UniPort 数据库中的 67 888 个蛋白序列作为数据集，将蛋白质序列用独热编码表示，将 CNN 模型与循环神经网络(recurrent neural network, RNN)模型进行结合，引入迁移学习(transfer learning)，最终成功完成 5 825 个蛋白家族的功能注释，是覆盖 GO 家族最多的模型之一。与此同时，研究人员发现，在没有牺牲“主要类别”蛋白质的预测性能的前提下，实现了对超大家族蛋白质进行功能预测的功能。

相比于传统的序列分析和特征提取方法，图神经网络(graph neural networks, GNN)在蛋白质功能预测中有一些特别之处。它可以通过对蛋白质的结构和相互作用图进行建模，捕捉蛋白质之间的关系和结构信息，从而实现对蛋白质功能的推断和预测。Ioannidis 等[27]提出的广义回归神经网络(general regression neural network, GRNN)则是利用多关系图(multirelational graphs)进行半监督学习，通过可学习参数对不同关系进行加权并预测蛋白功能。此外，为了综合利用蛋白质的局部和全局信息，Lai 等[33]提出了 GAT-GO。该方法基于图神经网络利用测序的残基间接触图和蛋白质序列嵌入的组合进行功能预测，以提高蛋白质功能预测的准确性和效率。GAT-GO 使用的特征有 one-hot 蛋白序列、PSSM、HMM 和 ESM-1b[36]嵌入信息。GAT-GO 使用 RaptorX[37]预测的蛋白质的结构信息，并使用 Facebook 的 ESM-1b 生成其嵌入信息。实验结果表明[33]，在 PDB-mmseqs 测试集中，训练和测试蛋白质之间的序列相似性低于 15%的情况下，GAT-GO 模型在 MFO、BPO 和 CCO 领域上的 F_{max} 分别达到 0.508、0.416、0.501，在精确度-召回曲线下的面积(area under the recall curve, AUPRC)分别为 0.427、0.253、0.411；相比之下，传统的 BLAST 方法的 F_{max} 分别为 0.117、0.121、0.207，AUPRC 分别为 0.120、0.120、0.163，表明 GAT-GO 模型在性能上优于传统方法。

在现有的蛋白质数据库中，仍然存在大量蛋白质数据缺失功能标签的问题，而传统的监督学习方法容易受到该种现象的限制而导致预测的准确率不高。生成式对抗网络(generative adversarial network, GAN)则可以解决数据不足的问题。Wan 等[29]提出 FFPred-GAN，通过 FFPred 特征提取器学习蛋白质特征的分布，使用具有梯度罚分的 Wasserstein 生成式对抗网络，并通过生成合成样本来增强原始训练样本的方式，成功地在预测 GO 项的所有 3 个领域获得更高的准确性。此外，Ranjan 等[34]提出了一个多方面网络的模型 Global-ProtEnc。它使用多注意

力机制来关联不同功能域中的子序列，具有双向注意力机制层来捕捉高相关的蛋白质片段，并解决语义模糊性，从而实现子序列分类和蛋白质功能预测。Global-ProtEnc 和 Global-ProtEnc-Plus 方法在基准 CAFA3 数据集上的评估表现出色。与 DeepGOPlus 相比，Global-ProtEnc-Plus 在生物过程方面的 F_{max} 提高了 6.50%，在细胞组件方面提高了 1.90%[34]。

4 基于结构进行蛋白质功能预测的方法

该方法是指通过蛋白质的结构信息来进行功能预测，包括结构相似性比对、蛋白质结构模拟和模型预测等。已经有部分研究人员构建了基于蛋白质结构的蛋白质功能预测生物信息学工具(表 3)。

MultiPredGO[38]是一个基于深度学习的多模态方法，其基本思想是利用蛋白质序列和蛋白质二级结构两种不同的信息，设计两种 CNN 模型进行特征提取。为了加快预测效率，又集成了蛋白质相互作用信息，生成 256 维的知识图谱嵌入，之后利用这些提取的特征来训练预测蛋白质功能的层次分类模型。该方法的新颖之处是使用多模态的方法将多种信息融合，并使用 ResNet-50 将 3D 结构从蛋白质数据库(protein data bank, PDB)中提取出来用作 2D 体素。将蛋白质序列和蛋白质三维结构两种模式共同制作的 11 536 998 210 741 个蛋白质数据集进行训练后的结果与各种单模态以及 INGA[40]和 DeepGO 两种多模态蛋白质功能预测方法进行比较后的结果表明，在准确性、F-measure、精度和召回指标方面的整体性能均优于前人的方法；在 CC 和 MF 方面，与 DeepGO 相比，平均分别提高了 13.05%和 30.87%。但这一方法的缺陷是在预测蛋白质生物过程时准确率不高[38]。

