

基于多头注意力机制和残差神经网络的肽谱匹配打分算法

闵鑫, 王海鹏*, 牟长宁

(山东理工大学 计算机科学与技术学院, 山东 淄博 255000)

(*通信作者电子邮箱 hpwang@sdut.edu.cn)

摘要: 肽谱匹配打分算法在肽序列鉴定的过程中起着关键性作用, 而传统的打分算法无法充分有效地利用肽碎裂规律进行打分。针对这一问题提出了一种结合肽序列信息表征的多分类概率和式打分算法 deepScore- α , 该算法不需要考虑全局信息进行二次打分, 不存在理论质谱与实验质谱相似度计算方法的限制。deepScore- α 使用一维残差网络对序列底层信息进行抽取, 再通过多头注意力机制融合序列不同肽键位点对当前肽键位点断裂产生的影响从而生成最终的碎片离子相对强度分布概率矩阵, 结合肽序列碎片离子的实际相对强度计算出最终的肽谱匹配得分。该算法与常用开源鉴定工具 Comet 以及 MSGF+ 进行了比较: 在人类蛋白组数据集上错误发现率 (FDR) 为 0.01 时, deepScore- α 保留的肽序列数量提升了约 14%, Top1 命中率 (正确肽序列在得分最高的谱图所占比例) 最大提升约 5 个百分点。使用人类蛋白组数据集训练的模型在 ProteomeTools2 数据集上进行泛化性能测试, 结果表明, 在 FDR 为 0.01 的条件下 deepScore- α 保留的肽序列数量提升了约 7%, Top1 命中率提升了约 5 个百分点, Top1 中来自 Decoy 库的鉴定结果减少约 60%。实验结果证明, deepScore- α 在较低 FDR 值情况下保留更多的肽序列并提升 Top1 的命中率, 且具有较好的泛化性能。

关键词: 打分算法; 肽序列鉴定; 注意力机制; 残差网络; 多分类概率和

中图分类号: TP391 **文献标志码:** A

Peptide spectrum match scoring algorithm based on multi-head attention mechanism and residual neural network

MIN Xin, WANG Haipeng*, MOU Changning

(School of Computer Science and Technology, Shandong University of Technology, Zibo Shandong 255000, China)

Abstract: Peptide spectrum match scoring algorithm plays a key role in the peptide sequence identification, and the traditional scoring algorithm cannot effectively make full use of the peptide fragmentation pattern to perform scoring. In order to solve the problem, a multi-classification probability sum scoring algorithm combined with the peptide sequence information representation called deepscore- α was proposed. In this algorithm, the second scoring was not performed with the consideration of global information, and there was no limitation on the similarity calculation method of theoretical mass spectrum and experimental mass spectrum. In the algorithm, a one-dimensional residual network was used to extract the underlying information of the sequence, and then the effects of different peptide bonds on the current peptide bond fracture were integrated through the multi-attention mechanism to generate the final fragmentation relative intensity distribution probability matrix, after that, the final peptide spectrum match score was calculated by combining the actual relative intensity of the peptide sequence fragmentation. This algorithm was compared with Comet and MSGF+, two common open source identification tools. The results show that when False Discovery Rate (FDR) was 0.01 on humanbody proteome dataset, the number of peptide sequences retained by deepScore- α is increased by about 14%, and the Top1 hit ratio (the proportion of the correct peptide sequences in the spectrum with the highest score) of this algorithm is increased by about 5 percentage points. The generalization performance test of the model trained by human ProteomeTools2 dataset show that the number of sequences peptide retained by deepScore- α at FDR of 0.01 is improved by about 7%, the Top1 hit ratio of this algorithm is increased by about 5 percentage points, and the identification results from Decoy library in the Top1 is decreased by about 60%. Experimental results prove that, the algorithm can retain more peptide sequences at lower FDR value, improve the Top1 hit ratio, and has good generalization performance.

Key words: scoring algorithm; peptide sequence identification; attention mechanism; residual network; multi-classification probability sum

收稿日期: 2019-11-04; **修回日期:** 2019-12-17; **录用日期:** 2019-12-18。 **基金项目:** 国家自然科学基金资助项目 (31500669); 山东省自然科学基金资助项目 (ZR2014FQ024); 山东省高等学校优秀青年创新团队支持计划项目 (2019KJN048)。

作者简介: 闵鑫 (1995—), 男, 四川成都人, 硕士研究生, 主要研究方向: 深度学习、生物信息学; 王海鹏 (1980—), 男, 山东淄博人, 副教授, 博士, 主要研究方向: 机器学习、生物信息学; 牟长宁 (1990—), 男, 山东淄博人, 硕士研究生, 主要研究方向: 深度学习、生物信息学。

