

# 基于深度学习的蛋白质从头测序方法研究进展

荆 昊, 贾晓田, 柳 楠

(山东建筑大学 计算机科学与技术学院, 山东 济南 250101)

**摘要:** 蛋白质组学的核心任务之一为揭示蛋白质的氨基酸序列。蛋白质从头测序方法可直接分析串联质谱数据, 无需依赖蛋白质数据库信息, 进而鉴定新蛋白质并提供翻译后的修饰信息, 在蛋白质组学研究领域备受青睐。DeepNovo系列、Denovo-GCN、GraphNovo等模型凭借深度学习的强大能力, 为蛋白质从头测序提供了多样化且创新的方法, 有效开辟了提高蛋白质鉴定准确性与效率的新途径, 为生物信息学研究提供了更为精准可靠的工具。尽管蛋白质从头测序方法具有独特优势, 串联质谱数据的复杂性和方法本身的局限性使鉴定准确性受到一定程度的影响。深入总结蛋白质从头测序方法的发展历程, 探讨基于深度学习的新兴方法的优势与局限性, 以便充分利用深度学习在分析串联质谱信息方面的优势不断优化算法和模型, 进一步提升蛋白质鉴定的准确性。

**关键词:** 蛋白质从头测序; 蛋白质组学; 串联质谱技术; 深度学习

DOI: 10.11907/rjtk.241459

中图分类号: TP301

文献标识码: A

开放科学(资源服务)标识码(OSID):

文章编号: 1672-7800(2025)008-0011-06



## Research Progress on Protein De Novo Sequencing Methods Based on Deep Learning

JING Hao, JIA Xiaotian, LIU Nan

(School of Computer Science and Technology, Shandong Jianzhu University, Ji'nan 250101, China)

**Abstract:** One of the core tasks of proteomics is to reveal the amino acid sequence of proteins. It has become the dominant means of large-scale protein research. The protein de novo sequencing method can directly analyze tandem mass spectrometry data without relying on protein database information. Due to its ability to recognize new proteins and provide post-translational modification information, it is highly favored. The fusion of DeepNovo series, Denovo GCN, GraphNovo and other models provides diverse and innovative methods for de novo peptide sequencing of proteins. Combining the powerful capabilities of deep learning and the needs of proteomics, it opens up new ways to improve the accuracy and efficiency of protein identification, providing more accurate and reliable tools for bioinformatics research. Although de novo sequencing methods for proteins have unique advantages, their identification accuracy is still affected to some extent due to the complexity of tandem mass spectrometry data and the limitations of the method itself. After summarizing the development process of protein de novo sequencing methods in depth, explore the advantages and limitations of emerging methods based on deep learning. These methods fully utilize the advantages of deep learning in analyzing tandem mass spectrometry information. By continuously optimize algorithm design and deep learning models, in order to achieve the goal of improving the accuracy of protein identification.

**Key Words:** protein de novo sequencing; proteomics; tandem mass spectrometry technology; deep learning

## 0 引言

蛋白质是生命活动的基本执行者,其序列、结构和功能的研究对于理解生物体的复杂性至关重要<sup>[1]</sup>。传统的

蛋白质测序方法,如Edman降解法和质谱法,虽然在蛋白质序列分析方面具有一定效果,但存在操作复杂、样品需求量大、纯度要求高、耗时长和通量低等局限性<sup>[2-3]</sup>。随着生物信息学和测序技术的快速发展,蛋白质序列数据呈现爆炸式增长,开发新的高效测序技术迫在眉睫<sup>[4]</sup>。

收稿日期: 2024-05-14

扫描二维码阅读全文:



基金项目: 国家自然科学基金青年科学基金项目(61902221)

**作者简介:** 荆昊(1997-),男,山东建筑大学计算机科学与技术学院硕士研究生,研究方向为蛋白质从头测序;贾晓田(2000-),女,山东建筑大学计算机科学与技术学院硕士研究生,研究方向为多肽特征提取;柳楠(1980-),女,博士,山东建筑大学计算机科学与技术学院教授、博士生导师,研究方向为算法复杂度。本文通讯作者:柳楠。

深度学习作为人工智能的一个重要分支,已经在蛋白质结构预测、蛋白质-蛋白质相互作用分析以及化合物-蛋白质相互作用预测等多个领域展现出强大的功能<sup>[5-6]</sup>。在蛋白质从头测序领域,深度学习技术的应用有望突破传统方法的限制,实现高通量、低成本、高效率的蛋白质序列分析<sup>[7]</sup>。深度学习在蛋白质从头测序中的应用不仅能够提高测序的准确性和效率,而且能够揭示蛋白质序列与其功能之间的复杂关系,为疾病机理研究、药物开发和生物标志物的发现提供新的视角和工具<sup>[8-9]</sup>。此外,随着单分子蛋白质测序技术的发展,深度学习算法在处理大规模蛋白质组数据方面将发挥更加重要的作用,从而推动个体化医疗和精准医疗的发展<sup>[10]</sup>。