CNN model[39]是一种基于卷积神经网络的结构-功能预测方法，用于从血红素蛋白的活性位点的三级结构中预测蛋白质功能，以研究结构与功能之间的关系。该方法通过将血红素结合位点的三级结构转换为 xy 平面，并将空间划分为小的立方体区域(体素)。研究者从 PDB (protein data bank，一个存储蛋白质、核酸和其他生物分子结构信息的公共数据库)中 3 206 个不同的蛋白质结构条目中收集 6 866 个血红素分子，使用 CNN 模型学习血红素蛋白结构与功能之间的关联，并将输出作为蛋白质功能的类别标签[39]。通过训练模型并使用大量的血红素蛋白数据，研究者能够准确地预测血红素蛋白的功能。然而，这种方法仍需要进一步改进和开发，以适用于大量未知功能蛋白质的预测。AlphaFold2 是一种用氨基酸序列预测蛋白质三级结构的深度学习算法，可以准确预测血红素蛋白中血红素结合位点的结构。如果能够克服从氨基酸序列中预测血红素结合位点的挑战，就可以利用血红素蛋白的氨基酸序列通过深度学习的方法直接预测蛋白质的功能。

表 3 基于结构的蛋白质功能预测的生物信息学工具

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	–	–	–	–

5 基于相互作用网络进行蛋白质功能预测的方法

该方法是指运用蛋白质与其他生物分子之间的互作关系来推断蛋白质的功能。已经有部分研究人员构建了基于相互作用网络的蛋白质功能预测方法的生物信息学工具(表 4)。

DeepFunc[41]是一种能够从蛋白质序列和网络信息中准确预测蛋白质功能的一种深度学习框架。具体来说，DeepFunc 是将从 InterPro 工具收集到的与输入蛋白质序列相关联的结构域、家族和基序的相关特征信息转变为一个维度为 35 000 维的高维二进制向量，然后使用两个全连接层对该向量进行降维，获取到一个低维向量。与此同时使用 EggNOG[15]获取功能性连接并且与 STRING[45]工具中的相互作用结合构建 PPI 网络。之后使用 Deepwalk[46]算法来提取底层 PPI 网络的全面拓扑特征集合，该向量再与该拓扑特征进行融合，形成全连接网络，最后进行功能分类。总的来说，DeepFunc 使用神经网络从蛋白质序列和网络派生的信息中做出准确的预测，该方法结合了 PPI 网络的拓扑特征和基于子序列的特征，主要使用深度学习技术高效地对从 InterProScan 中提取的高维向量进行简化，再与 PPI 网络中提取的拓扑特征相结合之后进行功能预测。最终在 CAFA3 数据集上进行测试，与 DeepGO、FFPred3 等方法相比，曲线下面积(area under curve, AUC) (该值越接近 1.0，检测方法真实性越高)达到了最高(0.94)。

Graph2GO[42]是一种基于多模态图的前馈神经网络架构，用于预测蛋白质功能。它整合了蛋白质结构、序列、亚细胞位置和相互作用网络等多种数据类型，并利用变分图自编码器(variational graph auto-encoders, VGAE)和图卷积神经网络(graph convolutional network, GCN)在基因本体上进行功能推断。该模型由两部分组成：无监督图表示模型和深度神经网络(deep-learning neural network, DNN)分类器。研究者从 SwissProt 中筛选出的 15 133 个人类蛋白质、亚细胞位置以及蛋白质结构域信息转换并拼接为向量送入 STRING 中获取的 1 713 652 个蛋白质相互作用网络(protein-protein interactions, PPIs)以及在蛋白质序列相似性网络(sequence similarity network, SSN)中的 843 212 个蛋白质相互作用边缘，将这两个网络生成的嵌入串联作为 DNN 框架的输入，以预测蛋白质的功能[42]。最终对比实验结果表明，Graph2GO 在 CC、MF 和 BP 下的精度、召回率以及 F1 分数都高于基于序列的

表 4 基于相互作用网络的蛋白质功能预测的生物信息学工具

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

BLAST 的方法。同时，该方法在测试果蝇、小鼠等其他物种时，也取得了较好的性能，可以看出其鲁棒性较强。但该方法仍有改进的空间，例如考虑 GO 术语层次关系的问题。