0 引言

现如今主流的蛋白质序列鉴定方法中,序列数据库搜索法^[1]因其对不同的质谱类型表现出很强的鲁棒性且性能良好而成为最常用的鉴定方法。序列数据库搜索法的目的是在蛋白质序列数据库中搜索匹配出最可能产生给定实验质谱的序列,具体步骤一般分为三部分:对于每一张实验质谱,首先,从序列数据库中提取出与产生该实验质谱的母离子质量偏差在一定范围内的所有肽段形成肽谱匹配;然后,计算出每一个候选肽段的理论质谱和对应的实验质谱的得分;最后,再根据分数对肽谱匹配进行排序,取出分数最高者作为鉴定结果。其中,有效准确地计算出肽谱匹配得分对最终的鉴定结果起着决定性的作用。大多数打分算法都依赖于将候选肽转换成理论质谱,然后计算理论质谱与实验质谱的某种相似度分数,再通过全局信息进行二次打分。目前比较新的常用数据库搜索鉴定工具主要包括 Comet^[2-3]、MSGF+^[4]、Mascot^[5]、MaxQuant^[6]等,其中 Comet 和 MSGF+ 作为代表性的开源鉴定工具被广泛使用,但是这些传统的鉴定工具中所使用的肽谱匹配打分方法都依赖于已知的肽碎裂规律及相关信息,存在着理论质谱准确度和相似度计算方法有效性等限制。

使用非深度学习方法解决肽谱匹配打分问题一直是一个备受关注的问题, Bai 等^[7]使用最大二分匹配模型进行肽谱匹配打分并取得了不错的效果,但由于最大二分匹配模型存在的问题导致肽谱匹配打分仍旧存在很大的问题。为改进此类问题, Bai 等^[8]又提出了子模块广义匹配(Submodular Generalized Matching, SGM)模型,该模型在匹配准确率上有一定的提升,但是其充分利用质谱离子峰所携带的信息以及肽碎裂规律的能力还略显不足。深度学习领域不断的发展与完善,也为解决蛋白组学中的一些问题提供了新的途径。许多深度学习模型被用来解决蛋白组学的一些重要问题,如利用深度学习预测蛋白质结构^[9-10]。蛋白组学中的许多问题可以看作是序列信息处理的延伸和拓展,而序列信息处理在深度学习领域中也是一个备受关注的问题。循环神经网络常常被用作序列信息处理, Zhou 等^[11]提出了基于双向长短期记忆网络(Bidirectional Long Short-Term Memory network, Bi-LSTM)的碎片离子预测模型用于预测理论质谱,该模型具有良好的预测性能,但是并未进行进一步的肽谱匹配打分实验。随着深度学习的发展,众多学者发现 ResNet (Residual neural Network)^[12]不仅能有效解决计算机视觉领域许多重要问题,也能被用来有效地处理序列信息,并且已经有一些学者成功将 ResNet 应用到了蛋白组学的相关领域^[13-14]。在现今的自然语言处理领域,许多取得不错效果的深度学习模型都使用注意力机制^[15],注意力机制的核心目标是从众多信息中选择出对当前任务目标更关键的信息加以利用。在肽碎裂事件中,某一肽键的断裂不只与相邻两个氨基酸相关,也会与其他位置的氨基酸存在一定的关联,因此利用注意力机制能有效地利用这样一些相关性,进而提高模型的准确度。而多头自注意力机制(multi-head self-attention mechanism)^[16]是对普通注意力机制的改进。因此本文提出了基于深度学习的打分算法:deepScore- α 。该算法使用 ResNet 与多头注意力机制相结合的模型用于肽谱匹配打分,有效地结合了两者的优点,利用深

度学习自动学习肽碎裂相关规律,能获得比较好的打分效果。

1 模型与算法

1.1 deepScore- α 算法打分流程及模型结构

deepScore- α 打分算法使用深度学习模型学习肽碎裂规律对肽谱匹配进行打分,输入为肽谱匹配(Peptide Spectrum Match, PSM)中肽序列以及该肽序列在对应谱图中已标注的碎片离子离散化相对强度。算法分为两个阶段,输入的肽序列通过特征提取算法转换为特征序列与肽序列对应的碎片离子离散化相对强度一同输入模型,不同阶段输入相同,但是会产生不同的输出(具体流程如图1所示)。

1)在训练阶段,模型输出预测的碎裂离子离散化相对强度以评估模型学习肽碎裂规律的效果,并使用交叉熵作为模型损失进行权重更新。

2)在打分阶段,模型输出肽序列对应的碎片离子相对强度概率分布矩阵,再结合实际的碎片离子相对强度计算最终的肽谱匹配分数,而不是给出预测的碎片离子相对强度。

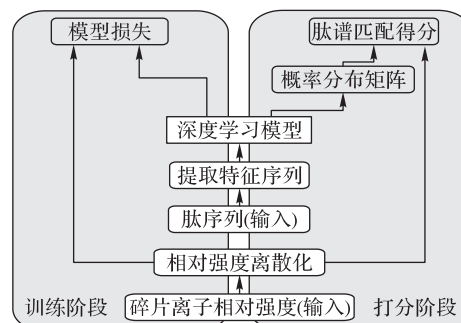
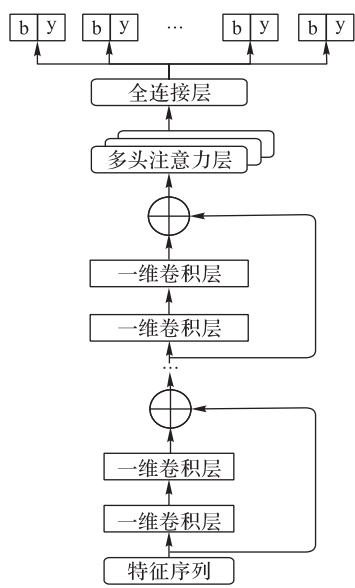


