PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2018, 45(12): 1214~1223
www.pibb.ac.cn

深度学习方法在生物质谱及蛋白质组学中的应用 *

赵新元[1] 秦伟捷[2)**] 钱小红[2)**]

([1]北京工业大学生命科学与生物工程学院，北京 100022；
[2]蛋白质组学国家重点实验室，北京蛋白质组研究中心，国家蛋白质科学中心(北京)，军事医学研究院生命组学研究所，北京 102206)

摘要 深度学习是近年来机器学习领域最热门的研究方向，尤其是在图像及语音识别、自然语言处理、自动驾驶等方面取得了突破性进展．生物质谱是当今生命科学领域重要的研究工具，尤其在蛋白质组学、代谢组学、生物制药等领域发挥着关键作用．近年来，基于深度学习方法的发展，以生物质谱为核心的蛋白质组学大数据分析将迎来发展新契机．本文综述了深度学习方法在生物质谱数据解析及蛋白质组学研究方面的最新应用．

关键词 深度学习，生物质谱，蛋白质组学，大数据
学科分类号 Q51，TP39
DOI: 10.16476/j.pibb.2018.0165

深度学习属于机器学习方法中的一种，其结构框架往往类比人类大脑，由大量神经元构成多层结构．作为机器学习领域的一个重要研究方向，深度学习近年来受到越来越多的关注．随着新型算法的快速开发以及以英伟达公司(NVIDIA corporation)为代表的图形处理器硬件产业的飞速发展，深度学习的研究成井喷之势，尤其是在图像识别、自然语言处理及自动驾驶方面取得了多个重大突破．深度学习即机器学习中的多层神经网络方法，主要结构包括输入层、隐藏层及输出层．图 1 示意了深度学习方法的基本结构．深度学习通过使用如反向传播算法等来发现大型数据中的复杂结构，以指示算法内部应如何优化其内部参数，经过模型参数的层层优化，从而提高数据处理的准确度[1-2]．按照神经网络隐藏层间关系，目前各学科领域主要使用两种神经网络结构，分别为卷积神经网络(convolutional neural network，CNN)和循环神经网络(recurrent neural network，RNN)．除此以外，亦有其他一些神经网络模型，如自动编码机、受限玻尔兹曼机及深度信念神经网络等．其中，CNN 在设计之初主要用于处理具有类似网格状结构的数据类型，例如传统的图片格式，其包括了对应像素的三颜色通道中的像素强度信息．CNN 主要有两种类型网络层，即卷积层与池化层．其中，卷积层用来提取数据的各种特征，可以使得模型在保留数据之间关系的同时大大降低参数数量，而池化层可以对提取到的特征进行抽象并降维处理，从而减少训练参数．通过使用如反向传播算法进行模型训练后，CNN 可以对网格状数据进行有效地理解并进行诸如图片分类、语音识别及自动驾驶等任务．CNN 最早由 Yann LeCun 等[3]提出并应用在手写字体识别上，继而在处理图像、视频、语音和音频方面取得了突破．其中，2016 年美国谷歌公司推出的 AlphaGo[4-5]即采用了 CNN 结合蒙特卡洛搜索树算法，一举击败了韩国九段围棋高手李世石，使得深度学习这一概念走进普通大众的视野．RNN 是另一种常见的深度学习架构，其从输入层得到信息后，会在隐藏层传递过程中包含上一层信息，实现信息的隐藏性传递，使得下一层单元中包含之前所有层的信息，

* 国家重点研发计划(2016YFA0501403, 2017YFC0906703), 国家自然科学基金(21675172)和蛋白质组学国家重点实验室自主课题(SKLP-K201706)资助项目．
** 通讯联系人．Tel: 010-61777107
钱小红．E-mail: qianxh1@163.com
秦伟捷．E-mail: aunp_dna@126.com
收稿日期：2018-06-13，接受日期：2018-09-10

从而使得模型更适合处理语言文字一类前后关联的数据类型[1]. 因此，RNN 在连续及关联数据中(如文本和语音类型数据)应用广泛.

质谱(mass spectrometry)是当前蛋白质组学研究中的核心技术，其分析能力近十几年来飞速发展[6-7]. 从电离技术上区分，当前的生物质谱主要分为基质辅助激光解吸附电离[8](matrix assisted laser desorption/ionization，MALDI)和电喷雾电离[9](electrospray ionization，ESI)两种方式. 得益于电喷雾电离质谱在肽段二级谱图获得上的高通量优势，高效液相色谱串联电喷雾电离质谱的分析方式成为蛋白质组学研究的主要工具[10-12]，其产出的生物质谱数据主要包括一级母离子谱图及二级碎片离子谱图，亦有针对特定分析需求而采取的三级及以上离子碎裂的质谱采集方式[13-14]，但总体比较少见. 二级质谱图为质谱中特定母离子经过如碰撞诱导碎裂(CID、HCD)或电子转移诱导碎裂(ETD)后产生的碎片离子图，这些二级谱图可用于对应肽段的序列鉴定[15]. 目前蛋白质组学中常用的数据检索方法为数据库依赖的检索方法，以 Mascot、Maxquant、Sequest 等搜索引擎为代表[16]，将二级谱图中的碎片信息与数据库中蛋白质肽段序列所能产生的理论碎片进行比对，通过一系列数据统计分析，得到检索结果.

