

摘要

近年来,基于串联质谱技术的蛋白质鉴定,逐渐成为蛋白质组学研究中的关键问题,而质谱数据的预处理是其中不可忽略的问题之一,因为质谱中的大量噪声使得鉴定时间的增加和精度的下降,而且同位素峰的存在也会使肽鉴定的准确性和可靠性受到影响。本文主要针对低精度的离子阱质谱数据,研究了串联质谱数据的预处理问题,包括质谱中的强度基线水平的确定和同位素质峰的认识问题。

第二章首先简单介绍了质谱技术的基础,包括质谱仪的原理、组成和数据处理等,随后详细介绍了基于串联质谱的蛋白质鉴定方法,分析了质谱数据预处理的必要性,并总结了现有预处理算法不足,比如不能根据每个质谱的不同特点来获得基线。

在第三章中,论文提出了一种自适应地确定谱峰强度基线的方法。通过大量观察发现,不同质谱中谱峰强度的分布范围是有差别的,但是在形式上低强度的噪音峰都近似服从高斯分布,而离子峰则近似服从伽马分布。据此给出了一个期望最大化即 EM 学习算法和一个快速近似算法自适应地估计每个质谱中噪音峰强度的分布参数,从而确定噪音基线水平。实验结果表明,与固定基线方法相比,自适应基线方法能够在相同数据缩减率下提高肽鉴定的准确度。

在高精度质谱仪产生的质谱数据中,同位素质峰非常明显,相对容易识别。而在低精度的离子阱质谱中,同位素质峰则非常不明显甚至缺失,因此尚缺少有效的识别方法。本研究发现,离子阱数据中的同位素质峰实际上仍有规律可循。论文第四章首先给出了第一同位素与单同位素质峰在质量和强度方面的多种约束关系,并据此构造出谱峰的特征向量表示,然后提出利用机器学习中的决策树方法对谱峰类别进行学习和预测。在多个数据集上的实验表明,通过将识别出的同位素质峰去除,在相同的肽鉴定假阳性率下,鉴定出的肽段数量明显增加。

论文最后将新算法集成到 pFind 系统的质谱数据预处理模块中,并进行了大量实验,实验结果表明,进行预处理后, pFind 在鉴定速度上提高 15% 的同时,在相同肽鉴定假阳性率下,鉴定出的肽段数量有所提高。