朱宇翀, 陈德华, 潘乔. 基于 Bi-LSTM 和 Transformer 的谱图预测模型[J]. 智能计算机与应用,2025,15(3):203-206. DOI: 10.20169/j.issn.2095-2163.250330

基于 Bi-LSTM 和 Transformer 的谱图预测模型

朱宇翀, 陈德华, 潘 乔
(东华大学 计算机科学与技术学院, 上海 201620)

摘 要: 数据非依赖采集(DIA)近年来发展迅速,在蛋白质组学中也有着广泛的应用。DIA 数据的蛋白质鉴定通常需要使用由数据依赖采集(DDA)得到的谱图数据库。然而该数据库含有的信息有限,为了在搜索过程中覆盖更多的蛋白质,目前采用深度学习模型的预测结果对该数据库进行补充。针对谱图预测任务,不同模型在不同数据集上的表现存在差异,且仅有少量模型展示了其在四维(4D)质谱数据上的性能。本文比较不同序列模型在 4D-DIA 血浆数据上的表现,提出了一个新的模型结构,该模型使用门控结合了双向长短期记忆网络(Bi-LSTM)和 Transformer 的特征,在较长的氨基酸序列上拥有更好的表现。

关键词: 数据非依赖采集技术; 谱图预测; 双向长短期记忆网络; Transformer
中图分类号: TP399 **文献标志码**: A **文章编号**: 2095-2163(2025)03-0203-04

Bi-LSTM and Transformer based model for spectrum prediction

ZHU Yuchong, CHEN Dehua, PAN Qiao
(School of Computer Science and Technology, Donghua University, Shanghai 201620, China)

Abstract: Data Independent Acquisition (DIA) has developed rapidly in recent years and is widely applied in proteomic studies. A sample-specific spectral library from Data Dependent Acquisition (DDA) is used to identify proteins in DIA data. Because of the limitation of the DDA spectral library, predictions results of deep learning models are used to enrich the library, thereby covering more proteins during the searching. For spectrum prediction, different models perform inconsistently on different datasets, and only a small number of models have demonstrated their performance on 4D mass spectrometric data. The performance of different sequence models on 4D-DIA plasma data is compared and a new model structure is proposed. The model employs gate combining features extracted by Bi- LSTM and Transformer, which works better on long amino acid sequences.

Key words: DIA; spectrum prediction; Bi-LSTM; Transformer

0 引 言

液相色谱-质谱联用仪可以鉴定生物基质中的蛋白质,在蛋白质组学有着极其重要的应用。由于 4D-DIA 技术有着重现性高以及定量准确度高等优点[1],近年来备受青睐。相比 3D 质谱数据,4D 质谱数据新增了离子淌度,能够反映离子空间结构上的差异。DIA 技术则是根据母离子质荷比范围无偏地设置隔离窗口,将窗口内的所有母离子共碎裂,由此得到混合了多个母离子共碎裂信息的二级谱图[2]。尽管 DIA 有着很多优点,但是 DIA 数据的复杂性无疑增加了蛋白质鉴定的难度。这个问题也同样存在于 4D-DIA 数据。为了得到准确的分析和定量结果,常用的做法是采集部分待测样本的 DDA 数据,使用由 DDA 数据得到谱图数据库鉴定 DIA 数据中的蛋白质。然而,DDA 方法自身存在的局限性会对 DIA 数据的蛋白质鉴定结果产生不利的影响。

得益于近年来深度学习技术的飞速发展,目前可以借助神经网络的预测结果,进一步丰富谱图数据库,从而减轻 DDA 方法带来的不利影响。预测谱图是其中一项难度较高的任务,许多现有的研究工作使用序列模型来解决这个问题[3-9]。虽然这些模型在不同数据集上取得了一定的效果,但并不存在一种最优的模型结构能在不同数据集上都优于其他模型。此外,绝大多数模型没有展示其在 4D 质谱数据上的表现。因此,何种模型能够更好地预测 4D 质谱数据的谱图仍是一个有待研究的问题。