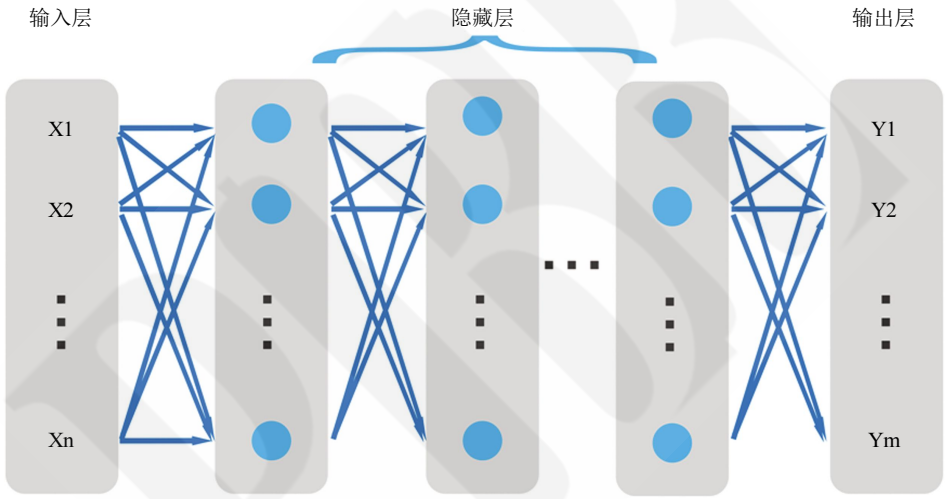

Fig. 1　The model of deep learning

图 1　深度学习示意图

尽管近年来以生物质谱为核心的蛋白质组学技术飞速发展，但在一些特定质谱数据的分析中仍面临巨大挑战，如面对蛋白质数据库中未收录蛋白质以及抗体药物的序列测定，现有以数据库检索方式为主的数据解析方法束手无策. 另外，传统的数据分析方法在质谱中肽段碎裂规律研究及肽段理化性质预测方面也无能为力. 当前，机器学习是数据挖掘中重要的研究方法，在当前各学科发挥着重要作用. 机器学习分为监督式学习及无监督式学习，常见的方法有线性回归、逻辑回归、决策树、支持向量机、朴素贝叶斯、K 邻近算法、随机森林、人工神经网络等一些算法. 在质谱及蛋白质组学领域，机器学习发挥着重要作用，蛋白质组学常用的如 k-means 方法等的聚类方法，则是一种无监督式学习方法，可将样本依据蛋白质学数据特征分类成特定组群，或者通过聚类找到功能差异相似基因. 另外，在蛋白质结构预测、蛋白质翻译后修饰、蛋白质相互作用网络等领域，亦有大量的基于机器学习方法的出现，为生物学功能研究提供重要的参考价值. 近年来，得益于计算机硬件及深度学习算法的飞速发展，生物医药领域已经有大量的使用深度学习的实例出现[17-18]. 深度学习相比传统的机器学习方法，可包含众多隐藏层及非线性变换，能更有效地解读质谱及蛋白质组学数据中复杂的关系，具有比传统机器学习方法更有效的数据解读能力，从一开始出现就引起这一领域的高度关注，国内外众多课题组也开始将这一最新的数据分析方法应用到蛋白质组学中，既有关于质谱数据的分析，亦有对蛋白质功能、细胞定位的预测等，均得到了良好的效果.

本文简要综述了深度学习方法应用于生物质谱及蛋白质组学的最新进展，对这些进展做简要的总结，并进一步展望深度学习方法在未来生物质谱及蛋白质组学领域的应用前景，以期为从事相关领域的研究人员提供深度学习在本学科应用的参考.

1 深度学习在生物质谱领域的应用

质谱仪器在近年来发展迅速，扫描速度、质量精度等性能日益提高．然而，与质谱仪器高速发展不相匹配的是数据处理方式发展有限，成为蛋白质组学研究新的技术瓶颈．随着深度学习技术的发展，这一最前沿的机器学习方法势必将在质谱数据解析领域发挥越来越重要的作用．当前在从头测序(*de novo*)、肽段谱图预测以及质谱成像等领域已经有深度学习方法应用的出现，下文将对这一领域作初步的介绍.

1.1 深度学习在肽段从头测序中的应用

在蛋白质组学研究中，*de novo* 测序方法是发现未知序列新蛋白质的重要方法．由于无数据库序列作参考，*de novo* 测序只能利用二级谱图信息以及一级母离子精确质量数进行肽段氨基酸序列的确认．在近几十年的研究中，诞生了许多工具软件，如 PEAKS[19]、pNovo[20]、Novor[21]及 MSNovo[22]等，极大地促进了这一领域的发展．但需要注意的是，*de novo* 在许多方面仍存在技术难题，如混合谱的存在会干扰目标肽段的鉴定、重要肽段碎片的缺失以及大量肽段中性丢失峰的存在使得精确的肽段序列鉴定存在困难等．另外，由于对解析出的序列缺乏有效的验证手段，*de novo* 方法在结果的准确性确认方面也面临许多问题．近年来，随着越来越多抗体药物的上市，对抗体进行精确的序列测定在药物研发、试剂开发中扮演着日益重要的角色[23]．抗体互补决定区(complementarity-determining regions，CDR)的氨基酸一级序列是抗体分子发挥生物学功能的核心区域，对这一区域序列的精确分析具有重要的药物开发价值．因此 *de novo* 方法在抗体药物测序领域发挥着重要作用．鉴于此，当前生物质谱研究领域迫切需要开发高效率、高精度的 *de novo* 数据分析方法.

