doi:10.6043/j.issn.0438-0479.202005034

数据不依赖获取的质谱数据的深度学习分析方法

何情祖[1],钟传奇[2],李 翔[1],帅建伟[1,3*],韩家淮[2,3*]
(1.厦门大学物理科学与技术学院,福建 厦门 361005;2.厦门大学生命科学学院,福建 厦门 361102;
3.厦门大学健康医疗大数据国家研究院,福建 厦门 361102)

摘要:近年来,数据不依赖获取(data-independent acquisition,DIA)质谱技术在蛋白质组学领域内被广泛关注.然而DIA质谱数据具有维度高、背景噪声大、多种信号混合等特点,这使得DIA质谱数据的分析成为一大挑战.本文提出一种基于深度学习的可直接处理DIA质谱数据的算法:Ultra-DIA.该算法使用深度变分自动编码器提取离子信号的特征来区分不同肽段产生的子离子,最终生成虚拟谱图,进而对肽段和蛋白进行定性和定量分析.对于测试数据,该算法找到的肽段数量和蛋白数量比主流算法DIA-Umpire分别多61.4%和64.5%.此外,相较于DIA-Umpire,该算法能够找到更多低浓度的蛋白.

关键词:深度学习;变分自动编码器;数据不依赖获取;质谱数据
中图分类号:Q 633 **文献标志码**:A **文章编号**:0438-0479(2021)01-0097-07

蛋白质谱仪的采集策略可以分为数据依赖获取(data-dependent acquisition,DDA)和数据不依赖获取(data-independent acquisition,DIA)两种.鸟枪(shotgun)法是一种典型的DDA采集模式[1].在鸟枪法实验中,质谱仪自动选择强度排名前N名的肽离子进行破碎,并记录所得的碎片离子质谱图.由于鸟枪法只选择强度排名靠前的肽段进行打碎,具有一定随机性,这使得鸟枪法检测到的肽段重复性较差.

为了解决DDA检测到的肽段重复性差的问题,2012年瑞士Gillet等[2]与AB-SCIEX公司联合开发出DIA相应的采集模式——全碎片离子顺序窗口化获取(sequential windowed acquisition of all theoretical fragment ions,SWATH).SWATH是一种典型的DIA模式,与传统的质谱数据采集方法相比,SWATH模式通过高速扫描每个荷质比窗口内所有的肽离子,并将其裂解成子离子,从而获得肽段和碎片完整的信息.相比于DDA,DIA模式具有通量高、特异性好、灵敏度高、定量结果重复性高等特点.但是DIA的质谱数据极其复杂,不同肽段的信号混合在同一张谱图内,加大了数据分析的难度.

为了分析DIA质谱数据,相关算法的开发一直是研究热点.2014年Rost等[3]提出一种文库依赖的SWATH质谱数据分析方法:OpenSWATH.该方法先使用DDA质谱数据进行数据库搜索,建立肽段的目标文库,然后根据目标文库对DIA质谱数据进行定量分析.2015年Tsou等[4]提出一款开源软件DIA-Umpire,直接分析DIA模式下的质谱数据,省去了DDA实验.在使用质谱技术研究生物学问题时,一般会制备多组样品,并且研究者主要关注不同样品间有差异的蛋白.同年,厦门大学韩家淮课题组[5]提出一种DIA的质谱数据分析方法.该方法通过计算不同组实验样品之间肽段产生的母离子和子离子信号的相关性,从而生成虚拟谱图并通过搜索引擎来鉴定肽段.同时Wang等[6]提出另一种用于DIA质谱数据的匹配分析方法MSPLIT-DIA(mixture-spectrum partitioning using libraries of identified tandem mass spectra-DIA).