本文比较了目前广泛使用的序列模型 Bi-

作者简介: 朱宇翀(1998—),男,硕士研究生,主要研究方向:深度学习。Email:1649719302@qq.com; 陈德华(1976—),男,博士,教授,主要研究方向:深度学习,数据科学与智慧医疗; 潘 乔(1977—),男,博士,副教授,主要研究方向:人工智能,大数据与智慧医疗。

收稿日期: 2023-09-07

LSTM 和 Transformer 在 4D-DIA 数据上的表现。同时,研究提出了一个新的模型结构,该结构使用门控结合了 Bi-LSTM 和 Transformer 的优势,使得该模型在预测较长的氨基酸序列所对应的谱图时,仍能取得较好的效果。

1 相关工作

1.1 序列模型

序列模型用于处理存在序列关系的数据,例如语音、文本、DNA 等。Elman[10] 最早提出了循环神经网络(Recurrent Neural Network,RNN)的基本结构,该网络以上一个时刻的隐藏层的信息作为当前时刻的输入,以此实现了信息的传递。在 Elman Network 的基础上,Hochreiter 等学者[11] 提出长短期记忆神经网络(Long Short-Term Memory, LSTM)。LSTM 通过引入记忆细胞和门控单元解决了 RNN 因梯度消失而无法学到长期依赖的问题。为了提升运算效率,Chung 等学者[12] 提出了门控循环单元(Gated Recurrent Unit, GRU)。GRU 不再沿用 LSTM 中的记忆细胞结构,而是通过改进门控单元以达到储存长期信息的目的。在 2017 年,Vaswani 等学者[13] 提出了 Transformer,其中结合了多头注意力机制和残差连接,在众多领域都有出色表现,目前许多 state-of-the-art 的模型结构都使用了 Transformer。

1.2 谱图预测

谱图预测的一种主要方法是根据标注的离子序列种类预测该离子的离子强度。Sven 等学者基于梯度提升回归树(Gradient Boosting Regression Tree, GBRT)提出 MS2PIP[14],该方法在谱图预测任务上取得了一定的成果。随着深度学习技术的发展,近年来的绝大部分工作主要使用深度学习来预测谱图。Zhou 等学者基于 Bi-LSTM 提出了 pDeep[3],该模型的性能超过了传统的机器学习方法。以此为基础构建的 pDeep2[4] 将质谱仪的型号作为输入提升了模型的泛用性,而后提出的 pDeep3[5] 则添加了小样本学习策略以解决训练样本不足的问题。Tiwary 等学者[7] 提出的 DeepMass: Prism[6] 也同样基于 LSTM,但是放宽了对于输入序列长度的限制。Gessulat 等学者[6] 基于 GRU 提出了 Prosit,该模型将碰撞能量作为模型的输入特征。自 Transformer 问世之后,Lou 等学者[8] 基于 LSTM 和 Transformer 的组合提出了 DeepPhospho,该模型可以准确预测含有磷酸化修饰的肽段的谱图。Zeng 等学者[9] 提出的 AlphaPeptDeep 则展示了仅通过 Transformer 便能在多个数据集上取得很好的预测效果。

2 模型结构

模型采用了编码器-解码器的结构,其整体结构如图 1 所示。模型接受一个长度不超过 50 的氨基酸序列和母离子电荷作为输入。首先对氨基酸序列和母离子电荷进行嵌入,随后通过编码器提取特征,最终由解码器将特征映射为离子强度的预测结果。模型的输出是一个 98 维的向量,表示 49 个带单个电荷的 b 离子和 49 个带单个电荷的 y 离子的离子强度预测结果。在离子强度不可能存在的情况下,其值将由-1 表示。