图1 deepScore- α 打分算法流程

Fig. 1 Flowchart of deepScore- α scoring algorithm

deepScore- α 使用的模型由一维 ResNet 模型和多头自注意力层组成: ResNet 由卷积层和残差链接构成,主要用于提取和处理序列特征,多头自注意力机制则利用注意力机制的特性更加全面准备地学习肽碎裂规律以提升模型预测效果。ResNet 是深度卷积网络的改进,深度卷积网络随着网络层数的不断增加,在网络能够收敛的前提下,由于网络优化问题其表现会出现下降。ResNet 通过残差连接能将较浅的网络层的输出直接输入到较深的网络层,在一定程度上解决了这个问题。deepScore- α 中所使用的 ResNet20 参考了 He 等^[12]提出的 ResNet,并且在此基础上进行了一些改进以适应肽序列信息处理,将其改为一维卷积且控制输入输出长度保持不变。

多头自注意力机制由谷歌于 2017 年提出,不同于之前的注意力机制,多头自注意力机制采用了更加新颖的 multi-head 机制,即多次计算的注意力结果进行拼接,再通过线性变换获得最终的多头注意力结果。在长距离依赖上,因为自注意力需要计算每个输入位点与其他所有位点之间的关联,所以能够忽略位点之间的距离影响从而完整地学习到一个序列的内部结构信息。本文对多头注意力层中头的数量对最终模型预测结果的影响进行了比较实验,选取能产生最优实验结果的结构作为 deepScore- α 最终的模型结构。本文所使用的 multi-head self-Attention ResNet20 如图2所示,最终输出的 b 、 y 代表 b 离子和 y 离子的离散化相对强度。

图2 deepScore- α 模型结构Fig. 2 Structure of deepScore- α model

1.2 deepScore- α 分数计算方法

deepScore- α 所使用的打分算法是一种基于概率和的打分方法。假设候选肽氨基酸序列为 $X = \{A_1, A_2, \dots, A_n\}$ (其中 n 为肽序列总长度), 通过特征提取后获得对应的特征向量 $F = \{x_1, x_2, \dots, x_{n-1}\}$ 。将特征向量序列作为输入, 通过模型获得概率矩阵 P 。使用候选肽序列在谱图中进行谱峰标注, 获得 b 离子及 y 离子的相对谱峰强度 $N = \{n_1, n_2, \dots, n_o\}$ (o 为该肽序列产生的碎片离子总数), 再通过以下公式进行离散化:

$$Y(n_o) = \begin{cases} \lceil n_o * 10 \rceil, & n_o \neq 0 \\ 0, & n_o = 0 \end{cases} \quad (1)$$

得到最终 b/y 离子对应的离散化相对强度序列 $Y = \{y_1, y_2, \dots, y_o\}$ 。使用以下公式计算出最终的肽谱匹配得分:

$$\text{score}(Y, P) = \sum_{i=1}^o P_{iy_o} \times Q(Y) \quad (2)$$

其中: n 为肽序列产生的所有 b/y 离子总数; P 为模型输出的碎片离子相对强度分布概率矩阵, P_{iy_o} 表示第 i 个碎片离子实际的离散化相对强度 y_o 在概率矩阵中对应的概率; $Q(Y)$ 为肽谱匹配质量系数计算函数。 $Q(Y)$ 具体计算式如下:

$$Q(Y) = (L - L_0 + 1) / (L + 1) \quad (3)$$

其中: L 为离散化相对强度序列 Y 的长度, L_0 为其序列中相对强度为 0 的数量。

1.3 Comet 及 MSGF+ 打分算法

本文使用开源鉴定工具 Comet 以及 MSGF+ 作为对比。其中, Comet 工具所采用的打分算法将实验质谱谱图转换为稀疏矩阵, 并使用了快速交叉关联 (Fast Cross Correlation) 算法来计算实验质谱和理论质谱的相似度得分, 再通过得到的相似度得分结合全局信息计算出 e -value 作为最终肽谱匹配得分; MSGF+ 工具所采用的打分算法则将肽序列转化为肽序列向量, 再使用概率模型将质谱转换为质谱向量, 通过计算肽序列向量与质谱向量的点乘给出最终的肽谱匹配得分。以上两种鉴定工具所使用的打分算法均可看作是一种计算肽序列与实验质谱相似度的打分算法, 本文提出的 deepScore- α 则通过深度学习模型获得肽序列对应的碎片离子相对强度的概率和