结合最新发展的深度学习方法，2017 年加拿大滑铁卢大学的李明教授团队[24]开发了基于深度学习方法的肽段从头测序框架 DeepNovo．DeepNovo 采用两组 CNN 与一组长短期记忆人工神经网络(long short-term memory，LSTM)进行肽段的 *de novo* 测序，其中两组 CNN 分别称为 ion-CNN 与 spectrum-CNN，均以 Tensorflow 框架为核心搭建而成．在这项研究中，任意一张质谱图可以被理解为柱状图数据，即质荷比与该质荷比下的强度信息．由于一张谱图中存在大量的质荷比分布，因此数据呈现出高度复杂度．在结果输出即肽段预测方面，本方法将所有的肽段序列理解为 20^L (L 为肽段长度，20 为理论上氨基酸种类)的排列形式．综上，这一模型选取质谱二级谱图作为输入层，而按氨基酸顺序排列的肽段作为输出层，框架结构如图 2 所示．其中，在 spectrum-CNN 部分中，DeepNovo 学习谱图的一般特征，结果如图中 prefix lookup 部分所示，并将其传递到 LSTM 神经网络中；而在 ion-CNN 部分中，通过学习当前谱图特征，给出预测的氨基酸候选可能．如此，可以得到两组信息，即从 ion-CNN 学习得到的信息以及经过 spectrum-CNN 和 LSTM 后得到的信息，将这两组信息进行整合，可以得到未知肽段的序列预测．经过 DeepNovo 进行分析后，准确度大幅领先目前已有的 *de novo* 方法，在氨基酸和肽段精确度上分别提高 7.7%～22.9%和 38.1%～64.0%.进一步将 DeepNovo 应用于抗体轻链与重链的全序列重构，在无蛋白质数据库辅助的前提下，实现了 97.5%～100%的序列覆盖度以及 97.2%～99.5%的准确度．以上结果展示了深度学习方法应用于谱图解析上巨大的成功，也为将来深度学习应用于其他类型质谱数据的解析提供了有益参考.

1.2 深度学习在肽段二级谱图碎片离子预测中的应用

实验中得到的肽段二级质谱碎片谱图与理论碎片谱图的比对，是当前 Mascot、Sequest、pFind、PEAKS 等常见的蛋白质组学数据库检索软件进行质谱数据解析的基础，因此肽段碎裂谱图的理论预测对于数据库检索策略而言至关重要．在一张肽段的二级谱图中信息包含两个维度，即质荷比 (m/z)与信号强度．当前的主要数据库检索工具往往只利用了质荷比信息，而忽略了信号强度信息．在这一领域，目前已经有了一些利用机器学习方法进行肽段谱图预测的尝试，如 MassAnalyzer [25]、MS-Simulator[26]以及 PeptideART[27]等，但总体效果并不理想．鉴于肽段谱图数据中各碎片信息之间存在前后关联形式，因此 RNN 以及由此衍生出的长短期记忆网络(long short-term memory，LSTM)在谱图预测方面具有可预见的优势．这是由于在 RNN

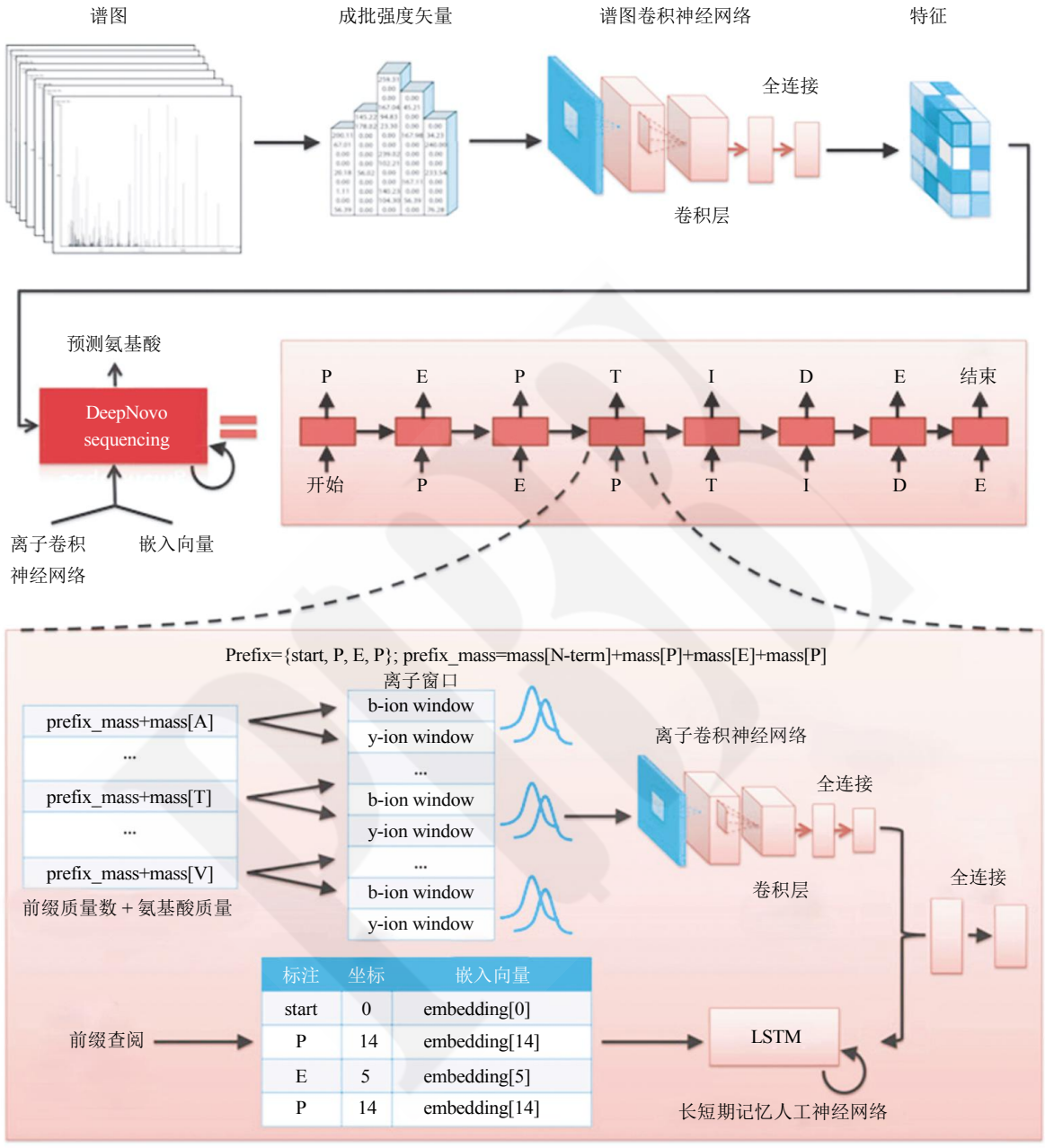

Fig. 2　Schematic diagram of de novo peptide sequencing in DeepNovo

图 2　DeepNovo 用于肽段 de novo 从头测序示意图[24]

