
摘 要

当前蛋白质鉴定的研究中，通过串联质谱鉴定多肽序列从而鉴定蛋白质是最广泛使用的技术。实验中，从色谱中分离出来的多肽经过 CID 过程被裂解成碎片离子，这些离子的质量/电荷比值(m/z)被质谱仪器检测到，形成串联质谱。采用数据库搜索方法或者 De novo 从头解序方法，可从这些碎片离子的 m/z 值中鉴定出多肽的序列来。

然而，不管采用什么计算方法进行多肽序列的鉴定，高分辨率串联质谱数据的特点决定了其在计算上的困难。质谱中大量的物理噪声和离子的同位素峰，增加了多肽序列鉴定过程的计算量，而且使得随机匹配的可能性增高，从而导致鉴定的结果可靠性降低。此外，质谱中数据中的质量测量误差直接影响多肽鉴定结果。因此，在进行多肽序列鉴定之前，对质谱数据进行预处理非常关键。

本文从理论、算法和应用三个层次来讨论对串联质谱数据的预处理技术，实现对质谱数据的多种预处理，包括过滤质谱中的物理噪声、过滤同位素峰、预测离子对应的分子式、识别质谱的测量误差等。通过这些预处理，最终降低序列鉴定过程的计算量、提高单个质谱的鉴定可靠性、以及提高能鉴定出多肽的质谱个数，从而提高蛋白质鉴定的可靠性。

本文首先提出一个关键的同位素模式概念，可以定量地刻画离子的一系列同位素在质量、丰度上的特征；并给出了计算离子的理论和实验同位素模式的公式，在此基础上可进行多种讨论，比如区分质谱中离子和噪声对应的谱峰、预测离子的分子式、估计质谱质量测量误差等。

基于同位素模式概念，本文提出了从质谱中挑选潜在的离子单同位素峰的算法 *PeakSelect*。本文从理论上讨论了噪声和离子谱峰的本质区别以及质谱中离子同位素峰重叠的分布情况，并讨论了噪声谱峰在强度上的分布。在此基础上，本文提出多个有效的特征来区别噪声、孤立的离子谱峰、重叠的离子谱峰，并建立谱峰分类的决策树，从质谱中挑选潜在的离子的单同位素峰。实验结果表明 *PeakSelect* 能准确地挑选质谱中的离子单同位素谱峰，不仅能够大大缩短鉴定软件在多肽序列鉴定上所需的计算时间，并且能大大增加可靠鉴定出的多肽的个数，提高了质谱的利用率，也提高了所鉴定的蛋白质序列的覆盖率，从而提高了鉴定结果的可靠性。此外，本算法性能较之现有的商用软件，比如 *ProteinLynx*TM *Global Server* 对质谱有效峰选取的预处理效果更好。

在估计质谱质量误差之前，本文先提出了预测离子分子式的算法 *FFP*。*FFP* 通过比较分子式对应的理论同位素模式与质谱中实际出现的实验同位素模式间的差异来预测离子的分子式，并将分子式预测问题转化为优化问题。结合优化建模和统计分析，*FFP* 对小质量段内的离子分子式预测的五选正确率达 95% 以上。在此基础上，本文提出了估计质谱质量测量误差的算法 *QMass*，并将 *QMass* 应用到 Q-TOF 数据上。

质谱的质量测量误差包含两个部分，一是随机误差，二是系统误差。随机误差服从正态分布，而根据仪器的测量原理，可以得到系统误差的理论分布函数，比如，TOF 仪器的测量误差与离子理论质量近似成线性关系。对一个具体的质谱分析其测量误差就是要得到这个系统误差的分布函数的具体参数。*QMass* 借助 *FFP* 的预测结果可得到一些准确的测量误差样本点。并且，*QMass* 还通过计算质谱中各种潜在的连续或同源离子谱峰间的质量差的理论值与实验值之间的差异而得到相应的测量误差样本点。在这些样本点数据上 *QMass* 通过估计误差分布函数的参数从而得到整个质谱的质量测量误差估计。实验结果表明，95% 以上的质谱误差估计值与实际的误差值相差不超过 40ppm。据我们所知，*Qmass* 是第一个不依赖于内标或外标参考，也不依赖数据库搜索结果，直接从质谱数据中分析其质量测量误差的方法。