进行打分, 无需计算肽序列与实验质谱的相似度, 打破了相似度计算方法的限制。

2 实验与结果分析

2.1 数据集介绍及处理

本文用于模型训练的数据集为人类蛋白组数据集 (HumanProteome), 来自 Wilhelm 等^[17]在 2014 年关于人体蛋白组鉴定的相关工作 (PRIDE 数据集标识符: PXD000865), 包含人体 26 个组织的蛋白质二级串联质谱谱图及相应的鉴定结果。在一般的肽序列鉴定流程中, 蛋白质样品首先通过质谱仪获得原始质谱谱图数据 (数据文件格式通常为 raw), 再使用相应的鉴定工具给出谱图的对应得分最高的肽段作为鉴定结果。由于鉴定软件给出的鉴定结果仍会存在错误, 于是本文通过设置阈值条件 $q\text{-value} \leq 0.001$ 及 $PIF \geq 0.7$ 对鉴定结果进行过滤, 并且在谱图存在多个鉴定结果时选取后验错误率 (Posterior Error Probability, PEP) 较低者作为最终的鉴定结果以保证鉴定结果的可信度 (q -value、蛋白质水解诱导因子 (Proteolysis Inducing Factor, PIF) 和 PEP 均为相应的统计学指标)。碰撞能量 (Norm Collision Energy, NCE) 为质谱仪器一种重要参数, 本文使用碰撞能量对该数据集进行划分, 为确保模型最终的训练效率, 去除掉数据量过少的碰撞能量为 40 的部分, 最终获得两个训练数据集: 790 271 条 NCE 为 30 的肽序列数据集, 101 847 条 NCE 为 35 的肽序列数据集。随后使用标注误差限 (碎片离子理论质荷比与实际谱图种谱峰质荷比的差值) 为 20 ppm 的条件对过滤处理后得到的数据集进行标注, 获得肽序列对应的碎片离子谱峰强度, 肽段碎裂会产生多种碎片离子, 本文只考虑占比较大的 b/y 型碎片离子。

在进行打分效果评估时, 对所有的原始谱图文件 (raw) 进行抽样, 并利用 pParse (v2015, 一种谱图提取工具) 提取出抽取原始谱图文件中的谱图, 再使用 Comet (v2018014) 和 MSGF+ (v2019.07.03) 对提取的谱图进行鉴定。由于本文只针对肽谱匹配打分算法进行研究, 所以为保证算法效果评估的有效性, 在进行打分时对鉴定工具输出的每张谱图的前 50 个较高分数候选肽段进行重新打分。在进行打分 Top1 命中率 (在已知正确肽序列的情况下正确肽序列在该谱图所有候选肽中得分最高的谱图所占比例) 实验时, 为了保证命中率计算结果的准确性, 分别从碰撞能量为 30 和 35 的模型测试集中抽取 10 000 张原始鉴定结果可信度比较高的谱图进行打分实验。用于算法泛化性能测试的数据集来自 ProteomeTools2 (PRIDE 数据集标识符: PXD010595)^[18], 同样对所有原始谱图数据进行随机抽样, 再使用 Comet 和 MSGF+ 进行鉴定, 最后使用 deepScore- α 对 Comet 以及 MSGF+ 输出的候选肽段进行重新打分并评估打分结果, 由于 ProteomeTools2 数据集原始鉴定结果较为准确, 所以直接进行 Top1 命中率评估。

2.2 特征提取及模型训练结果

在自然语言处理 (Natural Language Processing, NLP) 中, 对单个字或词进行编码是十分重要的, 因此逐渐发展出了 WordEmbedding^[19-20] 这样的特征编码方式。而本文所涉及的肽碎裂事件中肽键可以被看作一个基本单位, 即可以借鉴自然语言处理中的编码方式对其进行特征编码, 通过类似 one-hot 类型的编码将氨基酸残基表示成 22 维的向量。在肽键碎裂过程中, 肽键左右相邻的氨基酸种类及肽序列中碱性氨基

酸和肽序列所带电荷的数量对肽键的碎裂起着至关重要的作用。本文综合考虑了以上因素,最终形成了以下 105 维的特征集(以某一肽键为例),具体特征集如表 1 所示。

表 1 肽键特征集
Tab. 1 Feature set of peptide bonds

名称	长度
C 端及 N 端氨基酸种类	46
肽键相邻氨基酸种类	46
肽键是否靠近肽序列某一端	1
肽序列中碱性氨基酸的数量	1
b 离子中碱性氨基酸的数量	1
y 离子中碱性氨基酸的数量	1
肽键距离肽序列 C 端和 N 端的距离	2
肽序列的总长度	1
肽序列所带电荷量	5
肽序列的电子迁移率	1

本文也探索了多头自注意力机制中头(head)的数量对最终模型的准确率产生的影响,在人类蛋白组数据集上分别进行了不同 head 数量的实验,并且与逻辑回归以及支持向量机(Support Vector Machine, SVM)进行了比较,实验结果如表 2 所示。实验结果表明,本文采用的模型预测准确率明显优于逻辑回归和支持向量机,更有效地学习到了肽碎裂规律,且在不加入多头注意力机制时 ACC(Accuracy, 预测的相对强度值与真实相对强度值相等的样本占所有样本的比例)为 81.56%;加入多头注意力机制后,随着 head 数量的增加,模型预测结果逐步提升;当 head 为 8 时,模型准确率达到最优,相较于未加入注意力机制的测试结果提升了约 2.6 个百分点;随着 head 数量的进一步增多,模型准确率出现下降,于是 deepScore- α 后续采用了 head 为 8 的模型进行打分实验。

表 2 不同模型在测试集上的预测准确率
Tab. 2 Prediction accuracy of different models on test set

模型	准确率/%
ResNet	81.56
ResNet+Attention(head=1)	82.27
ResNet+Attention(head=2)	82.52
ResNet+Attention(head=4)	83.69
ResNet+Attention(head=8)	84.16
ResNet+Attention(head=16)	83.51
Logistics regression(multiclass)	69.61
SVM	66.11