此外,近两年也出现了一些基于深度学习的质谱数据分析算法.2019年Tran等[7]提出了DeepNovo-DIA算法,将深度学习和从头测序(de-novo sequencing)

收稿日期:2020-05-27 **录用日期**:2020-08-12
基金项目:国家自然科学基金(11874310,11675134)
*通信作者:jianweishuai@xmu.edu.cn(帅建伟);jhan@xmu.edu.cn(韩家淮)
引文格式:何情祖,钟传奇,李翔,等.数据不依赖获取的质谱数据的深度学习分析方法[J].厦门大学学报(自然科学版),2021,60(1):97-103.
Citation:HE Q Z,ZHONG C Q,LI X,et al. Deep learning analysis for data-independent acquisition mass spectrometry data[J]. J Xiamen Univ Nat Sci,2021,60(1):97-103. (in Chinese)

法结合起来,对 DIA 质谱数据直接进行肽段氨基酸序列的鉴定. 同年,Gessulat 等[8]提出了名为 Prosit 的基于深度学习的算法,直接从蛋白质数据库中理论预测肽段的驻留时间和子离子强度,同样摒弃了 DDA 实验. 2020 年 Demichev 等[9]提出 DIA-NN 算法,使用深度学习替代 OpenSWATH 工作流中打分步骤,以期发现更多的肽段和蛋白. 同年,Yang 等[10]提出了 DeepDIA 算法,同样使用深度学习从蛋白质数据库中预测肽段的驻留时间和子离子强度,其性能要强于同类的 Prosit 算法.

为了能够直接处理 DIA 质谱数据,生成虚拟谱图,不需要 DDA 数据进行建库,且输出文件能够与现有的肽段定性定量工作流兼容,本研究提出了基于深度学习的 DIA 质谱数据分析算法:Ultra-DIA.

1 Ultra-DIA 算法分析流程

1.1 深度变分自动编码器

自动编码器是深度学习中一种重要的无监督学习模型,主要用于数据的降维和特征抽取,它的本质是利用非线性神经网络生成的低维特征来代替高维输入. 常用的自动编码器有深度变分自动编码器(deep variational autoencoder,DVAE)[11-12]、堆栈自动编码器(stacked autoencoder,SAE)[13]和降噪自动编码器(denoising autoencoder,DAE)[14]等. DVAE、SAE 和 DAE 的结构分别如图 1 所示,其编码器和解码器都为多层全连接网络. 其中,SAE 和 DAE 常用的损失函数为均方差损失函数(mean square error loss,MSELoss). 相较于 SAE 的简单结构,DAE 通过随机擦除输入数据的值来学习鲁棒性更好的特征,而 DVAE 则进一步考虑高斯随机采样的操作,结合贝叶斯定理和 KL 散度(Kullback-Leibler divergence)来约束低维特征分布,从而提升了特征的表达能力和样本重构的表现. Ultra-DIA 采用 DVAE 作为深度学习模型.

自动编码器实现的映射为:

$$\phi: X \to Z,$$

$$\varphi: Z \to \hat{X},$$

其中,Z 为隐变量,也表示输入数据的特征. 编码器 ϕ 将输入数据 X 映射到 Z 分布,解码器 φ 将 Z 分布重构成 $\hat{X}$. 神经网络的训练目标就是使得 $\hat{X}$ 不断接近 X,公式如下:

$$\arg\min_{\phi,\varphi} L(X,\hat{X}).$$

如图 1 所示,DVAE 主要分成以下 3 大模块:编码器、高斯随机采样器和解码器.

假设存在 Z 的分布,但无法确定它的具体形式,只能期望在给定 X 的情况下推出 Z 分布,即 $p(Z|X)$. 根据贝叶斯定理,可得:

$$p(Z \mid X) = \frac{p(X \mid Z)p(Z)}{p(X)}.$$

由于 $p(X)$ 表示输入数据 X 的未知的概率分布,所以 $p(Z \mid X)$ 无法直接计算,那么可以使用一个可解的分布 $q(Z \mid X)$ 去近似 $p(Z \mid X)$. 而 KL 散度 D 可以用来描述两个分布的近似程度,其形式如下:

$$D(p(X) \parallel q(X)) = \int p(X) \ln \frac{p(X)}{q(X)} \mathrm{d}X.$$

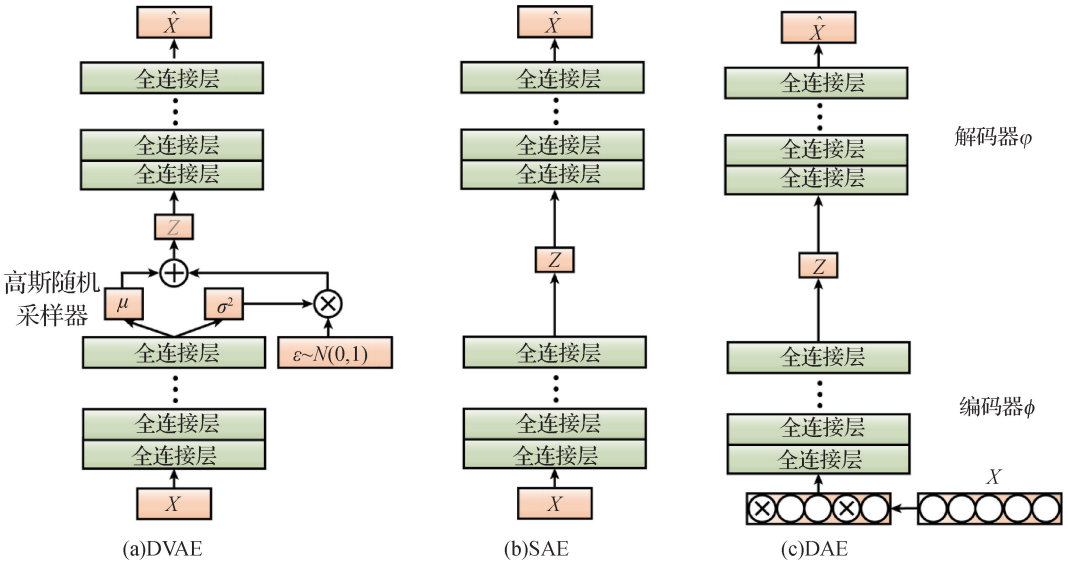

图 1 3 种自动编码器结构示意图

Fig. 1 Schematic diagram of three kinds of autoencoder

结合贝叶斯定理和 KL 散度,最终可以得到 DVAE 的损失 L:

$$L = -D(q(Z \mid X) \parallel p(Z)) + E_{Z\sim q(X\mid Z)}[\log\ p(X \mid Z)],$$

其中 E 表示期望. 假设 $q(Z \mid X)$ 服从均值为 μ,方差为 σ^2 的高斯分布 $N(\mu,\sigma^2)$,而 $p(Z)$ 服从标准正态分布 $N(0,1)$,那么 L 的第一项为:

$$-D(q(Z \mid X) \parallel p(Z)) = \frac{1}{2}(\log \sigma^2 - \mu^2 - \sigma^2 + 1).$$

假设输入数据服从正态分布,则 L 第二项中的 $p(X|Z)$等价于 $p(X \mid \hat{X})$,可得:

$$E_{Z\sim q(X\mid Z)}[\log\ p(X \mid Z)] = \frac{1}{N}\sum \parallel X - \hat{X} \parallel^2.$$

由于对 Z 的"采样"操作不可求导,故 DVAE 引入重参数化技巧(reparameterization trick),从正态分布 $\varepsilon\sim N(0,1)$中获取随机数 ε,使得采样的操作不参与梯度下降,只让采样的结果参与即可. 重参数化操作为:

$$Z = \mu + \varepsilon\ \sqrt{\sigma^2}.$$

综上所述,DVAE 的前向传播过程为:输入 X 给编码器,将编码器输出等分为两份,其中一份看成输入数据分布的均值 μ,另一份看成方差 σ^2. 通过重参数化操作获得隐变量 Z 并输入给解码器,最终由解码器重构出 $\hat{X}$. 通过计算 L 的值,使用梯度下降法将误差反向传播来调整编码器和解码器的权重参数.

1.2　Ultra-DIA 分析流程

如图 2 所示,Ultra-DIA 的分析流程分为 5 步.