## 1 蛋白质从头测序问题描述

蛋白质从头测序作为生物质谱学中的一项关键任务,旨在通过分析质谱数据来推导未知蛋白质的氨基酸序列,对于生物学研究和疾病诊断具有极其重要的意义<sup>[11-14]</sup>。尽管面临诸如搜索空间过于庞大、质谱数据的复杂性以及不确定性等挑战,串联质谱技术(Tandem Mass Spectrometry, MS/MS)的应用显著提升了蛋白质组学分析的准确性<sup>[15-16]</sup>。MS/MS技术通过揭示母离子与子离子之间的关联信息,实现了对复杂生物样品中蛋白质的特异性鉴定。MS/MS与色谱技术的联用进一步提升了分析能力,使得对热稳定性差、强极性和高分子量的蛋白质进行分析成为可能,为解决蛋白质从头测序难题提供了强有力的工具。尽管质谱数据可能受到噪声、碎片缺失和修饰的影响,MS/MS技术与先进算法的结合正不断提高还原蛋白质原始序

列的能力,推动着蛋白质组学研究的深入<sup>[17]</sup>。

近年来,研究者们提出DeepNovo系列、Denovo-GCN、GraphNovo、PEAKS、Novor、PepNovo等多种先进蛋白质从头测序方法,这些方法被广泛应用于质谱数据的深度解析,实现了对肽段序列的高精度鉴定、蛋白质翻译后修饰的精准定位以及蛋白质相互作用网络的高效构建<sup>[18-20]</sup>。解决蛋白质从头测序问题不仅需要创新的实验技术,而且需要结合先进的计算方法。因此,该研究领域需不断提高算法的效率、增强模型对噪音和变异的鲁棒性,并进一步探索深度学习技术与图谱方法融合的可能性,以更全面、准确地还原蛋白质的氨基酸序列。

## 2 基于深度学习的蛋白质从头测序方法——DeepNovo系列

### 2.1 DeepNovo

DeepNovo是用于蛋白质从头测序的深度学习模型,以其卓越的性能和创新的网络设计而闻名,为蛋白质组学的研究提供了一种强大工具<sup>[21]</sup>。该模型的突出特点在于其采用卷积神经网络(Convolutional Neural Network, CNN)架构,可以从质谱数据中提取更深层次、更全局的特征。与传统的手工设计特征相比,DeepNovo的CNN架构能够更有效地捕捉蛋白质肽段的模式与规律,从而提高蛋白质从头测序的准确性。如图1所示,DeepNovo采用序列标记与解码的方法,逐步推断蛋白质的氨基酸序列。这种端到端的训练框架使其能够直接从原始数据中学习特征,无需对从质谱仪得到的蛋白质质谱数据进行特征的人工设计与提取,提高了训练效率。

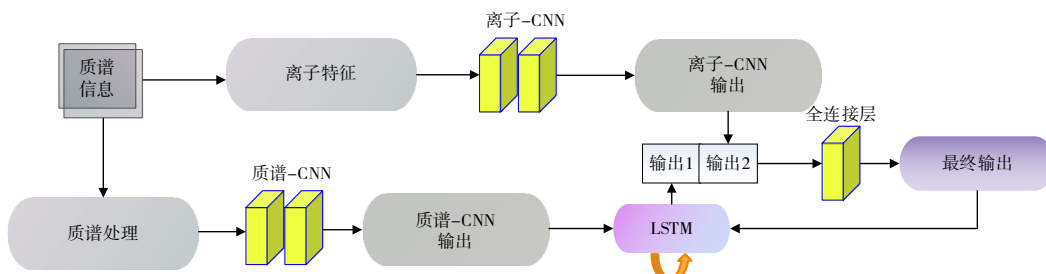


Fig. 1 DeepNovo sequencing flow

图1 DeepNovo测序流程

为确保模型的性能和泛化能力,DeepNovo采用大规模的蛋白质质谱数据集进行训练,从而更好地适应多样的生物学样本和实验条件,提高了其在真实场景中的可靠性。DeepNovo模型提供了一个完整的端到端训练与预测方案,能够适应任何数据源,为研究人员提供了极大的灵活性。此外,DeepNovo无需依赖外部数据库便可独立完成蛋白质序列的重建,这一点在处理新的或未知蛋白质序列时尤为重要。同时,DeepNovo的设计允许针对特定类型的数据进行再训练,增强了其适应性和泛化能力。