模型训练中学习率设置为 0.000 1,使用 Adam 优化器,为避免模型过拟合,权重衰减设置为 0.000 1,最终在测试集上测试结果如图 3 及图 4 所示。因为模型在训练时预测的是离散化的碎片离子相对强度,所以预测值与真实值相差大小为 ± 1 时也可以认为预测有效。因此在评估模型预测效果时,除了将传统的 ACC 作为评价标准,还需要考虑到预测值与真实值偏差范围为 ± 1 的部分,从测试结果可以看出,模型在两个数据集上均获得了不错的预测效果,ACC 与 ± 1 的部分所占比例之和都达到了 97%。

2.3 肽谱匹配质量系数

对于某一实验质谱谱图,产生该谱图的正确肽序列相较于其他错误的候选肽序列往往能在谱图中标注出更多的非零

相对强度的离子峰,有更多的相对强度不为零的碎片离子存在,即非零相对强度的占该肽序列产生的所有碎片离子数量的比例更大。由于 deepScore- α 利用碎片离子离散化相对强度的概率和对候选肽进行打分,因此当正确肽序列长度短于错误候选肽序列时,即使错误候选肽序列有更多的相对强度为零的碎片离子,其概率和也可能超过正确肽序列。本文提出了肽谱匹配质量系数(PSM Quality Coefficient, PQC)用于降低肽序列长度及谱峰相对强度为零的碎片离子对打分算法的影响。

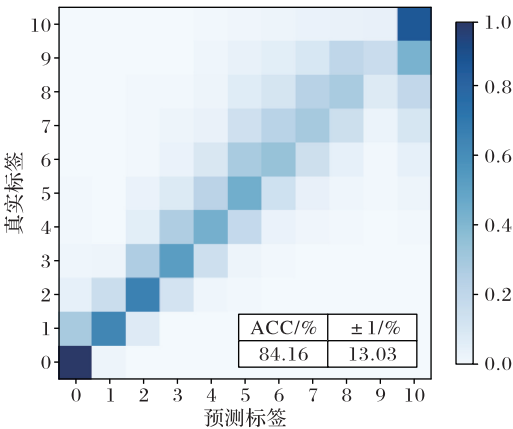


图 3 碰撞能量为 30 的数据集上模型训练结果
Fig. 3 Model training results on dataset with NCE of 30

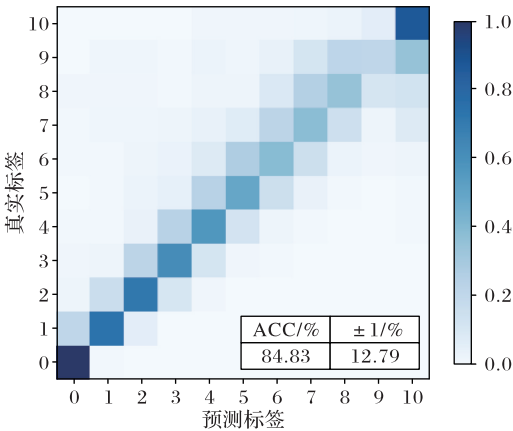


图 4 碰撞能量为 35 的数据集上模型训练结果
Fig. 4 Model training results on dataset with NCE of 35

2.4 deepScore- α 与 Comet 以及 MSGF+ 的打分效果比较

本文使用 Comet 和 MSGF+ 对碰撞能量为 30 及 35 的数据集中的随机抽取的原始数据文件(raw)进行鉴定,再使用 deepScore- α 对 Comet 以及 MSGF+ 输出的前 50 条鉴定结果进行重新打分并排序,最终的错误发现率(False Discovery Rate, FDR)曲线如图 5 和图 6 所示。通过比较可以看出,deepScore- α 在 FDR=0.01 时保留的肽序列数量相较于 Comet 和 MSGF+ 最高提升了约 14%,在 30 碰撞能量的情况下完全优于 Comet 和 MSGF+,而在 35 碰撞能量的情况下某些部分基本与 Comet 和 MSGF+ 持平。其原因应该是 35 碰撞能量的训练数据过少导致,30 碰撞能量最终用于训练测试的包含 790 271 条肽序列,而 35 碰撞能量训练集只有 101 847 条肽序列,并且由于 deepScore- α 是利用概率进行打分,其对标签分类的概率更加敏感,因此交叉熵(Cross Entropy)更能反映模型对肽碎裂规律的学习程度。在最终用于打分的模型中,碰撞能

量 30 的模型在测试集上的最小交叉熵为 35.92,而在碰撞能量 35 中测试集的最小交叉熵为 40.84,所以在利用概率进行打分时,deepScore- α 在 35 碰撞能量的数据集上的表现比 30 碰撞能量差一些,如果能有效地扩大 35 碰撞能量的数据集规模,使其与 30 碰撞能量的数据集规模一致,使交叉熵降低至相同水平,其最终打分表现也应该一致。