中，上一隐藏层的信息可有效传递至下一隐藏层，从而更好地利用了质谱数据中质荷比的前后相关性．基于此，Zhou 等[28]开发了 pDeep 算法，利用双向长短时记忆循环神经网络(bidirectional long short-term memory，BiLSTM)进行了肽段二级谱图的预测．模型是采用 Keras 平台下 Tensorflow 运行环境搭建而成，通过采用图形处理器(GPU)进行运算，通过前期模型考察，选取两层隐藏层构建 BiLSTM 模型．最终，Zhou 等收集了近 400 万张高精度的二级质谱图，包含了 HCD、ETD 及 EThcD 等不同的碎裂模式．通过进行模型训练，pDeep 可以实现对 3 种碎裂方式进行谱图预测，并且结果中平均皮尔森相关系数大于 0.9，证明了方法的稳定可靠．另外，这一策略也发现，神经网络

中间层得到的信息可以揭示肽段的一些理化性质，可以从一个侧面理解肽段理化性质对于肽段二级碎裂具有影响．另外，此策略对于质量数相近的氨基酸组成，例如 GG 与 N、AG 与 Q、甚至 I 与 L 等这些在普通数据检索策略中无法区分的情况都可以实现有效判别，其中对于 I 与 L 可在 HCD 谱图中实现 0.67 的分辨准确率，而在 EThcD 谱图中准确率更是高达 0.76．pDeep 的成功构建说明使用机器学习方法对理论肽段序列进行二级谱图预测具有广阔的应用前景，未来势必会对肽段数据检索算法的提升提供重要指导．

1.3 深度学习应用于肽段性质预测

肽段的定性与定量是基于“bottom-up”方法的蛋白质组学研究中的核心内容，而肽段的准确定性定量与其质谱特征的提取及理化性质的预测密切相关．在液相色谱 - 质谱串联分析中，肽段的保留时间预测在非标定量研究以及多反应监测方法开发中具有重要的应用价值．肽段的色谱保留时间主要由肽段的理化性质决定，之前已有一些对其进行预测的算法出现，如 SSRCalc、ELUDE、GPTime 等，但存在较大的预测误差，实际应用存在不少困难．鉴于此，Ma 等[29]发展了 DeepRT 方法，将深度学习方法用于液相色谱中肽段保留时间的预测．DeepRT 利用了 CNN 与 RNN 两种方法提取肽段特征，其中 CNN 过程使用了 4 层卷积网络，用来处理肽段序列．RNN 则处理每一个肽段序列，并把每个氨基酸视作长度为 20 的向量．经过两种神经网络的特征提取后，利用主成分分析进行降维，继而采用常见的 3 种机器学习方法(支持向量机、随机森林及梯度提升)进行保留时间的预测，由于神经网络方法并无特定的特征提取方式，更少的依赖人为经验，而是机器根据数据类型自动学习处理，因此往往能够实现更为有效的特征提取，从而实现保留时间的精确预测．通过将已发表的数据集按比例分成训练集、验证集及测试集(8 ∶ 1 ∶ 1)，DeepRT 模型得到有效训练，实现了理论预测值与真实值相关性接近 0.99，在与其他保留时间预测软件 ELUDE 和 GPTime 的对比中，具有更精确的保留时间预测．

在液相色谱串联质谱分析中，肽段的色谱及质谱性质提取是最基本的处理步骤，其包含三个维度的数据信息，即肽段的质荷比、保留时间以及肽段信号强度，从而形成一个 3D 图形．目前的肽段色谱及质谱性质提取是根据预先设置好的参数，利用先验知识进行特定类型的特征提取，在特征识别的灵敏度上仍存在许多问题．鉴于此，Zohora 等[30]开发了一套基于深度学习的肽段色谱及质谱性质提取策略 DeepIso，采用 CNN 扫描液相谱图，监测肽段特征，并将其应用于肽段的色谱及质谱性质提取及定量研究．DeepIso 框架分为三部分，即 CNN 训练、利用训练出的 CNN 进行数据的肽段色谱及质谱性质提取以及提取结果的验证．利用该策略对采用 Asp-N、Chymotrypsin 和 trypsin 酶切处理的抗体轻重链共 6 个样品的肽段进行色谱及质谱性质提取，实现了数据库检索中 93.21%鉴定结果可与特征提取峰相匹配，并实现了高达 99.44%的特征提取的准确度．这一应用揭示了深度学习方法在基于液相色谱 - 质谱串联分析的蛋白质组定性及定量流程中肽段色谱及质谱性质提取方面的应用潜力．