### 2.2 DeepNovo-DIA

DeepNovo-DIA模型的设计考虑了数据非依赖型采集

(Data Independent Acquisition, DIA)质谱的特点,通过深度学习技术提高了蛋白质鉴定性能<sup>[22]</sup>。该模型的创新点在于其可对蛋白质质谱数据的图结构进行建模,质谱图中的每个峰被视为图中的一个节点,而节点之间的边代表峰之间的相关性。这种图结构的表示方式使得模型能够捕捉到质谱数据中的复杂相互作用与模式。技术路线如图2所示。

为训练图结构,DeepNovo-DIA采用图卷积神经网络(Graph Convolution Neural Networks, GCN)的变体。GCN通过聚合邻居节点的信息来更新每个节点的表示,从而学习节点之间的复杂关系。如图3所示,DeepNovo-DIA模型结

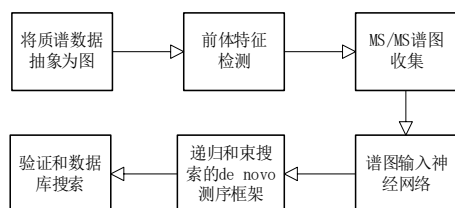


Fig. 2 DeepNovo-DIA technology roadmap

图2 DeepNovo-DIA 技术路线

合CNN来提取质谱图中的局部特征,以及长短期记忆网络(Long Short-Term Memory, LSTM)来处理序列数据,形成一个端到端的学习框架。其中,CNN有助于识别质谱图中的模式,而LSTM则维持序列信息,这对于正确重建氨基酸序列至关重要。

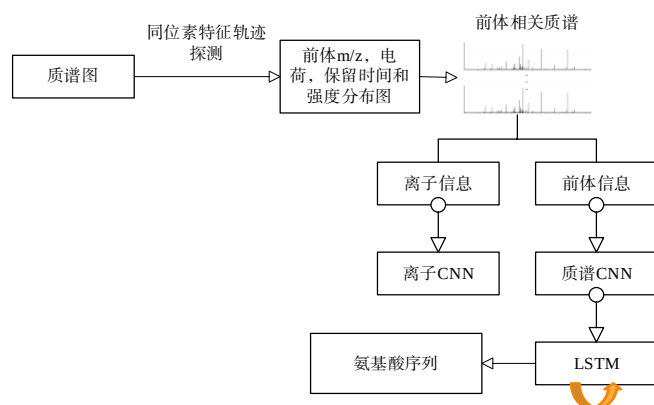


Fig. 3 DeepNovo DIA workflow

图3 DeepNovo-DIA 工作流程

通过这种结合GCN、CNN与LSTM的深度学习架构,DeepNovo-DIA模型能够同时处理图结构数据和序列数据,有效应对质谱数据的复杂模式,提高从头测序的准确性。此外,DeepNovo-DIA在准确性和可扩展性方面也表现出色,不仅能够精准预测已知肽段序列,而且还能发现数据库中未记录的新肽段,因此适用于新兴的临床样本分析。

DeepNovo-DIA模型针对DIA数据的特点进行了优化,能够处理高度多重化的质谱数据,这是DeepNovo在设计时未特别考虑的,这种创新性设计使其成为研究蛋白质组学的前沿方法。

### 2.3 DeepNovoV2

个性化癌症疫苗被广泛认为是下一代具有合理性和前景的癌症免疫疗法。采用DeepNovoV2对蛋白质质谱数据进行从头测序是识别肿瘤细胞表面特异性抗原的有效方法,也是开发个性化癌症疫苗的关键步骤<sup>[23]</sup>。来自不同物种的实验数据表明,DeepNovoV2模型相较于以往方法在肽水平上实现了13.01%~23.95%的更高测序准确度。

DeepNovoV2模型采用一种创新的谱图表示方法,直接将质谱图表示为(m/z,强度)对,避免了DeepNovo中使用的强度向量表示方法,解决了精度、速度与内存之间的权衡问题。更值得一提的是,DeepNovoV2模型采用一种顺

序不变网络结构T-Net(Transformation Network),这种结构对输入数据的不同排列具有不变性,特别适合处理质谱图中的峰值序列,提高了模型对不同峰值之间关联的建模能力。T-Net是PointNet的组成部分,其主要作用为实现输入数据的变换,使质谱图的峰值序列对于模型的输入具有不变性。如图4所示,T-Net结构包括3个1D卷积层,紧接着是全局最大池化层和3个全连接层。这种设计确保了模型对于输入数据的排列顺序不敏感,使其在处理质谱图等无序数据时更为有效。