1) 制备样品并通过 AB SCIEX 公司的 TOF-5600 质谱仪采集 SWATH 数据. 由于原始数据为闭源的 wiff 格式,无法直接读取数据内容,所以需要使用质谱数据分析软件 ProteoWizard 包含的 msconvert[15] 转换工具将 wiff 文件转化为开源的 mzXML 格式文件.

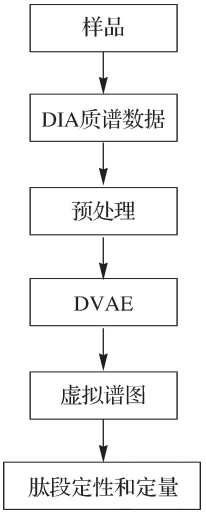

图 2　Ultra-DIA 分析流程

Fig. 2　Analysis workflow of Ultra-DIA

2) 对 DIA 质谱数据进行预处理,将噪音信号通过信噪比阈值进行过滤,再对过滤后的一级质谱(MS1)和二级质谱(MS2)峰型信号进行归一化,消除数量级对后续分析的影响.

3) 使用 DVAE 提取母离子和子离子的特征,进而对不同子离子进行分类. 经过深度学习的识别后,可以得到母离子和子离子的组合和对应的驻留时间,即可生成虚拟谱图.

4) 基于虚拟谱图,使用 Comet[16-17] 和 X!Tandem[18] 等搜索引擎对这些谱图进行数据库搜索,得到肽段的定性结果.

5) 最后使用 OpenSWATH 工作流对肽段的定性结果进行定量分析,最终得到肽段和蛋白的定性和定量结果.

2　结果与分析

2.1　数据情况

为了验证算法的性能,本研究制备了一个小鼠细胞的 SWATH 数据,梯度时间为 30 min. 这个数据包含 100 个 MS1 窗口,MS1 质荷比范围为 400～1 200,MS2 的质荷比范围为 100～1 800. 在扫描时间上,MS1 的扫描时间为 250 ms,MS2 的扫描时间为 33 ms,一个循环的总时间约为 3.6 s.

质谱数据包含 3 个维度的信息:质荷比、驻留时间和强度. 质谱仪会采集不同时刻的谱图信号,将这些谱图按照时间排列就可以得到具有不同质荷比的离子的信号随着时间的变化曲线,称为色谱曲线.

在 SWATH 实验中,混合蛋白质样品被酶解后进入色谱柱进行初步分离,随后再依次进入质谱仪. 因此,肽段的信号会在驻留时间方向上具有色谱峰的特征. 色谱峰是判断肽段是否存在的重要依据. 肽段离子的真实色谱峰呈现出高斯曲线的特征,但其一般不对称且在峰开始或者结束位置会出现干扰峰. 一般而言,高信噪比的色谱峰表明此时记录到了离子信号,而低信噪比的色谱峰则是背景离子产生的干扰信号. 在数据预处理阶段要针对色谱峰进行过滤,保留高信噪比的离子.

不同质荷比的母离子产生的色谱峰可能互不相同,如图 3(a)所示,图中包含了 5 000 个母离子在 1 056～2 816 s 内的提取离子流(XIC)色谱曲线. 不同

颜色的曲线表示不同荷质比的母离子. 每条曲线都可能包含多个色谱峰,每个峰都代表不同的肽段离子. 在同一条色谱曲线上的不同峰表示具有相同质荷比的不同肽段. 从图中可以看出,SWATH 数据的 MS1 信号非常复杂,直接通过母离子来推断肽段序列非常困难,所以需要 MS2 的信息来共同鉴定肽段. 而在图 3(b)中,这是第 50 个 MS1 窗口内由母离子打碎产生的子离子的 XIC 色谱曲线. 可以看出,子离子色谱曲线与母离子色谱曲线具有相似的特征,并且信噪比要高于母离子色谱曲线.