在打分阶段比较打分算法有效分辨正确候选肽与错误候选肽的能力时,评估比较其 Top1 命中率是十分有必要的。本

文从划分为模型测试的数据集中随机抽取 10 000 张谱,再使用 Comet 和 MSGF+ 对这 10 000 张谱图进行重新打分,deepScore- α 对 Comet 和 MSGF+ 输出前 50 的鉴定结果进行重新打分排序,最终的 Top1 命中率比较如图 7 所示。经过比较发现,deepScore- α 的 Top1 命中率在 30 碰撞能量以及 35 碰撞能量的情况下都比 Comet 和 MSGF+ 更高,在 30 碰撞能量的情况下更高出了 5 个百分点,因此 deepScore- α 识别正确肽序列的能力要强于 Comet 和 MSGF+。

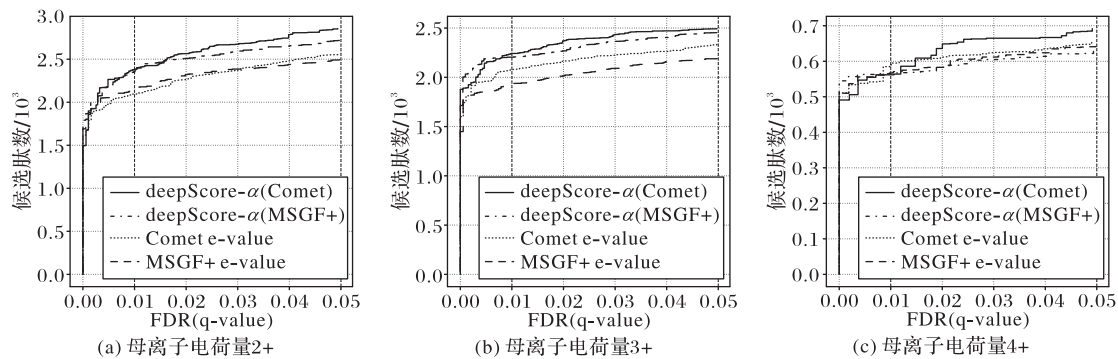


图5 deepScore- α 、Comet和MSGF+在碰撞能量为30的人类蛋白组数据集上的FDR曲线

Fig. 5 FDR curves of deepScore- α , Comet and MSGF+ on humanbody preteome dataset with NCE of 30

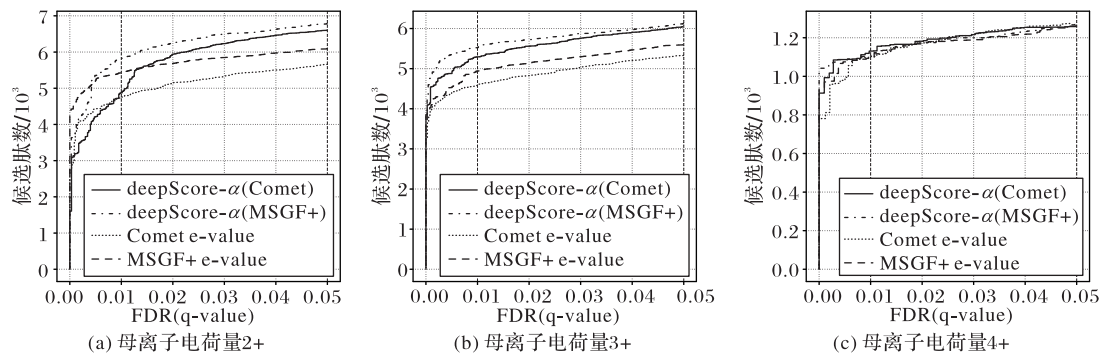


图6 deepScore- α 、Comet和MSGF+在碰撞能量为35的人类蛋白组数据集上的FDR曲线

Fig. 6 FDR curves of deepScore- α , Comet and MSGF+ on humanbody preteome dataset with NCE of 35

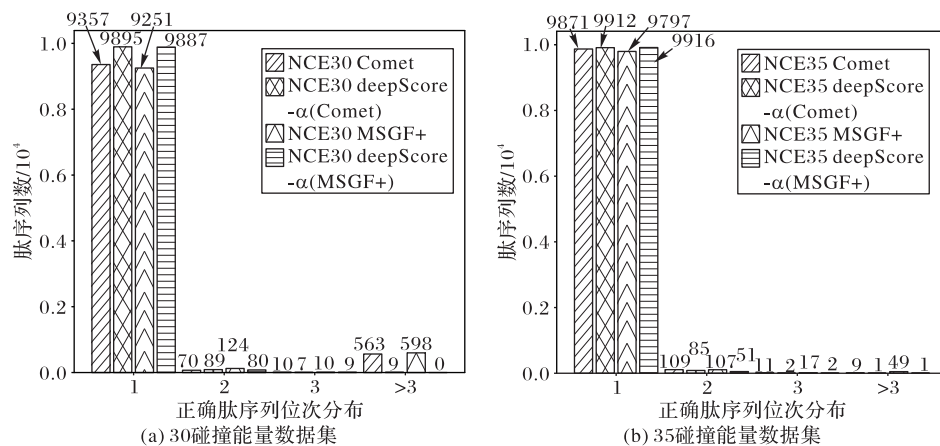


图7 deepScore- α 、Comet和MSGF+打分结果中正确候选肽的分布

Fig. 7 Distribution of correct candidate peptides in deepScore- α , Comet and MSGF+ scoring results