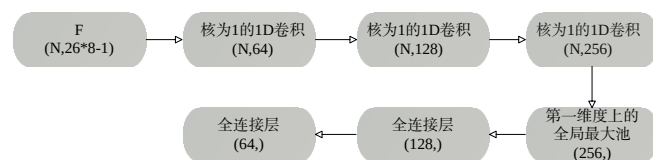


Fig. 4 T-Net structure

图4 T-Net 结构

### 2.4 DeepNovo A+

DeepNovo A+模型通过引入Transformer模型来增强对蛋白质质谱数据的处理能力<sup>[24]</sup>。Transformer模型最初在自然语言处理领域中取得巨大成功,其核心优势在于能够有效捕捉序列数据中的长距离依赖关系<sup>[25]</sup>。在DeepNovo A+中,Transformer的自注意力机制被用于分析质谱数据中的各个氨基酸残基,其将每个残基的质谱信号视为序列中的一个元素,从而识别并整合序列中的远程信息。此外,DeepNovo A+中的Ion-CNN网络是专门为提取质谱信号特征而设计的。该网络通过卷积层识别质谱图中不同区域的信号模式,并将这些局部特征整合起来,形成对整个肽段的全面表征。为了更全面地利用二级质谱中的信息,DeepNovo A+扩展了Ion-CNN的输入规模,纳入了包括a、b、y等不同类型的离子特征信息,以及它们之间的关联性。DeepNovo A+模型的另一个关键特点为引入了验证集自动控制迭代次数的策略。该策略通过监控验证集损失函数值的变化来实时调整模型的训练过程,以避免欠拟合或过拟合的问题。当验证集损失函数值在连续5次迭代中未能改善时训练将自动停止,从而确保模型的泛化能力。

通过结合Transformer模型的自注意力机制、改进的Ion-CNN网络以及验证集自动控制迭代次数策略,DeepNovo A+模型对质谱数据的理解进一步加深,对蛋白质从头测序的准确性和鲁棒性显著提升,成为生物信息学领域中一种先进、全面的分析方法。

## 3 适合处理长序列的蛋白质从头测序方法

### 3.1 Denovo-GCN

Denovo-GCN是一种结合GCN与CNN的模型,其可充分利用二级质谱数据,通过构建基于质谱谱峰质量差异的无向图将肽段的匹配信息与质谱数据结合起来<sup>[26]</sup>。Denovo-GCN模型首先使用CNN提取质谱图中每个峰(节点)的特征,这些特征捕捉了质谱图中的局部信息,如峰强度和

质量—电荷比 ( $m/z$ ), 被用于构建图的节点特征矩阵, 为后续的图卷积操作提供输入。其次, GCN 利用节点特征矩阵, 通过聚合邻居节点的信息来更新每个节点的表示, 使模型捕捉到节点之间的复杂关系, 从而更全面地理解质谱数据中的全局模式。GCN 的输出为更新后的节点特征, 这些特征包含了原始 CNN 提取的局部信息以及通过图卷积操作获得的全局信息。最后, 通过若干个线性层和 Softmax 层, Denovo-GCN 输出预测的氨基酸序列。具体结构如图 5 所示。这种端到端的训练和预测框架使得 Denovo-GCN 模型能有效处理长序列蛋白, 并在氨基酸精度、召回率和肽段召回率方面优于传统的 DeepNovo 模型<sup>[27]</sup>。

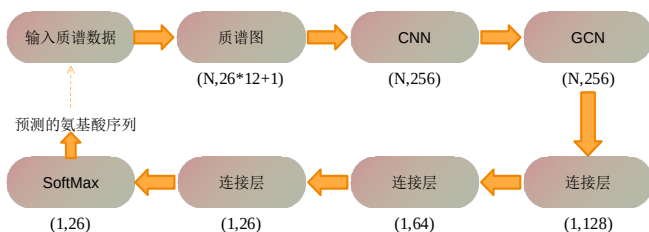


Fig. 5 Denovo-GCN model structure

图5 Denovo-GCN 模型结构

### 3.2 SeqNovo

SeqNovo 模型巧妙整合了 Seq2Seq 结构、多层感知机 (Multilayer Perceptron, MLP) 和注意力机制, 提高了对氨基酸序列的预测能力<sup>[28]</sup>。其中, Seq2Seq 结构的优势在于能够处理长序列, 尤其适用于物联网医学领域中的连续数据。SeqNovo 模型采用一层 GRU (Gated Recurrent Unit) 作为循环单元, 克服了传统 Seq2Seq 在处理长序列时的长程依赖性不足问题, 使得模型能够更好地捕捉氨基酸序列中远距离的相互作用。SeqNovo 模型在编码器中加入由 3 个全连接层组成的 MLP, 这些 MLP 层在特征维度上与传统的 Seq2Seq 编码器输出拼接, 增强了模型对特征的提取能力。注意力机制解决了 Seq2Seq 在将长序列转换为固定长度向量时可能导致的信息损失问题。通过在解码器中引入注意力机制, 模型能够在预测当前元素时集中注意力于先前预测的元素, 从而更好地捕捉序列中的长距离依赖性。在氨基酸序列预测任务中需要特别关注远距离氨基酸之间的相互作用, 而注意力机制有效提高了模型对这种长程依赖性的感知能力。