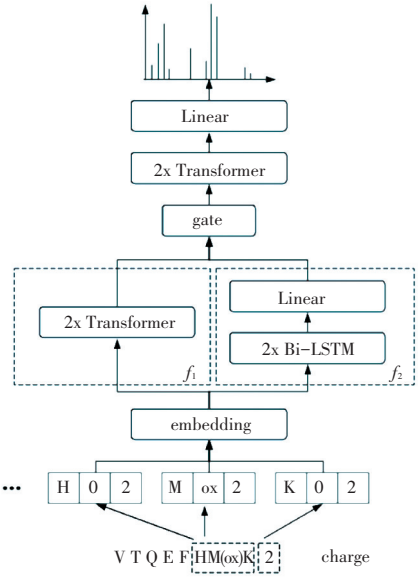


图 1 模型结构图

Fig. 1 Overview of the model structure

2.1 嵌入

氨基酸序列包含氨基酸和修饰。本文的模型首先对氨基酸、修饰以及母离子电荷进行嵌入。研究中采用和词嵌入类似的方式嵌入氨基酸和修饰,并通过单层全连接神经网络将母离子电荷映射成一个高维向量,由此得到母离子电荷的嵌入。最终,将三者进行拼接便得到了 $x = \{x_1, x_2, \cdots, x_n\}$,其中 n 表示氨基酸序列的长度,序列中的某个元素 x_i 可以由以下方式表示:

$$x_i = [\boldsymbol{W}_a(a_i), \boldsymbol{W}_m(m_i), \boldsymbol{W}_C(c)] \quad (1)$$

其中,$\boldsymbol{W}_a, \boldsymbol{W}_m, \boldsymbol{W}_C$ 表示一个可以学习的参数矩阵;a_i 表示一个氨基酸在氨基酸词典中的索引;m_i 是其对应的修饰在修饰词典中的索引;c 表示母离子电荷。

2.2 编码器

编码器提取特征的过程可以分为 3 步。

(1)是采用不同的模型结构 f_1 和 f_2 得到 x 的不同特征表示 e_l 和 e_t。该操作可以由以下方式表示：

$$e_l = f_1(x) \tag{2}$$

$$e_t = f_2(x) \tag{3}$$

其中，f_1 表示 2 层 Transformer 编码器，每个 Transformer 模块的 d_{model} 为 128，前馈神经网络的神经元个数为 512，且共有 8 个自注意力机制头。f_2 表示 2 层 Bi-LSTM，其隐藏状态的大小为 128。研究中，额外增加了一层全连接神经网络用于调整向量的维度。

(2)采用门控来合并 e_l 和 e_t，由此得到 e_{merge}，该过程的数学公式分别如下：

$$z = \sigma(\boldsymbol{W}_z([e_l, e_t])) \tag{4}$$

$$e_{\text{merge}} = z \times e_l + (1 - z) \times e_t \tag{5}$$

其中，$\boldsymbol{W}_z$ 表示一个可以学习的参数矩阵，σ 表示 Sigmoid 激活函数。

该操作结合了 Bi-LSTM 和 Transformer 各自的优势，使模型可以得到更好的特征表示。

(3)受 DeepPhospho[8] 的启发，研究使用了 Transformer 进一步提炼特征。在这里，还引入了残差链接，即先将 x 与 e_{merge} 进行相加，而后经过 2 层 Transformer 得到最终的特征 e。此处，Transformer 的模型参数设置与 f_1 相同。

2.3　解码器

解码器采用了 2 层全连接神经网络，每层包含 128 个神经元，并使用 ReLU 作为激活函数。

3　实验

3.1　数据集

本文采集了 10 份心理疾病患者以及 10 份无心理疾病人士的血浆样本，而后通过 Bruker 公司的 TimsPro2 质谱仪得到 10 条 4D-DDA 质谱原始数据。接着，研究使用 FragPipe-18.0 对 DDA 数据进行分析，并基于软件生成的 DDA 谱图数据库构建数据集。数据集中氨基酸序列的长度范围在 7~50 之间，其中 97%的氨基酸序列长度不超过 30。数据集总共包含 11 161 个样本，实验中随机划分数据集中 85%的数据作为训练集，余下的 15%作为测试集。