SDN2GO[43]模型是一种利用卷积神经网络来学习、提取和整合蛋白序列、蛋白质域和 PPI 网络等特征，并用于蛋白质功能预测的方法。具体而言，从 GOA 数据库得到 13 882 个注释蛋白的 13 704 个人类蛋白质和 4 796 个注释蛋白质的 6 623 个酵母蛋白，通过 STRING 获取的人类和酵母的蛋白网络信息以及从 Interpro 获得的蛋白质域信息作为训练数据，通过使用基于卷积神经网络的子模型来提取蛋白质序列信息、PPI 网络信息和蛋白质域的特征。随后，权重分类器接收来自这 3 个子模型的输出向量，并通过训练，学习和优化每个 GO 分类器接收到的特征的权重，以实现对多标签分类的效果。最终在 CAFA 上进行测试，结果表明 BP、MF 和 CC 上的 AUC 分别为 0.917、0.964、0.948，是 BLAST、DeepGO 和 NetGO 没有达到的[43]。该方法的显著优点是整合了蛋白质域的通用特征，如类型、数量和位置信息等。然而，该方法也存在一些缺陷，例如蛋白质序列仅基于 3 元组进行编码，这可能会忽略一些更长范围的序列相关特征。因此，在进一步改进此方法时，可以考虑更全面的序列编码策略来捕获更广泛的序列信息。

DeepGraphGO[44]是一个基于图神经网络的端到端的模型。该模型集成了蛋白质序列和蛋白质网络信息，并采用多物种策略进行训练，以适应不同的物种。它作为 NetGO[47]的一部分进行集成，进一步提高了性能。该模型的结构包括输入层、图卷积层和输出层。输入层接收蛋白质网络图 G 或加权邻接矩阵 A 以及蛋白质的 N 个二进制特征向量。图卷积层通过更新节点的表示向量来捕捉图边的高级信息。节点的更新可以通过加权平均邻居节点的信息或多层 GCN 网络进行迭代更新的方式实现。输出层使用全连接层来预测每个蛋白质的 GO 项得分，将节点的表示向量映射到 GO 项的得分，表示该蛋白质可能具有这些功能的概率。在最终的实验结果中发现，DeepGraphGO 在 3 个方面都实现了最佳性能，尤其在 BPO 和 CCO 方面。例如，DeepGraphGO 在 BPO 中实现了最高的 F_{max}(0.327)，比 Net-KNN[47](0.305)和 DeepGOPlus (0.290)分别提高了 7.2%和 12.8%[44]。这一结果表明，DeepGraphGO 通过图神经网络有效整合了蛋白质序列和网络信息。总体而言，DeepGraphGO 能够充分利用蛋白质网络信息和蛋白质特征向量，通过图卷积层捕捉蛋白质之间的高级关系，并通过输出层预测 GO 项，即蛋白质功能。

6 总结与展望

近年来，随着计算能力的提升和生物数据的快速扩展，利用生物信息学和深度学习算法解决生物学问题在科学界引起了广泛兴趣。特别是在蛋白质功能预测领域，生物信息学被认为是不可或缺的工具。本文对利用生物信息学预测蛋白质功能的方法进行了归纳整理，并分析了相关方法的特点和局限性。

尽管已经取得了一些进展，蛋白质功能预测领域仍然面临着一些挑战和限制。首先，目前的方法在功能预测和建模复杂蛋白质之间的关联性方面仍有局限性。复杂蛋白质往往由多个结构域组成，不同结构域之间的相互作用和功能关系复杂多样，因此需要更加精确和准确的方法来解析这种复杂性。其次，在处理多关系蛋白质相互作用网络数据时也存在一定困难。现有方法往往难以有效地处理和整合不同类型的关系，这限制了对蛋白质功能的全面预测和理解。最后，针对不同种类的生物数据，缺乏针对性的蛋白质功能

预测工具。本研究团队主要关注病毒组中不同蛋白质功能的预测，在最近进行的土壤病毒辅助代谢相关蛋白质功能预测的研究中[48]，只能将鉴定得到的未知功能蛋白质和相应代谢基因数据库进行一一比对，以此预测蛋白质功能，过程费时费力。因此，需要开发新的方法，更好地利用不同数据源的信息来处理这些问题。