通过 3 种神经网络组件的有机整合, SeqNovo 模型相较 DeepNovo 模型在准确性和效率方面取得显著提升。此外, SeqNovo 模型在物联网医学领域还有很多潜在应用, 例如在实验室设备和疾病检测设备中的集成, 以及通过个性化肽段测序预测提供高度个性化的风险评估和治疗建议等。

Denovo-GCN 和 SeqNovo 在设计上均考虑了处理长序列蛋白质的挑战, 两者相较传统方法 (如 DeepNovo) 均取得了显著的性能提升。Denovo-GCN 构建的图结构可以更好地捕捉全局节点关系, 从而有效处理长序列蛋白质数据, 在肽段水平的召回率上相较 DeepNovo 有一定提升。Se-

qNovo 通过整合 Seq2Seq 结构、MLP 和注意力机制, 使得模型在序列建模的灵活性和动态特征捕捉方面更具优势, 相较于传统 Seq2Seq 和 DeepNovo 在准确性和性能方面均取得显著提升, 尤其是在处理长序列、提取特征和关注长距离依赖性方面取得了显著改善。

## 4 对质谱数据进行图建模的蛋白质从头测序方法

### 4.1 GraphNovo

GraphNovo 是一种创新的两阶段蛋白质序列预测模型, 专门用于解决质谱数据中的缺失碎片问题<sup>[29]</sup>。该模型的第一阶段为噪声分析与特征提取, 在该阶段, GraphNovo 模型分析了噪声对质谱解析的影响, 并引入噪声信号比 (Noise-to-Signal Ratio, NSR) 来衡量谱图的噪声水平。通过比较实验谱图与理论谱图, 研究者发现在缺失碎片较多的情况下, 噪声水平的增加反而可能导致算法性能提高。这一发现挑战了传统认知, 即通常认为噪声的增加会降低信号的清晰度, 从而降低分析的准确性。该阶段的分析为后续模型提供了关键的输入特征。第二阶段为深度学习模型训练与预测, 在该阶段, GraphNovo 模型使用图神经网络 (Graph Neural Network, GNN) 捕捉质谱数据中的复杂关系。GNN 能处理图结构数据, 其中节点代表质谱图中的碎片峰, 边表示这些峰之间的相互作用。通过训练, GraphNovo 模型可根据给定的质谱图和噪声水平准确预测肽段的氨基酸序列。

GraphNovo 模型的创新之处在于其能够容忍质谱数据中的缺失碎片, 因为 GNN 可以通过消息传递机制在节点之间传播信息, 从而推断缺失的碎片。这种两阶段处理方法使得 GraphNovo 在不同噪声水平和序列长度条件下均表现出色, 尤其是在处理含有较多缺失碎片的质谱数据时具有一定优势。

### 4.2 PepNet

PepNet 模型是专门针对高能碰撞解离质谱 (HCD-MS/MS) 数据进行蛋白质肽段从头测序的模型, 在人类和多种非人类生物的蛋白质组学研究方面表现出色<sup>[30]</sup>。PepNet 的主要结构包括: ①输入表示。MS/MS 光谱被表示为一维向量, 对光谱进行分箱, 确定质荷比 ( $m/z$ ) 范围为 0~2 000。在默认情况下, 使用 0.1 Th 的箱宽生成一个维度为 20 000 的向量。去除每个光谱中的前体峰, 并通过将每个峰的强度除以该光谱中最大峰的强度来进行归一化; ②时间卷积网络 (Temporal Convolutional Network, TCN) 模块。PepNet 包含 5 个连续的 TCN 块和下采样层, 这些块在处理输入光谱时具有递增的扩张率, 能够提取输入光谱的大部分信息, 同时在不同分辨率上工作, 以获得互补信息。同时, TCN 块在堆叠更多扩张层时可保持相对较低的计算成本; ③底部上分支。PepNet 引入底部上分支, 将所有 TCN 块的输出融合成一个单一的特征张量, 有助于整合全局与局部信息, 从而提高蛋白质从头测序的准确性; ④元信息融合。