3.2　训练过程

训练过程在 Ubuntu20.04 LTS 上进行。研究中使用了 Adam 作为优化器，*L*1 *loss* 作为损失函数。用于训练模型的超参数：*epochs* 为 100，初始学习率为 5e-4，学习率预热为 40 个 *epoch*，*batch size* 为 64。

为了比较不同模型之间的表现，本文还训练了以下 3 个使用不同编码器的模型：

(1) Model 1。4 层 Bi-LSTM，其隐藏状态的大小为 128。

(2) Model 2。4 层 Transformer 编码器，每个 Transformer 模块的 d_{model} 为 128，前馈神经网络的神经元个数为 512，共拥有 8 个自注意力机制头。

(3) Model 3。简单拼接 2 层 Bi-LSTM 与 2 层 Transformer 的模型，Bi-LSTM 和 Transformer 的模型参数设置与 Model 1 和 Model 2 相同。

训练这 3 个模型的超参数设置和训练本文所提出的模型相同，除了在训练 Model 2 时，*epochs* 被更改为 250。

3.3　实验结果

研究使用皮尔逊相关系数(Pearson correlation coefficient，*PCC*)和光谱角(Spectral Angle，*SA*)来评价模型的预测结果。*PCC* 可以衡量 2 个向量之间的线性相关性，其取值范围在-1 到 1 之间，越接近 1 表示 2 个向量之间的相关性越高。*PCC* 计算方法如下式所示：

$$PCC = \frac{cov(p,t)}{\sigma_p \sigma_t} \tag{6}$$

其中，p 和 t 分别表示模型预测的离子强度和真实值；$cov(p,t)$ 表示 p 和 t 之间的协方差；σ_p 和 σ_t 分别表示 p 和 t 的标准差。SA 通过计算预测结果和真实值之间的角度来度量两者之间的相关性，其取值范围在-1~1 之间，越接近 1 表示预测结果越好。*SA* 的计算公式具体如下：

$$SA = 1 - 2\frac{\cos^{-1}(p \cdot t)}{\pi} \tag{7}$$

其中，p 和 t 分别表示模型预测得到的离子强度和真实值。

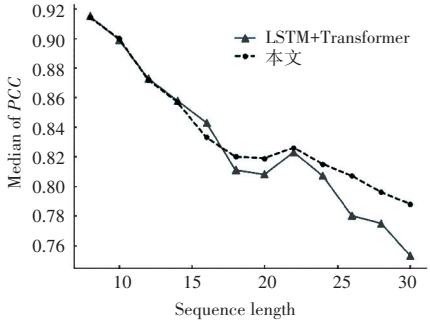
不同模型预测结果见表 1。从表 1 中可以观察到仅使用 Bi-LSTM 的模型要优于仅使用 Transformer 的模型，说明 Bi-LSTM 更易于学习到谱图的特征。此外，本文提出的模型和 Model 3 总体性能接近，都超过了仅使用 Bi-LSTM 和仅使用 Transformer 的模型，说明将 Bi-LSTM 与 Transformer 结合能够发挥这 2 种模型各自的优势，从而取得更加优异的表现。

表 1　不同模型的预测结果比较

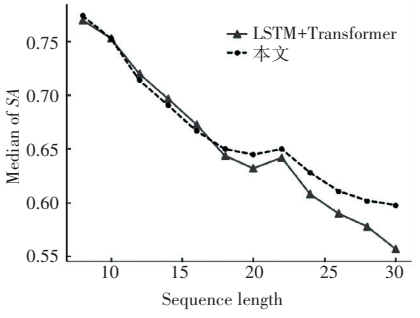
Table 1　Comparison of results of different models

模型	*PCC*	*SA*
本文	0.914	**0.776**
Model 1	0.902	0.759
Model 2	0.877	0.724
Model 3	**0.915**	0.772