1.4 深度学习在数据非依赖分析中的应用

目前数据非依赖采集(data independent acquisition, DIA)是定量蛋白质组学研究中发展迅速的技术．其可以将固定宽度质量窗口下所有母离子进行二级碎裂，采集得到所有母离子的二级碎片质谱图，从而实现高覆盖的定量方法．然而，正是由于较宽的一级质量窗口，导致二级谱图中存在来自大量不同肽段的碎片相互干扰，使得 DIA 方法的直接定性存在很大困难．另外，当前的 DIA 数据解析方法，如 DIA-umpire[31]、PIQED[32]等，仅能鉴定数据库中的蛋白质，无法发现未知序列肽段．因此，尽管 DIA 数据中包含丰富的蛋白质组信息，但由于其数据结构的高度复杂性，现有策略无法进行充分、准确的解析，限制了 DIA 技术的进一步推广．深度学习方法对于高复杂度的数据具有比传统方法更为有效的优势，近期也有关于深度学习方法应用于 DIA 研究的报道．据加拿大滑铁卢大学李明教授最新的综述文章透露，其研究小组正在开发基于深度学习的数据解析工具 DeepDIA，将可以对 DIA 扫描中的谱图进行 *de novo* 测序[33]．据介绍，前期的实验结果显示 DeepDIA 在 DIA 数据的 *de novo* 分析中展现出高准确度和可靠性．更详细的介绍需要在其文章发表后可见．另外，由 Vadim 等[34]开发的 DIA-NN (data independent acquisition by neural networks)工具亦利用人工神经网络提高 DIA 数据中母离子的鉴定能力，并进一步提高蛋白质定量的准确度．随着深度学习技术的发展，更多先进的分析方法与 DIA 相结合将使得 DIA 策略愈加强大和完善．

1.5 深度学习在质谱成像方面的应用

质谱成像技术(mass spectrometry imaging，MSI)是主要以 MALDI 为基础的成像方法，该方法通过质谱直接扫描生物样品，离子化并检测其表面分子进而成像，以实现在同一张组织切片或组织芯片上同时分析数百种化合物分子的空间分布特征．目前已进入了临床研究领域，其前所未有的细节呈现能力为分析组织样本提供了独特的优势[35]．在前期，已经有一些利用机器学习进行质谱成像数据分析的研究出现[36]，如 Boskamp 等[37]利用线性判别分析方法(linear discriminant analysis)区分不同类型肿瘤的特征．得益于质谱技术的发展，当前质谱成像数据的规模与信息量愈发庞大，因此急需发展相应高效数据解析方法．随着深度学习技术的发展，许多研究小组尝试将深度学习方法应用到质谱成像的数据解析．由于目视检查肿瘤组织并不能有效区分肿瘤及其亚型与健康组织的差异并揭示复杂代谢变化，Inglese 等[38]利用基于非监督神经网络方法的非线性降维方法(parametric t-SNE)进行数据降维，以更好的发现不同肿瘤样本中的生物异质性．Behrmann 等[39]尝试将 CNN 运用到质谱成像领域，对肺癌中腺癌与鳞癌两组样本进行分类，并通过交叉验证证实了深度学习方法的可靠性．以上尝试也证明了深度学习方法在未来大规模临床样本质谱成像的数据分析中会发挥越来越重要的作用．

2 深度学习应用于蛋白质组学研究

随着近年来质谱技术的飞速发展，分析灵敏度和通量显著提高，进而产生出越来越大的蛋白质组数据规模，进入了大数据时代．然而，面对与日俱增的数据规模，常规的数据分析方法愈发束手无策，无法有效解读．因此，机器学习方法开始逐渐进入蛋白质组学数据解析领域[40]．蛋白质组学中常用的质谱分析模式为自下而上(bottom-up)，即质谱采集分析被特定蛋白酶如胰蛋白酶(trypsin)酶解的蛋白质肽段片段，进而推导出其对应的蛋白质．蛋白质组学数据往往呈现多维特征，包含成百上千样本来源以及数以万计蛋白质种类的定量信息．常见的机器学习方法往往无法充分挖掘其包含的丰富信息，更需要深度学习技术提供更强大的数据挖掘能力，来满足日益增长的蛋白质组大数据的解析需求.

2.1 深度学习用于基于蛋白质组学策略的生物标志物筛选研究

深度学习可以用来对复杂多样的蛋白质组学数据进行深入的信息挖掘，An 等[41]从已发表的 259 例血浆样本的质谱数据中，利用深度信念神经网络(deep belief network，DBN)进行阿尔茨海默病诊断标志物蛋白质的筛选．经过深度学习的训练，发现了包含 20 个蛋白质的标志物群，其诊断精度超过 90%，并发现了 ACRP30 蛋白与阿尔茨海默病具有很强的相关性．在此项研究中，研究者还发现，采用深度学习方法，其模型准确度明显高于传统的机器学习方法如支持向量机、k 最邻近算法及普通的反向传播神经网络．

2.2 深度学习用于核酸结合蛋白预测

核酸(DNA 或 RNA)结合蛋白在细胞生物进程中发挥着重要作用[42-44]，当前已有众多生物化学及标记方法用于 DNA 及 RNA 结合蛋白的鉴定研究[45-49]，但实验的精确度和鉴定规模仍存在诸多局限，急需其他方法予以补充．目前已经有计算方法利用蛋白质的部分性质，如结构域序列等进行 DNA 或 RNA 结合蛋白的预测，但精确度仍然较为有限．为了研究 DNA 或 RNA 结合蛋白的序列特征，Alipanahi 等[50]利用深度学习方法从实验数据中进行学习，发展出了基于 CNN 的 DeepBind 方法，可用于预测蛋白质序列的核酸结合特性．经过对此方法的评测发现，即使利用体外实验数据训练模型，而用体内实验数据进行验证，此方法所得结果的准确性仍远超当前其他预测方法．由 DeepBind 确定的蛋白质结合位点的特异性甚至可作为权重指标，指示位点变异将如何影响其与特定序列的结合，为核酸结合蛋白质的研究提供了有力的工具支持．另外，Zheng 等[51]利用深度学习方法，基于 Tensorflow 框架进行 RNA 结合蛋白的预测，发展出 Deep-RBPPred 算法．此方法利用 CNN 训练 RNA 结合蛋白预测工具，取代其先前发表的 RBPPred 算法中的支持向量机算法[52]．在 Deep-RBPPred 中，作者利用蛋白质的一些理化性质如疏水性、极性、归一化的范德瓦尔斯体积、水溶性及侧链电荷性等去训练 11 层神经网络的权重，实现了 RNA 结合蛋白的高效预测．在未来，随着深度学习方法的改进以及大量生物学验证数据的支持，利用深度学习方法对未知结合蛋白及核酸结合序列的预测将会有非常光明的应用前景，为这一领域的发展提供强有力的生物信息学支持.