针对每个输入, 将元信息(例如光谱的电荷、前体的  $m/z$ 、碰撞能量等)通过线性层进行变换, 并与先前提到的特征张量连接起来, 然后经过 Softmax 解码层生成最终的概率矩阵, 表示每个位置的氨基酸概率; ⑤损失函数和辅助任务。训练时使用平均交叉熵(cross-entropy)作为主要损失函数, 监测非填充位置上的平均准确度。为确保模型能输出准确的氨基酸序列, 需考虑前体质量, 因此损失函数还包括预测序列与真实序列质量的平均绝对误差。此外, 引入辅助任务, 如预测肽段是否为胰蛋白酶肽段、肽段长度等, 以提高模型性能。

GraphNovo 和 PepNet 均用于解决在质谱数据中存在缺失碎片情况下的蛋白质测序问题。GraphNovo 强调了在 DIA、中分子和大分子质谱数据分析中的应用, 而 PepNet 则专注于通过 HCD-MS/MS 数据进行蛋白质肽段的从头测序。GraphNovo 采用两阶段建模策略, 其核心优势在于通过消息传递机制推断缺失信息, 在 thaliana、elegans 和 coil 数据集上的性能较 DeepNovo 提升显著。PepNet 通过 TCN 模块和多分辨率特征融合提升预测精度, 在 PXD019483 数据集上的性能优于 DeepNovo。

## 5 基于 Bin reclassification 的深度学习方法

Bin reclassification 是一项任务, 通常用于质谱数据分析, 特别是质谱肽段测序。在该任务中, 质谱实验的结果被分为不同的质荷比( $m/z$ )区间, 这些区间被称为“bins”, 每个 bin 包含一组  $m/z$  范围内的离子。

Spectralis 是一种用于高精度肽段测序的方法, 通过引入神经网络模型进行 bin reclassification, 从而显著提高了测序灵敏度<sup>[31]</sup>。该方法将离散化的  $m/z$  值分配给离子系列, 即使在没有峰存在的情况下也能完成。通过预测的 bin 类别, Spectralis 能够改进肽段一谱匹配的评分, 进而提高已有方法预测的肽段召回率。Spectralis 模型的输入为离散化的(binned)谱表示, 包括对每个 bin 的实验峰强度、Prosit 预测的输入候选肽段内的峰强度以及输入候选肽段的 b-ions 和 y-ions 的二进制类标签。模型采用 AA-gapped 卷积层, 卷积的间隙参数直接对应 20 种标准氨基酸及蛋白质 N-甲硫氨酸的分子质量。模型使用批归一化和 ReLU 激活函数, 并通过 Adam 优化器进行训练, 使用 focal loss 作为损失函数。最佳模型超参数为 16 个 AA-gapped 卷积层, 每层 20 个过滤器、0.3 的 dropout 率以及  $4 \times 10^{-5}$  的学习速率。

Spectralis 引入基于 Levenshtein 距离估计的肽段一谱匹配评分机制, 显著提升了从头测序的性能。Spectralis 的评分并不改变原始预测的肽段序列, 而是作为独立模块运行, 具有低计算成本的优势, 可对跨工具预测结果进行统一排序与集成。Spectralis 还整合了进化算法, 实现了总体召回率的增加, 支持变异调用(如单氨基酸变异检测)等应用场景, 并适用于病原体监测、免疫肽质谱学及宏蛋白质组学等多个领域。

## 6 基于 Transformer 模型的双向预测技术

在蛋白质组学领域, 研究者通过 MS/MS 数据来确定未知肽段的氨基酸序列。传统方法通常从肽段的第一个氨基酸开始进行序列预测, 但未能充分利用 MS/MS 数据中的所有信息。近年来, 深度学习技术, 尤其是基于 Transformer 的模型, 为这一问题提供了新的解决方案。例如, NovoB 模型通过双向预测策略显著提高了肽段序列的识别准确性<sup>[32]</sup>。具体技术路线如图 6 所示。

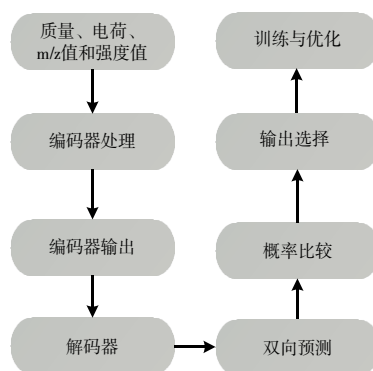


Fig. 6 Technology roadmap of NovoB model

图 6 NovoB 模型技术路线

NovoB 是一种基于 Transformer 的深度学习模型, 采用创新的双向预测方法进行肽段测序。该模型包含 1 个编码器和 2 个解码器, 其中编码器负责处理前体肽段的质量、电荷以及 MS/MS 谱中的  $m/z$  和强度数据; 两个解码器则分别从肽段的首尾两端开始预测氨基酸序列。通过这种方式, NovoB 能够同时考虑 MS/MS 谱中的 b 离子和 y 离子片段, 从而实现更准确的肽段序列预测。