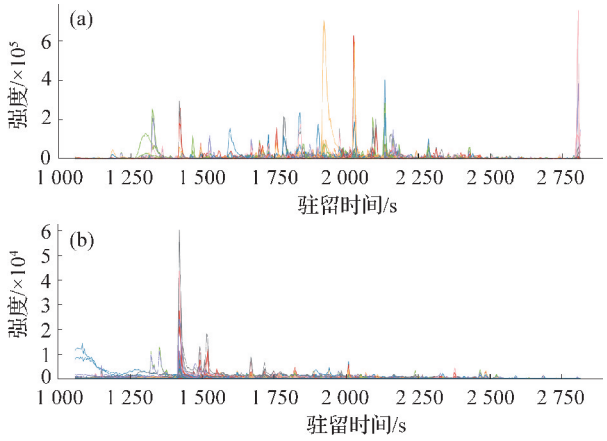

图 3 典型的母离子(a)和子离子(b)的色谱图

Fig. 3 Typical chromatograms of precursors (a) and fragments (b)

2.2 DVAE 的训练

深度学习的训练是使用神经网络的重要一步,在设计好误差函数后,就可以通过优化器不断减小误差,从而达到误差最小. 本研究对比了常见的 6 种深度学习优化算法,分别为 AdaDelta、Adagrad、自适应矩估计(adaptive moment estimation, Adam)、Adamax、Nadam 和随机梯度下降(stochastic gradient descent, SGD)算法[19].

如图 4 所示,除 SGD 算法外,其余优化算法的训练误差曲线在训练约 25 次之后开始平稳下降,并且降幅越来越小,最后趋于稳定. 由于 Adam 算法收敛速度快,本研究最终选取它作为模型的优化器. 对于 Adam 算法,参数设置为学习率 0.001, $\beta_1 = 0.900$, $\beta_2 = 0.999$.

2.3 识别结果

2.3.1 肽段和蛋白的定性结果分析

Ultra-DIA 的输出文件经过搜索引擎 Comet 和 X!Tandem 进行数据库搜索,再使用 PeptideProphet、iProphet、ProteinProphet 这 3 个软件对数据库搜索结果进行验证打分,最后使用 Mayu.perl 脚本计算肽段水平的 1%错误发现率(FDR)的得分,用这个得分过滤肽段和蛋白即得到定性分析结果.

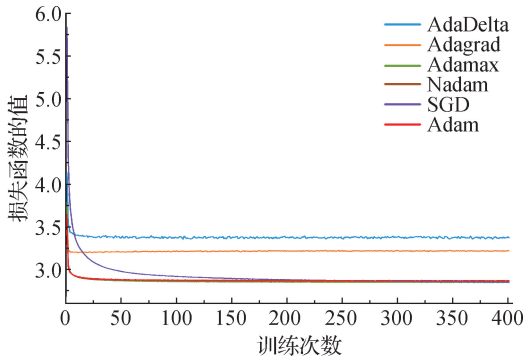

图 4 训练损失曲线

Fig. 4 Training loss curve

如图 5 所示,Ultra-DIA 分析 SWATH 数据后找到的肽段和蛋白数量均大于主流软件 DIA-Umpire. 在表示肽段的韦恩图中,交集部分的肽段覆盖了 DIA-Umpire 总肽段数的 78.7%;而在表示蛋白的韦恩图中,交集部分的蛋白对 DIA-Umpire 总蛋白数的占比达到 91.7%. 这说明在定性分析阶段,Ultra-DIA 能够重现出主流软件的结果,其找到的肽段和蛋白可信度高. 此外,在肽段水平上,Ultra-DIA 单独发现的数量约是 DIA-Umpire 单独发现的 3.6 倍,在蛋白水平上达到约 8 倍. 可见 Ultra-DIA 能够发现大量 DIA-Umpire 忽略的肽段和蛋白.