2.5 deepScore- α 泛化性能分析

打分算法的泛化性能是十分重要的,好的泛化性能能极大地提高打分算法在不同数据集上的表现而不只是在特定某一数据集下表现良好。本文首先在 Humanbody 数据集上训练模型,再使用训练完成的模型对 ProteomeTools2 中随机抽取

的谱图进行打分并与 Comet 和 MSGF+ 进行比较以评估其泛化性能。同样先使用 Comet 和 MSGF+ 进行鉴定,再使用 deepScore- α 对二者输出的前 50 条鉴定结果进行重新打分和排序, FDR 曲线如图 8 和图 9 所示。通过分析比较发现, deepScore- α 在 FDR=0.01 条件下保留的肽序列相较于 Comet

和 MSGF+ 最高提升了约 7%, 在 30 碰撞能量下完全优于 Comet, 在 35 碰撞能量下存在部分与 Comet 基本持平的情况, 与在 Humanbody 数据集上进行打分的表现一致, 应该为 35 碰撞能量数据量相较 30 碰撞能量不足所致。综合来看, deepScore- α 具有较好的泛化性能, 利用 Humabody 数据集训练的模型最终在 ProteomeTools2 上的打分效果基本与原数据集上的一致, 都优于 Comet 和 MSGF+。同样的, 打分算法的

Top1 的命中率也是其性能评价的一个重要指标。本文也评估了 deepScore- α 在 ProteomeTools2 数据上的 Top1 命中率, 实验结果如表 3 所示, deepScore- α 在 ProteomeTools2 数据集上的 Top1 命中率相较于另外两个鉴定工具均提高了 5 个百分点, Top1 的鉴定结果中 Decoy 的数量与另外两个鉴定工具相比减少了 60% 左右, 因此可以判断 deepScore- α 在打分性能上优于 Comet 和 MSGF+, 且具有较好的泛化性能。

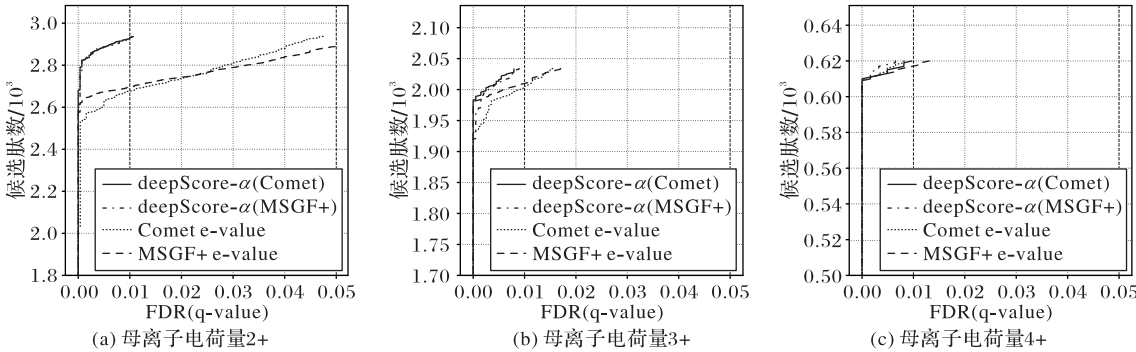


图 8 deepScore- α 、Comet 和 MSGF+ 在碰撞能量为 30 的 Proteome Tools2 数据集上的 FDR 曲线
Fig. 8 FDR curves of deepScore- α , Comet and MSGF+ on Proteome Tools2 dataset with NCE of 30

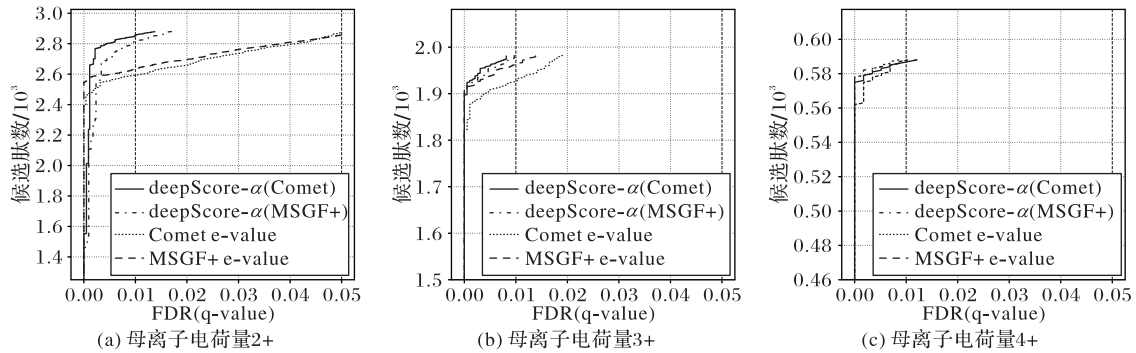


图 9 deepScore- α 、Comet 和 MSGF+ 在碰撞能量为 35 的 Proteome Tools2 数据集上的 FDR 曲线
Fig. 9 FDR curves of deepScore- α , Comet and MSGF+ on Proteome Tools2 dataset with NCE of 35

表 3 deepScore- α 、Comet 和 MSGF+ 在 ProteomeTools2 中打分结果比较
Tab. 3 Comparison of deepScore- α , Comet and MSGF+ scoring results in ProteomeTools2

