摘　要

随着生物医药技术以及计算机科学技术的发展，质谱分析技术在蛋白质组学及糖组学等多领域得到了应用，包括蛋白质鉴定、寡糖结构鉴定、生物标记物发现与疾病诊断建模等。

质谱检测技术主要分为两大类DDA技术（Data Dependent Acquisition数据依赖采集模式），DIA（Data Independent Acquisition，数据非依赖采集），具体分析方法可细分为库搜索方法和De Novo方法。理论谱预测作为基于质谱分析鉴定方法设计中的一个重要的环节，对于库搜索方法及De Novo 方法的鉴定准确性都有很大的影响。然而大部分蛋白质鉴定软件在理论谱预测环节的设计太粗糙，从而影响了最终鉴定结果的准确性。理论谱预测是指设计算法模型模拟蛋白质序列在质谱仪中碎裂的方式，从而构建一个理论的质谱用于跟实际质谱数据进行比较。

与蛋白质序列相比，寡糖树形分枝结构更为复杂，往往仅用二级质谱不足以实现寡糖分枝结构的准确鉴定。多级质谱分析可以通过对寡糖结构的连续多次碎裂，得到更多的结构信息从而实现对寡糖分枝结构的准确鉴定。但是，在鉴定过程中如何选择打谱路径产生多级质谱数据是一大难题。常规情况下，多级质谱的产生方式都是基于实验操作员的专业经验或者在当前质谱中选择峰强最强的离子峰产生下一级质谱的方法来得到多级质谱。但是，人工选峰的方法耗时且往往鉴定不准确，而且会增加样品用量。

针对上述问题，通过对生物大分子在质谱仪中碎裂机理以及对计算机算法和统计模型的深入研究，取得了如下成果：

（1）提出并实现了MS-Simulator理论谱预测模型

本研究提出了一种可用来预测给定肽段对应的y离子丰度的模型MS-Simulator。MS-Simulator模型基于质子迁移理论通过对肽段相邻位置糖苷键断裂形成y离子的丰度比的预测，然后根据丰度比信息计算离子峰的相对强度结合离子质量信息构建理论谱。跟已有的理论谱预测模型MassAnalyzer相比，MS-Simulator模型参数训练过程更加简单，而且预测的y离子理论谱准确率比MassAnalyzer高。通过对SEQUEST和X！Tandem等质谱鉴定软件的鉴定结果重打分，显著提高了质谱鉴定的准确性。

（2）提出并实现了TagNovo词条库模型理论谱预测模型

本研究提出了一种词条库模型TagNovo，通过收集已有鉴定结果的质谱数据，对每一肽段分成多个5-mer长度的肽片段，并从质谱中找其对应的局部质谱数据，每一5-mer片段及对应谱局部信息构成一个词条，对大量的词条进行聚类形成质谱“词条库”。TagNovo词条库模型可以用来实现理论谱的全谱预测，从而可以用来对小物种生物通过对蛋白序列的理论谱预测实现谱库扩充，解决了谱库搜索方法的发展瓶颈。

（3）提出并实现了多级质谱寡糖分枝结构鉴定模型GIPS

本研究提出了利用多级质谱实现寡糖分枝结构自动鉴定模型GIPS。 GIPS的创新点主要表现在两方面：1）提出一种新的选峰算法，通过计算选择含有潜在信息量最多的峰（SMI峰）来产生一级谱；2）提出了基于层次贝叶斯模型的寡糖候选结构打分算法，可来评估SMI峰的信息量以及对鉴定结果打分。实验结果表明，在对寡糖纯样品及从糖蛋白上分离的寡糖样品进行鉴定时， GIPS可以使用少量的打谱次数实现寡糖分枝结构的准确鉴定，而且对于有相似结构的寡糖同分异构体可以实现准确区分。GIPS寡糖鉴定策略避免

了多级质谱的手动选峰打谱，显著提高了寡糖鉴定的灵敏度和吞吐量且大大减少了样品的消耗。

（4）提出了基于深度学习的DIA糖蛋白鉴定算法

本研究提出了基于理论谱比对的质谱数据质量评价方法，用于评价质谱的数据质量。提出了基于多模式适应的数据结构设计方法用于重构质谱数据检索模式。提出了基于CNN+BI-LSTM的De Novo鉴定算法用于实现DIA质谱鉴定及糖蛋白序列预测。

本研究提高了理论谱预测的精度，通过理论谱预测对鉴定结果重打分提高了已有蛋白质鉴定软件的准确性。提出了新的质谱词条库模型，解决了现有谱库搜索方法的局限性。提出了多级质谱寡糖鉴定策略，使得自动、快速、准确地对寡糖分枝结构鉴定成为可能。

关键词：质谱分析，蛋白质，多肽，寡糖结构，算法，深度学习