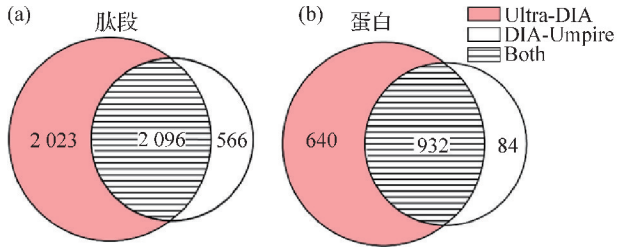

图 5 肽段(a)和蛋白(b)的定性分析结果

Fig. 5 Identification analysis results of peptides (a) and proteins (b)

此外,Ultra-DIA 总共发现了 4 119 个肽段和 1 572 个蛋白,DIA-Umpire 则仅为 2 662 个肽段和 1 016 个蛋白. 在肽段和蛋白的总数上,Ultra-DIA 均比 DIA-Umpire 增加了 54.7%,可见本文算法能够找到更多的肽段和蛋白.

2.3.2 肽段和蛋白的定量结果分析

定性分析后的肽段和蛋白存在着假阳性,仍然需

要进一步过滤,而过滤方式则是通过检测肽段能否定量来进行. 对于 SWATH 数据,能够定量的肽段和蛋白才是研究人员所关注的.

本研究将定性分析后生成的谱图库和 SWATH 数据输入给 OpenSWATH 软件. OpenSWATH 通过比较每个肽段对应的一组子离子色谱峰的相似度,对肽段进行评分;紧接着再通过 PyProphet 软件对 OpenSWATH 输出的评分结果进行综合,最终过滤子离子相似度低的肽段,并计算出通过筛选的蛋白数量. 通过筛选得到的肽段和蛋白即是可定量的肽段和蛋白.

如图 6 所示,Ultra-DIA 在肽段和蛋白的定量层次上,对 DIA-Umpire 的覆盖率分别为 79. 4% 和 94. 0%,相较于定性结果都有所上升. 这说明在定量层次上,Ultra-DIA 能够重现出 DIA-Umpire 的更多结果. 同时,Ultra-DIA 独自发现的肽段和蛋白数量也远大于 DIA-Umpire,这些被 DIA-Umpire 忽略的蛋白中可能隐藏着某些新蛋白. Ultra-DIA 共发现 3 830 个肽段和1 349个蛋白可以定量,而 DIA-Umpire 只发现 2 373 个肽段和 820 个蛋白可以定量. Ultra-DIA 在肽段总数上比 DIA-Umpire 多 61. 4%,在蛋白总数上多 64. 5%.

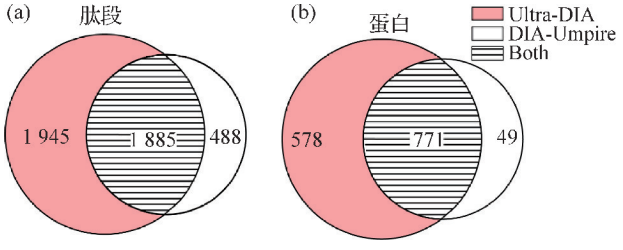

图 6　肽段(a)和蛋白(b)的定量分析结果

Fig. 6　Quantification analysis results of peptides (a) and proteins (b)

对比定性分析结果,Ultra-DIA 有 93. 0% 的肽段通过了定量筛选,而 DIA-Umpire 只有 89. 1% 的肽段通过了定量筛选. 同时,Ultra-DIA 发现的蛋白有85. 5% 可以定量,而 DIA-Umpire 发现的蛋白只有 80. 7% 可以定量. 这说明 Ultra-DIA 不仅找到了更多的肽段和蛋白,且其假阳性率更低,可信度更高.

图 7 展示了肽段的电荷分布,Ultra-DIA 和 DIA-Umpire 找到的肽段的电荷具有相似的分布,都是二价最多,三价次之,四价最少. 这样的分布符合仪器的设置和工作原理,说明 Ultra-DIA 找到的肽段可信度高.

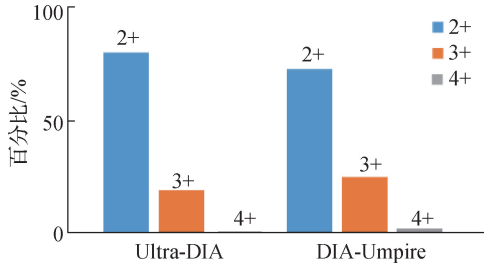

图 7　肽段电荷分布

Fig. 7　Peptide charge distribution

2. 3. 3　蛋白定量结果的丰度比较

在蛋白组学的研究中,样品中的低浓度蛋白一直是关注的重点. 许多关键蛋白或者新蛋白的浓度都很低,其质谱信号往往被噪声淹没,强度甚至接近噪声. 因此,发现的蛋白浓度越低说明算法的性能越优越.