为了进一步提高蛋白质功能预测的准确性和可靠性，可以从以下 3 个方面进行改进。首先，可以探索如何融合多种数据源的信息，包括蛋白质序列、结构和相互作用等多个方面的数据。通过综合分析这些信息，来提高预测模型的性能。其次，研究人员可以深入研究和理解深度学习模型在蛋白质功能预测中的决策过程，提高模型的可解释性。这将有助于科学家们更好地理解预测结果，并为进一步的实验设计和分子工程提供指导。此外，可以通过改进现有算法、寻找针对性的训练数据或者开发新的模型来提高预测性能和准确性，包括引入更先进的机器学习算法、优化特征选择和模型训练过程，以及增加更多高质量的训练数据等。特别是针对病毒组数据中蛋白质功能的预测，可以专门针对高质量的病毒蛋白质功能数据，进行训练、改进算法并开发新的模型。本团队正在尝试进行这部分工作。最后，可以进一步通过蛋白结构解析对蛋白质功能进行探索。虽然蛋白质结构的数量不足问题限制了通过利用结构决定功能的特性建立数据驱动的蛋白质功能预测模型，但是 AlphaFold2[49]预测的结构给这种情况提供了一个解决问题的角度，AlphaFold2 的核心部件 Evoformer 和 Structure Module，而它们的本质是注意力机制(attention mechanism)，其背后的逻辑是对于一个复杂的图结构，找到最相关的结点进行精化，从而降低样本复杂度。研究人员可以尝试引入注意力机制方法，或者将注意力机制与卷积操作、循环操作结合使用，通过实际的训练进一步完成功能预测。在预测蛋白质结构的同时，获取更多关于蛋白质功能的信息，也可以进一步推动蛋白质序列-结构-功能关系的研究。

总之，利用生物信息学进行蛋白质功能预测的研究已经取得较多进展。未来科学家们可以通过综合分析多种数据源、提高模型的可解释性，并探索新的算法和模型，以取得更加准确和可靠的蛋白质功能预测结果。

REFERENCES

[1] BOADU F, CAO H, CHENG J. Combining protein sequences and structures with transformers and equivariant graph neural networks to predict protein function[J]. Bioinformatics, 2023, 39(39 supplment 1): i318-i325.

[2] YUAN QM, CHEN S, RAO JH, ZHENG SJ, ZHAO HY, YANG YD. AlphaFold2-aware protein-DNA binding site prediction using graph transformer[J]. Briefings in Bioinformatics, 2022, 23(2): bbab564.

[3] XIA Y, XIA CQ, PAN XY, SHEN HB. GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues[J]. Nucleic Acids Research, 2021, 49(9): e51.

[4] YUAN QM, CHEN JW, ZHAO HY, ZHOU YQ, YANG YD. Structure-aware protein-protein interaction site prediction using deep graph convolutional network[J]. Bioinformatics, 2021, 38(1): 125-132.

[5] 关雯倩. 人血清转铁蛋白糖基化研究进展[J]. 检验医学, 2019, 34(6): 563-566.
GUAN WQ. Research progress of human serum transferrin glycosylation[J]. Laboratory Medicine, 2019, 34(6): 563-566 (in Chinese).

[6] ROST B, LIU J, NAIR R, WRZESZCZYNSKI KO, OFRAN Y. Automatic prediction of protein function[J]. Cellular and Molecular Life Sciences CMLS, 2003, 60(12): 2637-2650.

[7] 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. Gene ontology: tool for the unification of biology[J]. Nature Genetics, 2000, 25(1): 25-29.

[8] TETKOIV, RODCHENKOV IV, WALTER MC, RATTEI T, MEWES HW. Beyond the 'best' match: machine learning annotation of protein sequences by integration of different sources of information[J]. Bioinformatics, 2008, 24(5): 621-628.

[9] 滕志霞, 郭茂祖. 蛋白质功能预测方法研究进展[J]. 智能计算机与应用, 2016, 6(4): 1-4, 8.
TENG ZX, GUO MZ. A survey on computational methods of predicting protein functions[J]. Intelligent Computer and Applications, 2016, 6(4): 1-4, 8 (in Chinese).

[10] TIWARI AK, SRIVASTAVA R. A survey of computational intelligence techniques in protein function prediction[J]. International Journal of Proteomics, 2014, 2014: 845479.