2.3 深度学习用于蛋白质定位预测

真核生物蛋白质的亚细胞定位与其功能调控直接相关，由于现有实验技术和数据存在一定的局限

性，通过生物信息学手段对蛋白质的定位进行预测是当前该领域的重要研究内容．目前已有许多机器学习方法被应用在蛋白质定位预测中，但这些方法往往依赖已知数据库信息，对于存在位点突变的蛋白质及无同源信息的蛋白质无法提供有效的定位信息．鉴于此，Armenteros 等[53]利用 RNN 进行蛋白质定位的预测．在他们的模型中，依据最新的 uniprot 蛋白质数据库中的信息，利用 RNN 首先选取全蛋白质序列，再进一步选取蛋白质序列中对蛋白质定位起重要作用的序列，使用经过实验验证的蛋白质定位精确信息，进行模型训练，实现了出色的蛋白质定位预测效果(膜蛋白及可溶性蛋白质的预测准确度可达 92%)，远超过当前其他蛋白质定位预测软件．

3 总结与展望

从上文中可以看到，近年来大量以深度学习为基础的分析方法在生物质谱及蛋白质组学领域得到广泛应用．在质谱解析领域，面临着自动化高精度解析的难点，如 *de novo* 研究向来是谱图解析领域研究难点，已存在的解析方法在精度及易用性上都无法满足要求，而以 DeepNovo 为代表的研究出现，预示着深度学习在谱图解析领域将给这一领域带来重大变革．当前，质谱技术发展飞快，新型技术如 EThcD 以及 DIA 的快速发展需要强大的数据分析方法的跟进，传统的数据解析手段往往依赖数据检索及手工辅助相结合，技术门槛较高并无法有效地对富含多维度信息的质谱数据进行有效解析，而深度学习方法可避免这些问题，其利用已知的数据标注信息，通过多层神经网络，提取信息特征，并能对未知数据进行有效的注释与解析．当前在图像处理方面，深度学习取得了一系列重大进步，而质谱成像领域则与之关系最为相近，有望在接下来的几年中借助深度学习在图像识别领域的发展而受益，进一步在临床检测中发挥越来越重要的作用．在蛋白质组学研究领域，深度学习已广泛应用到生物标志物发现、核酸结合蛋白预测及蛋白质定位研究中，丰富了这一系列领域机器学习的深度，得到了比传统机器学习更好的预测能力，也为这些领域的数据挖掘带来崭新的思路．

与传统的机器学习相比，深度学习的优点在于可通过多层神经网络，经过如梯度下降方法的参数优化，可自动从复杂的数据中学习，能适应不同种类、多种类型的数据，并得到出色的预测结果．不过需要注意的是，深度学习也存在一定的问题，比如建立起来的模型由于经过多个隐藏层的训练，内部结果不易解读，往往无法实现结果的有效呈现．另外，由于深度学习往往含多个隐藏层，训练模型运算量巨大，往往需要高性能图形处理器的辅助，对硬件规格要求较高．尽管面临一部分问题，但令人欣慰的是深度学习方法的飞速发展，越来越多更有效的新型算法层出不穷，极大丰富了深度学习的应用范围．而随着以 Tensorflow 等多方开源平台的发展，以及大量互联网云服务的布局，深度学习的使用门槛越来越低，有利于将这一先进的数据处理方法应用到质谱相关的各学科领域．

随着质谱技术的高速发展，当前蛋白质学研究中蛋白质序列测定发展迅速，但是由于翻译后修饰如糖基化、磷酸化等，其化学计量值相对较低，质谱响应弱，二级质谱图往往碎裂较差，往往无法对翻译后修饰进行有效的鉴定．先前大量的机器学习方法已应用到蛋白质翻译后修饰的预测中，也取得了一定的效果，而深度学习方法可更有效地利用先前大规模质谱数据，将进一步对蛋白质翻译后修饰的研究推波助澜，实现更精确的蛋白质翻译后修饰研究．同样，在蛋白质结构解析等领域，伴随着以冷冻电镜为代表的技术对这一领域发展具有重大推进作用，将产生出越来越大量的数据信息，以此为资源，可进一步利用深度学习研究其他蛋白质的结构信息．在蛋白质组学研究中，依据蛋白质组表达以及基因组、转录组联合分析用于肿瘤及其他疾病分子分型也是研究热点．传统的机器学习方法往往无法对数据进行有效的解读，而深度学习得益于其自身算法的普适性，并能对多种数据格式进行有效的处理，在多组学分子分型研究中将发挥越来越重要的作用．另外，在蛋白质间相互作用、核酸结合蛋白的研究以及生物标志物发现等领域，深度学习可利用先前大量的实验数据进行训练，以求更好地利用已知数据去预测未知的信息．但同时也需要注意到，深度学习往往需要巨大的样本量，并需要将数据进行精确标记以帮助深度学习更好地工作，这方面涉及到谱图解析及标注、疾病样本评估、蛋白质组学表达谱数据产出等，都需要多学科专业人才的辅助．因此，专业的科学工作者与深度学习算法团队的合作将是高效的研究模式．