OpenSWATH 的定量结果与蛋白的真实浓度成正比. 本研究将 OpenSWATH 对蛋白信号的定量结果取以 10 为底的对数,绘制蛋白信号强度的分布直方图,用来比较两种算法发现的蛋白浓度分布.

如图 8 所示,DIA-Umpire 发现的蛋白以较高浓度居多,而 Ultra-DIA 则可以发现更低浓度的蛋白. 这说明在分析相同数据的情况下,Ultra-DIA 发现关键蛋白或者新蛋白的能力要强于 DIA-Umpire.

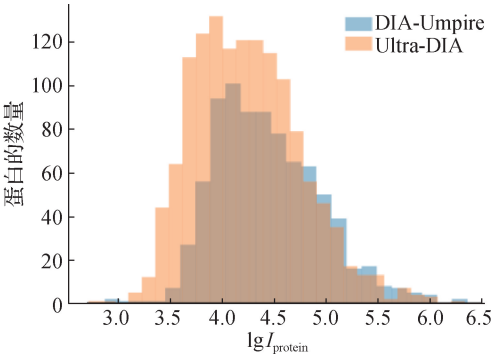

图 8　蛋白信号强度分布

Fig. 8　Protein signal intensity distribution

2. 3. 4　3 种自动编码器之间的比较

本研究对比了 DVAE、SAE 和 DAE 这 3 种自动编码器在测试数据上的肽段鉴定结果. 对于 SAE 和 DAE,还对比了 MSELoss 和二分类交叉熵损失函数(binary cross entropy loss, BCELoss)对结果的影响.

如图 9 所示,DVAE 鉴定到的肽段数量最多,其次是采用 BCELoss 的 DAE,因此在 Ultra-DIA 中本研究采用 DVAE 作为深度学习模型. 此外,采用 BCELoss 的 SAE 和 DAE 鉴定到的肽段均多于采用

MSELoss 的 SAE 和 DAE,这说明采用 BCELoss 提取到的特征更好.

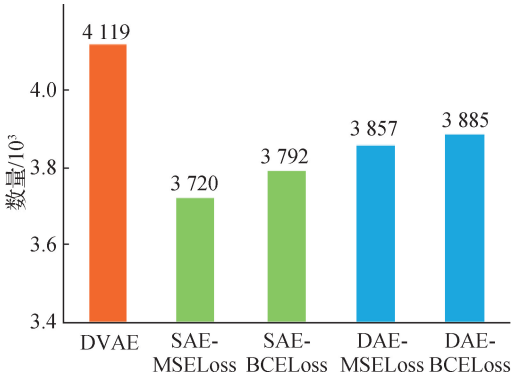

图 9 不同自动编码器的肽段鉴定结果

Fig. 9 Identification results of peptides by different autoencoders

3 结 论

本研究针对 DIA 数据,提出了基于深度学习的算法 Ultra-DIA. 该算法先对 DIA 数据进行预处理,然后使用深度 DVAE 分析 DIA 数据后生成虚拟谱图,再对虚拟谱图使用 Comet 和 X!Tandem 等搜索引擎进行数据库搜索,得到肽段和蛋白的定性分析结果和谱图库,最后使用 OpenSWATH 分析谱图库并对肽段和蛋白进行定量.

在测试数据集上,Ultra-DIA 的性能要优于 DIA-Umpire,但仍有许多可以改进的地方:1) 从质谱仪工作原理出发,对数据进行进一步的预处理;2) 神经网络的结构仍未达到最优,还可以设计更多不同类型的网络,通过对比最终找到的肽段和蛋白来确定哪种网络结构最合适;3) 神经网络的超参数仍未找到最优,需要不断测试;4) 测试数据的梯度时间较短,可以考虑使用不同梯度时间的数据来综合检验算法性能.