关键词：生物信息学，蛋白质鉴定，串联质谱，同位素模式，质谱数据预处理

On Preprocessing of Tandem Mass Spectra for Protein Identification

Zhang Jingfen (Computer Application Technology)

Supervised By Gao Wen

It has been a well-known method to identify proteins by identifying peptide sequences (or called peptide sequencing) using the tandem spectra. During experiments, the peptides separated from liquid chromatographers are fragmented and ionized by collision-induced dissociation (CID) and the ions are measured by mass spectrometer in mass/charge ratios (m/z). Consequently, the peptides can be identified by these m/z values of ions in tandem spectrum with a sequence database searching or *De novo* sequencing or the combining of the two above methods.

However, the numerous noise and isotopic peaks in high resolution tandem spectra (such as Q-TOF spectra) lead to a heavy computational cost in peptide identification. Furthermore, they can cause either false negative or false positive peptide identifications since they may match with the theoretical ions of an irrelevant peptide sequence. In addition, the measurement errors of ion masses in spectra puzzle the identification too. Therefore, the data preprocessing should be introduced before peptide sequencing.

This thesis aims to discuss the theory, algorithms and the application in preprocessing, and propose methods to preprocess tandem spectra in order to increase the accuracy of peptide identification and decrease the computation complexity.

Firstly, a key concept of Isotope Pattern Vector (*IPV*) which digitally characterizes the isotope cluster of a fragment ion universally is proposed in the thesis. Thus, the noise peaks and real peaks in spectra can be distinguished by the quantitative *IPV* value, the formulae of fragment ions can be predicted and the mass measurement errors can be analyzed.

Based on the concept of *IPV*, a new algorithm, *PeakSelect*, is proposed to find the monoisotope of ions in spectra which are crucial in peptide sequencing. In *PeakSelect*, we analyze the fundamental difference between noise peaks and ion peaks, the distribution of noise in intensity, and the complex overlapping of isotope peaks in spectra. By applying machine learning method, some features are proposed to distinguish the different information in spectra and a decision tree is constructed to classify the peaks into different categories such as noise, single ion peaks and overlapping peaks. Therefore, all of the potential monoisotopic masses of ions can be calculated. Experiments show that *PeakSelect* decreases greatly the computational times and increases the reliability of peptide identifications. In particular, *PeakSelect* performs well on complex spectra with a large number of peaks and from large peptides, and supports more sequence identification than other well-known systems such as

*ProteinLynx*TM Global Server.

To know the mass measurement error, we need know the theoretical masses of fragment ions in spectra. Therefore, we present a new method, *FFP* (Fragment ion Formula Prediction), to predict elemental component formulas of fragment ions and then know their theoretical masses. In *FFP*, we convert the prediction of the best formulas to the minimization of the distance between theoretical and experimental isotope patterns (*tIPV* and *eIPV*). Coupled with some local search method and a new multi-constraint filtering method, *FFP* can give accurate predicition for ions with low mass.

After *PeakSelect* and *FFP*, we propose a method *QMass* to analyze the mass measurement error in spectra and apply *QMass* in Q-TOF spectra. The measurement error can be divided into random error and system error, in which, the random error roughly follows the normal distribution and the distribution of the system error can be deduced by the measurement theory of the spectrometry. For example, the system errors in a TOF spectrum are linear with the ions' masses approximately. To predict the mass error for each spectrum is to find the parameters in the distribution function of errors. The monoisotopic peaks of some ions can be recognized by *PeakSelect* and the theoretical masses of some ions can be predicted by *FFP*. Therefore, some measurment error points can be obtained. Then, after having known the distribution function of the random and system errors, *QMass* can estimate the parameters in the distributions and analyze the measurement error for each spectrum. Experiments show that for over 95% spectra, the differences between the predicted error and the practical error are within 40ppm. To our best knowledge, *QMass* is the first method to analyze mass error directly from the spectrum and without any internal or external lock masses and without reference of database search results.

Keywords: bioinformatics, protein identification, tandem spectra, isotope pattern, preprocessing of spectra