近年来，随着生物质谱技术和相关定性定量方法的发展，为蛋白质组学研究注入了强大的活力，并逐渐产生出愈发庞大的数据规模，使得蛋白质组

学研究进入了大数据时代．因此，如何对质谱产出的组学数据进行有效的数据挖掘并解读将会是未来这一领域的研究热点．传统的机器学习方法往往需要依赖先验知识，根据特定方式去提取数据特征，从以上应用实例可以看出，深度学习方法的发展，不仅越来越多的在生物质谱及组学数据处理中发挥作用，作为更高级的机器学习方法，深度学习所展现出的优异的数据挖掘及分析能力，必将在未来生命科学相关领域的研究中发挥更大的作用．

参 考 文 献

[1] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, **521**(7553): 436−444

[2] 张军阳, 王慧丽, 郭 阳, 等. 深度学习相关研究综述. 计算机应用研究, 2018, **35**(7): 1−12
Zhang J Y, Wang H L, Guo Y, *et al*. Application Research of Computers, 2018, **35**(7): 1−12

[3] Cun Y L, Boser B, Denker J S, *et al*. Handwritten digit recognition with a back-propagation network [M]//DAVID S T. Advances in neural information processing systems 2. Morgan Kaufmann Publishers Inc. 1990: 396−404

[4] Silver D, Huang A, Maddison C J, *et al*. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, **529**(7587): 484−489

[5] Silver D, Schrittwieser J, Simonyan K, *et al*. Mastering the game of Go without human knowledge. Nature, 2017, **550**(7676): 354−359

[6] Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature, 2003, **422**(6928): 198−207

[7] Aebersold R, Mann M. Mass-spectrometric exploration of proteome structure and function. Nature, 2016, **537**(7620): 347−355.

[8] Tanaka K, Waki H, Ido Y, *et al*. Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. Rapid Commun Mass Sp, 1988, **2**(8): 151−153

[9] Fenn J B, Mann M, Meng C K, *et al*. Electrospray ionization for mass spectrometry of large biomolecules. Science, 1989, **246**(4926): 64−71

[10] Ding C, Jiang J, Wei J, *et al*. A fast workflow for identification and quantification of proteomes. Mol Cell Proteomics, 2013, **12** (8): 2370−2380.

[11] Wilhelm M, Schlegl J, Hahne H, *et al*. Mass-spectrometry-based draft of the human proteome. Nature, 2014, **509**(7502): 582−587

[12] Kim M S, Pinto S M, Getnet D, *et al*. A draft map of the human proteome. Nature, 2014, **509**(7502): 575−581

[13] Ting L, Rad R, Gygi S P, *et al*. MS3 eliminates ratio distortion in isobaric labeling-based multiplexed quantitative proteomics. Nat Methods, 2011, **8**(11): 937−940

[14] Macek B, Waanders L F, Olsen J V, *et al*. Top-down protein sequencing and MS3 on a hybrid linear quadrupole ion trap-orbitrap mass spectrometer. Mol Cell Proteomics, 2006, **5**(5): 949−958

[15] Shen Y, Tolic N, Xie F, *et al*. Effectiveness of CID, HCD, and ETD with FT MS/MS for degradomic-peptidomic analysis: comparison of peptide identification methods. J Proteome Res, 2011, **10** (9): 3929−3943

[16] Yuan Z F, Lin S, Molden R C, *et al*. Evaluation of proteomic search engines for the analysis of histone modifications. J Proteome Res, 2014, **13**(10): 4470−4478

[17] Cao C, Liu F, Tan H, *et al*. Deep learning and its applications in biomedicine. Genomics, Proteomics & Bioinformatics, 2018, **16**(1): 17−32

[18] 李渊, 骆志刚, 管乃洋, 等. 生物医学数据分析中的深度学习方法应用. 生物化学与生物物理进展, 2016, **43**(05): 472−483
Li Y, Luo Z G, Guan N Y, *et al*. Prog Biochem Biophys, 2016, **43**(05): 472−483

[19] Ma B, Zhang K, Hendrie C, *et al*. PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom, 2003, **17**(20): 2337−2342

[20] Chi H, Sun R X, Yang B, *et al*. pNovo: *de novo* peptide sequencing and identification using HCD spectra. J Proteome Res, 2010, **9**(5): 2713−2724

[21] Ma B. Novor: real-time peptide de novo sequencing software. J Am Soc Mass Spectrom, 2015, **26**(11): 1885−1894

[22] Mo L, Dutta D, Wan Y, *et al*. MSNovo: a dynamic programming algorithm for *de novo* peptide sequencing *via* tandem mass spectrometry. Anal Chem, 2007, **79**(13): 4870−4878

[23] Tran N H, Rahman M Z, He L, *et al*. Complete *de novo* assembly of monoclonal antibody sequences. Sci Rep-Uk, 2016, **6**: 31730

[24] Tran N H, Zhang X, Xin L, *et al*. *De novo* peptide sequencing by deep learning. Proc Natl Acad Sci USA, 2017, **114**(31): 8247−8252

[25] Zhang Z. Prediction of low-energy collision-induced dissociation spectra of peptides. Anal Chem, 2004, **76**(14): 3908−3922

[26] Sun S, Yang F, Yang Q, *et al*. MS-Simulator: predicting y-ion intensities for peptides with two charges based on the intensity ratio of neighboring ions. J Proteome Res, 2012, **11**(9): 4509−4516

[27] Li S, Arnold R J, Tang H, *et al*. On the accuracy and limits of peptide fragmentation spectrum prediction. Anal Chem, 2011, **83**(3): 790−796

[28] Zhou X-X, Zeng W-F, Chi H, *et al*. pDeep: predicting MS/MS spectra of peptides with deep learning. Anal Chem, 2017, **89**(23): 12690−12697