参考文献:

[1] ZHANG Y, FONSLOW B R, SHAN B, et al. Protein analysis by shotgun/bottom-up proteomics[J]. Chemical Reviews, 2013, 113(4): 2343-2394.

[2] GILLET L C, NAVARRO P, TATE S, et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis[J]. Molecular & Cellular Proteomics, 2012, 11(6): 016717.

[3] ROST H L, ROSENBERGER G, NAVARRO P, et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data[J]. Nature Biotechnology, 2014, 32(3): 219-223.

[4] TSOU C C, AVTONOMOV D M, LARSEN B, et al. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics[J]. Nature Methods, 2015, 12(3): 258-264.

[5] LI Y, ZHONG C, XU X, et al. Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files[J]. Nature Methods, 2015, 12(12): 1105-1106.

[6] WANG J, TUCHOLSKA M, KNIGHT J D R, et al. MSPLIT-DIA: sensitive peptide identification for data-independent acquisition[J]. Nature Methods, 2015, 12(12): 1106-1108.

[7] TRAN N H, 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.

[8] 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.

[9] DEMICHEV V, MESSNER C B, VERNARDIS S I, et al. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput[J]. Nature Methods, 2020, 17(1): 41-44.

[10] YANG Y, LIU X, SHEN C, et al. In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics[J]. Nature Communications, 2020, 11(1): 146-156.

[11] KINGMA D P, WELLING M. Auto-encoding variational bayes[EB/OL]. [2020-05-01]. http://arxiv.org/pdf/1312.6114.

[12] DOERSCH C. Tutorial on variational autoencoders[EB/OL]. [2020-05-01]. http://arxiv.org/pdf/1606.05908.

[13] BERNHARD S, JOHN P, THOMAS H. Greedy layer-wise training of deep networks[C]// Advances in Neural Information Processing Systems 19. Cambridge: MIT Press, 2007: 153-160.

[14] VINCENT P, LAROCHELLE H, LAJOIE I, et al. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion[J]. Journal of Machine Learning Research, 2010, 11(12): 3371-3408.

[15] CHAMBERS M C, MACLEAN B, BURKE R, et al. A cross-platform toolkit for mass spectrometry and proteomics[J]. Nature Biotechnology, 2012, 30(10): 918-920.

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

[17] 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.

[18] CRAIG R, BEAVIS R C J B. TANDEM: matching proteins with tandem mass spectra[EB/OL]. Bioinformatics, 2004, 20(9): 1466-1467.

[19] RUDER S. An overview of gradient descent optimization algorithms[EB/OL]. [2020-05-01]. http://arxiv.org/pdf/1609.04747.

Deep learning analysis for data-independent acquisition mass spectrometry data

HE Qingzu[1], ZHONG Chuanqi[2], LI Xiang[1], SHUAI Jianwei[1,3*], HAN Jiahuai[2,3*]

(1. College of Physical Science and Technology, Xiamen University, Xiamen 361005, China; 2. School of Life Sciences, Xiamen University, Xiamen 361102, China; 3. National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen 361102, China)

Abstract: In recent years, data-independent acquisition (DIA) mass spectrometry techniques have received wide attention in proteomics. However, DIA data are characterized with high dimensionality, large background noises, and mixing of multiple signals, which further challenge the analysis of DIA data. In this work, an algorithm based on deep learning that can directly process DIA mass spectrum data, namely Ultra-DIA, has been developed. It is combined with the deep variational auto-encoder and a variety of machine learning algorithms to directly process DIA data and to extract the features of MS ion signals, so that fragment ions generated by different peptides can be distinguished. Finally, Ultra-DIA generates pseudo-spectra to identify and quantify MS peptides and proteins. For the test data, our algorithm has found 61.4% more peptides and 64.5% more proteins than the mainstream algorithm of DIA-Umpire. In addition, our algorithm is capable of finding more proteins at low concentration compared to the DIA-Umpire.

Keywords: deep learning; variational autoencoders; data-independent acquisition; mass spectrometry data