NovoB 模型的主要优势在于其双向预测能力, 其能够更全面地利用 MS/MS 数据中的信息, 提高预测的准确性。在比较实验中, NovoB 模型在所有测试物种肽段水平上的平均准确率比 Casanovo 提高了约 9.8%, 并且在肽段水平和氨基酸水平的准确率上均优于 DeepNovo 和 PointNovo<sup>[33]</sup>。

## 7 结语

近年来, 基于深度学习的蛋白质从头测序方法研究取得了显著进展, 为蛋白质组学的发展提供了强有力的技术支持。DeepNovo 系列、Denovo-GCN、GraphNovo 等模型凭借其强大的数据处理能力和特征提取优势, 显著提升了蛋白质序列鉴定的准确性与效率。它们不仅能够处理复杂的质谱数据, 而且能够揭示蛋白质序列与功能之间的复杂关系。

尽管这些成就令人鼓舞, 蛋白质从头测序方法在实际应用中仍面临一些挑战。质谱数据的复杂性, 包括噪声、碎片缺失和修饰的影响对方法的准确性带来了一些挑战。此外, 现有方法在处理大规模数据集时的计算效率和可扩展

展性仍有待提高,模型的泛化能力和对不同实验条件的适应性也是当前研究中需要解决的问题。面对这些挑战,蛋白质从头测序方法的未来研究应聚焦于以下几个关键点:①进一步优化算法设计,提高对质谱数据的解析能力和鲁棒性;②开发更加高效的计算框架,以应对日益增长的蛋白质组数据规模;③加强对模型泛化能力的研究,使其能够适应更广泛的生物学样本和实验条件;④探索深度学习与其他计算方法的融合,如结合 GCN 和序列到序列学习,以进一步提升预测性能;⑤加强算法的解释性,使生物学家能够更好地理解和信任算法的预测结果。

通过以上技术优化与创新,相信基于深度学习的蛋白质从头测序方法将展现出更大的发展潜力和更广泛的应用前景,有望为生物医学研究带来更多突破性成果,并为人类健康福祉作出更大贡献。