[11] ZHOU NH, JIANG YX, BERGQUIST TR, LEE AJ, KACSOH BZ, CROCKER AW, LEWIS KA, GEORGHIOU G, NGUYEN HN, HAMID MN, DAVIS L, DOGAN T, ATALAY V, RIFAIOGLU AS, DALKıRAN A, CETIN ATALAY R, ZHANG CX, HURTO RL, FREDDOLINO PL, ZHANG Y, BHAT P, et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens[J]. Genome Biology, 2019, 20(1): 244.

[12] LIPMAN DJ, PEARSON WR. Rapid and sensitive protein similarity searches[J]. Science, 1985, 227(4693): 1435-1441.

[13] ALTSCHUL SF, GISH W, MILLER W, MYERS EW, LIPMAN DJ. Basic local alignment search tool[J]. Journal of Molecular Biology, 1990, 215(3): 403-410.

[14] ALTSCHUL SF, MADDEN TL, SCHÄFFER AA, ZHANG JH, ZHANG Z, MILLER W, LIPMAN DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs[J]. Nucleic Acids Research, 1997, 25(17): 3389-3402.

[15] HERNÁNDEZ-PLAZA A, SZKLARCZYK D, BOTAS J, CANTALAPIEDRA CP, GINER-LAMIA J, MENDE DR, KIRSCH R, RATTEI T, LETUNIC I, JENSEN LJ, BORK P, von MERING C, HUERTA-CEPAS J. eggNOG 6.0: enabling comparative genomics across 12 535 organisms[J]. Nucleic Acids Research, 2023, 51(D1): D389-D394.

[16] RANJAN A, FAHAD MS, FERNANDEZ-BACA D, DEEPAK A, TRIPATHI S. Deep robust framework for protein function prediction using variable-length protein sequences[J]. ACM Transactions on Computational Biology and Bioinformatics, 2019: 1.

[17] DEVOS D, VALENCIA A. Practical limits of function prediction[J]. Proteins: Structure, Function, and Genetics, 2000, 41(1): 98-107.

[18] DEVOS D, VALENCIA A. Intrinsic errors in genome annotation[J]. Trends in Genetics, 2001, 17(8): 429-431.

[19] KULMANOV M, HOEHNDORF R. DeepGOPlus: improved protein function prediction from sequence[J]. Bioinformatics, 2020, 36(2): 422-429.

[20] 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.

[21] KULMANOV M, KHAN MA, HOEHNDORF R. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier[J]. Bioinformatics, 2018, 34(4): 660-668.

[22] RADIVOJAC P, CLARK WT, ORON TR, SCHNOES AM, WITTKOP T, SOKOLOV A, GRAIM K, FUNK C, VERSPOOR K, BEN-HUR A, PANDEY G, YUNES JM, TALWALKAR AS, REPO S, SOUZA ML, PIOVESAN D, CASADIO R, WANG Z, CHENG JL, FANG H, et al. A large-scale evaluation of computational protein function prediction[J]. Nature Methods, 2013, 10(3): 221-227.

[23] YANG MG, CHEN SK, HUANG ZP, GAO S, YU TX, DU TT, ZHANG H, LI X, LIU CM, CHEN SH, LI HH. Deep learning-enabled discovery and characterization of *HKT* genes in *Spartina alterniflora*[J]. The Plant Journal: for Cell and Molecular Biology, 2023, 116(3): 690-705.

[24] JAYARAM B. Decoding the design principles of amino acids and the chemical logic of protein sequences[J]. Nature Precedings, 2008, 3: 1-1.

[25] KAUSHIK R, SINGH A, JAYARAM B. Whereinformatics lags chemistry leads[J]. Biochemistry, 2018, 57(5): 503-506.

[26] YOU RH, ZHANG ZH, XIONG Y, SUN FZ, MAMITSUKA H, ZHU SF. GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank[J]. Bioinformatics, 2018, 34(14): 2465-2473.

[27] IOANNIDIS VN, MARQUES AG, GIANNAKIS GB. Graph neural networks for predicting protein functions[C]//2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP). Le gosier, Guadeloupe. 2020: 221-225.

[28] DU ZH, HE YF, LI JQ, UVERSKY VN. DeepAdd: protein function prediction from k-mer embedding and additional features[J]. Computational Biology and Chemistry, 2020, 89: 107379.

[29] WAN C, JONES DT. Protein function prediction is improved by creating synthetic feature samples with generative adversarial networks[J]. Nature Machine Intelligence, 2020, 2(9): 540-550.