[29] Ma C, Zhu Z, Ye J, *et al*. DeepRT: deep learning for peptide retention time prediction in proteomics. Arxiv, 2017, 1705.05368

[30] Tuz Zohora F, Hieu Tran N, Zhang X, *et al*. DeepIso: a deep learning model for peptide feature detection. Arxiv, 2017, 1801.01539

[31] Tsou C C, Avtonomov D, Larsen B, *et al*. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. Nat Methods, 2015, **12**(3): 258−264, 257 p following 264

[32] Meyer J G, Mukkamalla S, Steen H, *et al*. PIQED: automated identification and quantification of protein modifications from DIA-MS data. Nat Methods, 2017, **14**: 646−647

[33] Tran N H, Zhang X, Li M. Deep omics. Proteomics, 2018, **18**(2): 1700319

[34] Demichev V, Messner C B, Lilley K S, *et al*. DIA-NN: deep neural networks substantially improve the identification performance of Data-independent acquisition (DIA) in proteomics. BioRxiv, 2018, doi: https://doi.org/10.1101/282699

[35] Aichler M, Walch A. MALDI Imaging mass spectrometry: current frontiers and perspectives in pathology research and practice. Lab Invest, 2015, **95**(4): 422-431

[36] Galli M, Zoppis I, Smith A, *et al*. Machine learning approaches in MALDI-MSI: clinical applications. Expert Rev Proteomics, 2016, **13**(7): 685-696

[37] Boskamp T, Lachmund D, Oetjen J, *et al*. A new classification method for MALDI imaging mass spectrometry data acquired on formalin-fixed paraffin-embedded tissue samples. Biochim Biophys Acta, 2017, **1865**(7): 916-926

[38] Inglese P, Mckenzie J S, Mroz A, *et al*. Deep learning and 3D-DESI imaging reveal the hidden metabolic heterogeneity of cancer. Chem Sci, 2017, **8**(5): 3500-3511

[39] Behrmann J, Etmann C, Boskamp T, *et al*. Deep learning for tumor classification in imaging mass spectrometry. Bioinformatics, 2018, **34**(7): 1215-1223

[40] Swan A L, Mobasheri A, Allaway D, *et al*. Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. Omics, 2013, **17**(12): 595-610

[41] An N, Ding H, Yang J, *et al*. Deep learning application in identifying proteomic risk markers for Alzheimer's. Disease. Alzheimer's & Dementia, 2017, **13**(7): P1133

[42] Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. Nat Rev Genet, 2014, **15**(12): 829-845

[43] Vaquerizas J M, Kummerfeld S K, Teichmann S A, *et al*. A census of human transcription factors: function, expression and evolution. Nat Rev Genet, 2009, **10**(4): 252-263

[44] Hudson W H, Ortlund E A. The structure, function and evolution of proteins that bind DNA and RNA. Nat Rev Mol Cell Biol, 2014, **15**(11): 749-760

[45] Castello A, Fischer B, Eichelbaum K, *et al*. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell, 2012, **149**(6): 1393-1406

[46] Huang R, Han M, Meng L, *et al*. Transcriptome-wide discovery of coding and noncoding RNA-binding proteins. Proc Natl Acad Sci USA, 2018, **115**(17): E3879-E3887

[47] Bao X, Guo X, Yin M, *et al*. Capturing the interactome of newly transcribed RNA. Nat Methods, 2018, **15**(3): 213-220

[48] Hu S, Xie Z, Onishi A, *et al*. Profiling the human protein-DNA interactome reveals ERK2 as a transcriptional repressor of interferon signaling. Cell, 2009, **139**(3): 610-622

[49] Qi L S, Larson M H, Gilbert L A, *et al*. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell, 2013, **152**(5): 1173-1183

[50] Alipanahi B, Delong A, Weirauch M T, *et al*. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol, 2015, **33**(8): 831-838

[51] Zheng J, Zhang X, Zhao X, *et al*. Deep-RBPPred: predicting RNA binding proteins in the proteome scale based on deep learning. BioRxiv, 2017, https://doi.org/10.1101/210153

[52] Zhang X, Liu S. RBPPred: predicting RNA-binding proteins from sequence using SVM. Bioinformatics, 2017, **33**(6): 854-862

[53] Almagro Armenteros J J, Sonderby C K, Sonderby S K, *et al*. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics, 2017, **33**(21): 3387-3395

Application of Deep Learning in Biological Mass Spectrometry and Proteomics*

ZHAO Xin-Yuan[1], QIN Wei-Jie[2]**, QIAN Xiao-Hong[2]**

([1] *College of Life Science and Bioengineering, Beijing University of Technology, Beijing* 100124, China;
[2] *National Center for Protein Sciences Beijing, State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Lifeomics, Beijing* 102206, China)

Abstract Deep learning is the most popular research area in the field of machine learning in recent years, especially in image and speech recognition, natural language processing, and automatic driving. Biological mass spectrometry is an important research tool in the field of life sciences and plays a key role in proteomics, metabolomics, and biopharmaceuticals. In recent years, based on the development of deep learning methods, the big data analysis in proteomics centered on biological mass spectrometry will usher into a new era. This article reviews the latest applications of deep learning methods in the analysis of biological mass spectrometry data and proteomics research.

Key words deep learning, biological mass spectrometry, proteomics, big data
DOI: 10.16476/j.pibb.2018.0165

* This work was supported by grants from National Key Program for Basic Research of China (2016YFA0501403, 2017YFC0906703), The National Natural Science Foundation of China (21675172) and National Key Laboratory of Proteomics Grant (SKLP-K201706).
**Corresponding author. Tel: 86-10-61777107
QIAN Xiao-hong. E-mail: qianxh1@163.com
QIN Wei-jie. E-mail: aunp_dna@126.com
Received: June 13, 2018 Accepted: September 10, 2018