尽管本文提出的模型和 Model 3 在整体表现上差异不大,但是本文提出的模型在预测较长的氨基酸序列时有着更好的表现。对于指定大小的 n = 8, 10, …, 30,从测试集中过滤了氨基酸序列长度小于 n 的测试样本,得到了 12 个新的测试集,并测试了本文的模型与 Model 3 在这些测试集上的性能,其结果如图 2 所示。图 2(a)和图 2(b)分别展示了 PCC 和 SA 与测试集中氨基酸序列长度最小值之间的关系。总体而言,模型的表现会随着氨基酸序列长度的增加而降低。当测试集中氨基酸序列的长度最小值超过 20 时,本文的模型的表现优于 Model 3。由于 2 个模型在预测较短的氨基酸序列所对应的谱图时表现很好,模型预测结果对后续蛋白质的搜索工作影响有限。相反,对于较长的氨基酸序列而言,模型表现越好,越有助于质谱分析软件鉴定出正确的蛋白质。



(a) *PCC* 与测试集中序列长度最小值的关系



(b) *SA* 与测试集中序列长度最小值的关系

图 2 模型性能与测试集中序列长度最小值之间的关系

Fig. 2 The relationship between performance and the minimum sequence length in the test set

4 结束语

本文比较了不同序列模型在 4D 质谱数据上预测谱图的表现,并提出了一个新的模型。该模型的表现超越了仅使用 Bi-LSTM 构建的模型以及仅采用 Transformer 构建的模型,并且对于较长的氨基酸序列输入也能得到较好的预测结果。

由于 4D-DIA 数据获取的成本较高,本文所使用的数据量有限,在未来研究中将会基于更多的数据提出更好的模型结构。同时,后期工作将尝试使用大模型来统一训练谱图预测和蛋白质结构预测等多个与氨基酸序列有关的任务。

参考文献

[1] MEIER F, BRUNNER A D, FRANK M, et al. diaPASEF: parallel accumulation-serial fragmentation combined with data-independent acquisition[J]. Nature Methods, 2020, 17(12): 1229-1236.

[2] 侯鑫行,周丕宇,宫鹏云,等. 基于数据非依赖采集的蛋白质组质谱数据解析方法研究进展[J]. 生物化学与生物物理进展, 2022,49(12): 2364-2386.

[3] ZENG Wenfeng, ZHOU Xiexuan, ZHOU Wenjing, et al. MS/MS spectrum prediction for modified peptides using pDeep2 trained by Transfer learning[J]. Analytical Chemistry, 2019, 91(15): 9724-9731.

[4] TARN C, ZENG Wenfeng. pDeep3: Toward more accurate spectrum prediction with fast few-shot learning[J]. Analytical Chemistry, 2021, 93(14):5815-5822.

[5] ZHOU Xiexuan, ZENG Wenfeng, CHI Hao, et al. pDeep: Predicting MS/MS spectra of peptides with deep learning[J]. Analytical Chemistry, 2017, 89(23): 12690-12697.

[6] GESSULAT S, SCHMIDT T, ZOLG D P, et al. Prosit: Proteome-wide prediction of peptide tandem mass spectra by deep learning[J]. Nature Methods, 2019, 16(6):509-518.

[7] TIWARY S, LEVY R, GUTENBRUNNER P, et al. High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis[J]. Nature Methods, 2019, 16(6):529-525.

[8] LOU Ronghui, LIU Weizhen, LI Rongjie, et al. DeepPhospho accelerates DIA phosphoproteome profiling through in silico library generation[J]. Nature Communications, 2021, 12(1): 6685.

[9] ZENG Wenfeng, ZHOU Xiexuan, WILLEM S, et al. AlphaPeptDeep: A modular deep learning framework to predict peptide properties for proteomics[J]. Nature Communications, 2022, 13(1): 7238.

[10] ELMAN J L. Finding structure in time[J]. Cognitive Science, 1990, 14(2):179-211.

[11] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[12] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv preprint arXiv, 1412.3555, 2014.

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. arXiv preprint arXiv, 1706.03762, 2017.

[14] GABRIELS R, MARTENS L, DEGROEVE S. Updated MS^2PIP web server delivers fast and accurate MS^2 peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques[J]. Nucleic Acids Research, 2019, 47(W1): W295-W299.