[30] KO CW, HUH J, PARK JW. Deep learning program to predict protein functions based on sequence information[J]. MethodsX, 2022, 9: 101622.

[31] XIA WQ, ZHENG LY, FANG JB, LI FC, ZHOU Y, ZENG ZY, ZHANG B, LI ZR, LI HL, ZHU F. PFmulDL: a novel strategy enabling multi-class and multi-label protein function annotation by integrating diverse deep learning methods[J]. Computers in Biology and Medicine, 2022, 145: 105465.

[32] DHANUKA R, TRIPATHI A, SINGH JP. A semi-supervised autoencoder-based approach for protein function prediction[J]. IEEE Journal of Biomedical and Health Informatics, 2022, 26(10): 4957-4965.

[33] LAI BQ, XU JB. Accurate protein function prediction *via* graph attention networks with predicted structure information[J]. Briefings in Bioinformatics, 2022, 23(1): bbab502.

[34] RANJAN A, TIWARI A, DEEPAK A. A sub-sequence based approach to protein function prediction *via* multi-attention based multi-aspect network[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2023, 20(1): 94-105.

[35] JIANG YX, ORON TR, CLARK WT, BANKAPUR AR, D'ANDREA D, LEPORE R, FUNK CS, KAHANDA I, VERSPOOR KM, BEN-HUR A, KOO DACE, PENFOLD-BROWN D, SHASHA D, YOUNGS N, BONNEAU R, LIN A, SAHRAEIAN SM, MARTELLI PL, PROFITI G, CASADIO R, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy[J]. Genome Biology, 2016, 17(1): 184.

[36] RIVES A, MEIER J, SERCU T, GOYAL S, LIN ZM, LIU J, GUO DM, OTT M, LAWRENCE ZITNICK C, MA J, FERGUS R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences[J]. Proceedings of the National Academy of Sciences of the United States of America, 2021, 118(15): e2016239118.

[37] XU JB, McPARTLON M, LI J. Improved protein structure prediction by deep learning irrespective of co-evolution information[J]. Nature Machine Intelligence, 2021, 3(7): 601-609.

[38] 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.

[39] 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.

[40] PIOVESAN D, GIOLLO M, LEONARDI E, FERRARI C, TOSATTO SCE. INGA: protein function prediction combining interaction networks, domain assignments and sequence similarity[J]. Nucleic Acids Research, 2015, 43(W1): W134-W140.

[41] 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.

[42] FAN KJ, GUAN YF, ZHANG Y. Graph2GO: a multi-modal attributed network embedding method for inferring protein functions[J]. GigaScience, 2020, 9(8): giaa081.

[43] CAI YD, WANG JC, DENG L. SDN2GO: an integrated deep learning model for protein function

prediction[J]. Frontiers in Bioengineering and Biotechnology, 2020, 8: 391.

[44] 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.

[45] SZKLARCZYK D, FRANCESCHINI A, WYDER S, FORSLUND K, HELLER D, HUERTA-CEPAS J, SIMONOVIC M, ROTH A, SANTOS A, TSAFOU KP, KUHN M, BORK P, JENSEN LJ, von MERING C. STRING v10: protein-protein interaction networks, integrated over the tree of life[J]. Nucleic Acids Research, 2015, 43(D1): D447-D452.

[46] PEROZZI B, AL-RFOU R, SKIENA S. DeepWalk: online learning of social representations[C]// Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. New York, USA. 2014: 701-710.

[47] YOU RH, YAO SW, XIONG Y, HUANG XD, SUN FZ, MAMITSUKA H, ZHU SF. NetGO: improving large-scale protein function prediction with massive network information[J]. Nucleic Acids Research, 2019, 47(W1): W379-W387.

[48] JI MZ, FAN XY, CORNELL CR, ZHANG Y, YUAN MM, TIAN Z, SUN KL, GAO RF, LIU Y, ZHOU JZ. Tundra soil viruses mediate responses of microbial communities to climate warming[J]. mBio, 2023, 14(2): e0300922.

[49] JUMPER J, EVANS R, PRITZEL A, GREEN T, FIGURNOV M, RONNEBERGER O, TUNYASUVUNAKOOL K, BATES R, ŽÍDEK A, POTAPENKO A, BRIDGLAND A, MEYER C, KOHL SAA, BALLARD AJ, COWIE A, ROMERA-PAREDES B, NIKOLOV S, JAIN R, ADLER J, BACK T, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596(7873): 583-589.

(本文责编　陈宏宇)