NCE	质谱数	打分算法	Top1 命中率/%	Top1 中来自 Decoy 库的候选肽数
30	5 632	Comet	75.31	169
		deepScore- α (Comet)	80.04	53
		MSGF+	76.22	214
		deepScore- α (MSGF+)	80.66	52
35	5 489	Comet	75.36	182
		deepScore- α (Comet)	80.09	74
		MSGF+	76.43	183
		deepScore- α (MSGF+)	80.62	62

3 结语

deepScore- α 通过深度学习模型学习肽碎裂规律预测出碎片离子离散化相对强度分布概率, 结合实际的碎片离子离散化相对强度形成肽谱匹配得分, 有效地利用了深度学习的优势。deepScore- α 在人类蛋白组数据集上的表现优于常用开源鉴定工具 Comet 和 MSGF+, 且使用人类蛋白组数据集训练的模型最终在 ProteomeTools2 上的表现也优于另外两个鉴

定工具, 均可以证明 deepScore- α 是一个打分效果优良且泛化能力较强的深度学习打分算法。下一步研究重点是拓展使用的碎片离子范围, 本文在 deepScore- α 打分算法中只使用了一价和二价的 b/y 离子, 虽然在肽序列碎裂后产生的碎片离子中 b/y 离子占大多数, 但是其产生诸如 a/x、c/z 离子以及相应的内部离子是否对打分算法产生影响还需要进一步验证。

参考文献 (References)

[1] KAPP E, SCHÜTZ F. Overview of tandem Mass Spectrometry (MS/MS) database search algorithms [J]. Current Protocols in Protein Science, 200749(1): 25. 2. 1-25. 2. 19.

[2] ENG J K, JAHAN T A, HOOPMANN M R. Comet: an open-source MS/MS sequence database search tool [J]. Proteomics, 2013, 13(1): 22-24.

[3] ENG J K, HOOPMANN M R, JAHAN T A, et al. A deeper look into Comet - implementation and features [J]. Journal of The American Society for Mass Spectrometry, 2015, 26(11): 1865-1874.

[4] KIM S, PEVZNER P A. MS-GF+ makes progress towards a universal database search tool for proteomics [J]. Nature Communications, 2014, 5: Article No. 5277.

[5] PERKINS D N, PAPPIN D J C, CREASY D M, et al. Probability-based protein identification by searching sequence databases using

- mass spectrometry data [J]. *Electrophoresis*, 1999, 20(18): 3551-3567.
- [6] 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): 1367-72.
- [7] BAI W, BILMES J, NOBLE W S. Bipartite matching generalizations for peptide identification in tandem mass spectrometry [C]// *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. New York: ACM, 2016: 327-336.
- [8] BAI W, BILMES J, NOBLE W S. Submodular generalized matching for peptide identification in tandem mass spectrometry [J]. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019, 16(4): 1168-1181.
- [9] BEPLER T, BERGER B. Learning protein sequence embeddings using information from structure [EB/OL]. [2019-03-22]. <https://arxiv.org/pdf/1902.08661.pdf>.
- [10] WANG S, PENG J, MA J, et al. Protein secondary structure prediction using deep convolutional neural fields [J]. *Scientific Reports*, 2016, 6: Article No. 18962.
- [11] ZHOU X, ZENG W, CHI H, et al. pDeep: predicting MS/MS spectra of peptides with deep learning [J]. *Analytical Chemistry*, 2017, 89(23): 12690-12697.
- [12] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]// *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 2016:770-778.
- [13] SHENG W, SUM S, LI Z, et al. Accurate de novo prediction of protein contact map by ultra-deep learning model [J]. *PLoS Computational Biology*, 2017, 13(1): Article No. e1005324.
- [14] WANG S, LI Z, YU Y, et al. Folding membrane proteins by deep transfer learning [J]. *Cell Systems*, 2017, 5(3): 202-211.
- [15] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [EB/OL]. [2019-03-22]. <https://arxiv.org/pdf/1409.0473.pdf>.
- [16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [EB/OL]. [2019-03-22]. <https://arxiv.org/pdf/1706.03762.pdf>.
- [17] WILHELM M, SCHLEGL J, HAHNE H, et al. Mass-spectrometry-based draft of the human proteome [J]. *Nature*, 2014, 509(7502): 582-587.
- [18] GESSULAT S, SCHMIDT T, ZOLG D P, et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning [J]. *Nat Methods*, 2019, 16(6): 509-518.
- [19] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [EB/OL]. [2019-03-22]. <https://arxiv.org/pdf/1301.3781.pdf>.
- [20] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// *Proceedings of the 26th International Conference on Neural Information Processing Systems*. Red Hook: Curran Associates Inc., 2013:3111-3119.

This work is partially supported by the National Natural Science Foundation of China (31500669), the Shandong Provincial Natural Science Foundation (ZR2014FQ024), the Support Program for Outstanding Youth Innovation Teams in Higher Education of Shandong Province (2019KJN048).

MIN Xin, born in 1995, M. S. candidate. His research interests include deep learning, bioinformatics.

WANG Haipeng, born in 1980, Ph. D., associate professor. His research interests include machine learning, bioinformatics.

MOU Changning, born in 1990, M. S. candidate. His research interests include deep learning, bioinformatics.