#### 参考文献:

- [1] ISPANO E, LAVEZZO E, BIANCA F, et al. An overview of protein function prediction methods: a deep learning perspective[J]. *Current Bioinformatics*, 2023, 18(8): 621–630.
- [2] ZHONG H G, XIAO L, JIA X C, et al. Hierarchical graph transformer with contrastive learning for protein function prediction [J]. *Bioinformatics*, 2023, 39(7): btad410.
- [3] CHEN H K, HAO R, TIAN R J. Progress in protein sequencing technology[J]. *Science Bulletin*, 2021, 66 (25): 3309–3317.  
陈厚凯,郝瑞,田瑞军. 蛋白质测序技术进展[J]. *科学通报*, 2021, 66 (25): 3309–3317.
- [4] DUAN Z W, XIN F, KAI L, et al. Identification of SH2 domain-containing proteins and motifs prediction by a deep learning method[J]. *Computers in Biology and Medicine*, 2023, 162: 107065.
- [5] WANG Q, XU T, XU K, et al. Prediction of transport proteins from sequence information with the deep learning approach[J]. *Computers in Biology and Medicine*, 2023, 160: 106974.
- [6] XU W, CHENG M, XU X Y, et al. Deep learning method on deformation prediction for large-section tunnels[J]. *Symmetry*, 2022, 14(10): 2019.
- [7] TABERNERO Z M, ALONSO C Á, PRIETA L D F, et al. An overview of machine learning and deep learning techniques for predicting epileptic seizures. [J]. *Journal of Integrative Bioinformatics*, 2023, 20(4): 20230002
- [8] ZHAO X Y, QIN W J, QIAN X H. Application of deep learning methods in biological mass spectrometry and proteomics [J]. *Progress in Biochemistry and Biophysics*, 2018, 45 (12): 1214–1223.  
赵新元,秦伟捷,钱小红. 深度学习方法在生物质谱及蛋白质组学中的应用[J]. *生物化学与生物物理进展*, 2018, 45(12): 1214–1223.
- [9] XIA B B, WANG J. Protein modeling and design based on deep learning [J]. *Journal of Biotechnology*, 2021, 37 (11): 3863–3879.  
夏彬彬,王军. 基于深度学习的蛋白质建模与设计[J]. *生物工程学报*, 2021, 37(11): 3863–3879.
- [10] WEN B, ZHANG B. Computational proteomics: focus on deep learning [J]. *Proteomics*, 2020, 20(21–22): e2000258.
- [11] SHUKEN S R. An introduction to mass spectrometry-based proteomics [J]. *Journal of Proteome Research*, 2023, 22(7): 2151–2171.
- [12] JIMMY K E, MCCORMACK A L, YATES J R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database[J]. *Journal of the American Society for Mass Spectrometry*, 1994, 5(11): 976–989.
- [13] AEBERSOLD R, MANN M. Mass spectrometry-based proteomics[J]. *Nature*, 2003, 422: 198–207.
- [14] COX J, MANN M. MaxQuant enables high peptide identification rates, individualized p. p. b. -range mass accuracies and proteome-wide protein quantification[J]. *Nature Biotechnology*, 2008, 26(12): 1367–1372.
- [15] TABB D L, MCDONALD W H, YATES J R. DTASelect and contrast: tools for assembling and comparing protein identifications from shotgun proteomics[J]. *Journal of Proteome Research*, 2002, 1(1): 21–26.
- [16] ALEXEY I N, ANDREW K, EUGENE K, et al. A statistical model for identifying proteins by tandem mass spectrometry[J]. *Analytical Chemistry*, 2003, 75(17): 4646–4658.
- [17] SANGTAE K, PAVEL P A. MS-GF+ makes progress towards a universal database search tool for proteomics[J]. *Nature Communications*, 2014, 5 (1): 5277.
- [18] MA B, ZHANG K Z, CHRISTOPHER H, et al. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry [J]. *Rapid Communications in Mass Spectrometry*, 2003, 17 (20): 2337–2342.
- [19] MA B. Novor: real-time peptide de novo sequencing software[J]. *Journal of the American Society for Mass Spectrometry*, 2015, 26(11): 1885–1894.
- [20] LI C, LI K L, LI K Q, et al. SWPepNovo: an efficient de novo peptide sequencing tool for large-scale MS/MS spectra analysis [J]. *International Journal of Biological Sciences*, 2019, 15(9): 1787–1801.
- [21] TRAN H N, ZHANG X L, XI N L, et al. De novo peptide sequencing by deep learning[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2017, 114(31): 8247–8252.
- [22] TRAN H N, QIAO R, XIN L, et al. Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry [J]. *Nature Methods*, 2019, 16(1): 63–66.
- [23] QIAO R, TRAN H N, LI M, et al. DeepNovoV2: better de novo peptide sequencing with deep learning [DB/OL]. <https://arxiv.org/abs/1904.08514>.
- [24] DI L, HE Y X, LU Y G. Deep Novo A+: improving the deep learning model for peptide sequencing with additional ion types and validation set [J]. *Current Bioinformatics*, 2020, 15(8): 949–954.
- [25] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//*Proceedings of the 31st International Conference on Neural Information Processing System*, 2017: 6000–6010.
- [26] WU R T, ZHANG X, WANG R T, et al. Denovo-GCN: de novo peptide sequencing by graph convolutional neural networks [J]. *Applied Sciences*, 2023, 13(7): 4604.
- [27] SANGTAE K, NITIN G, PAVEL P A. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases [J]. *Journal of Proteome Research*, 2008, 7(8): 3354–3363.
- [28] WANG K, ZHU M J, WADII B, et al. SeqNovo: de novo peptide sequencing prediction in IoMT via Seq2Seq[J]. *IEEE Journal of Biomedical and Health Informatics*, 2023, 29(4): 2377–2387.
- [29] MAO Z P, ZHANG R X, XIN L, et al. Mitigating the missing-fragmentation problem in de novo peptide sequencing with a two-stage graph-based deep learning model [J]. *Nature Machine Intelligence*, 2023, 5 (11): 1250–1260.
- [30] LIU K Y, YE Y Z, LI S J, et al. Accurate de novo peptide sequencing using fully convolutional neural networks [J]. *Nature Communications*, 2023, 14(1): 7974.
- [31] DANIELA K, JOHANNER H, YANIK B, et al. Deep learning-driven fragment ion series classification enables highly precise and sensitive de novo peptide sequencing[J]. *Nature Communications*, 2024, 15(1): 151.
- [32] LEE S, KIM H. Bidirectional de novo peptide sequencing using a transformer model[J]. *PLoS Computational Biology*, 2024, 20(2): e1011892.
- [33] YILMAZ M, FONDIRE W, BITTREMIEUX W, et al. De novo mass spectrometry peptide sequencing with a transformer model [C]//*International Conference on Machine Learning*, 2022: 25514–25522.

(责任编辑:尹晨茹)