Eco-Environmental Knowledge Web

环 境 化 学
ENVIRONMENTAL CHEMISTRY

第 41 卷第 10 期 2022 年 10 月
Vol. 41, No. 10 October 2022

DOI:10.7524/j.issn.0254-6108.2021070101

申珊齐, 李爱民, 黄雅娟, 等. 化学信息学在液相色谱高分辨质谱联用的化学品非靶向筛查中的应用[J]. 环境化学, 2022, 41(10): 3135-3143.

SHEN Shanqi, LI Aimin, HUANG Yajuan, et al. Cheminformatics in untargeted screening of liquid chromatography coupled to mass spectrometry data[J]. Environmental Chemistry, 2022, 41 (10): 3135-3143.

化学信息学在液相色谱高分辨质谱联用的化学品非靶向筛查中的应用*

申珊齐[1] 李爱民[1,2]** 黄雅娟[1] 施 鹏[1] 潘 旸[1]
李文涛[1] 张怀成[1] 吴 几[1]

（1. 南京大学环境学院，污染控制与资源化研究国家重点实验室，南京，210023；
2. 泉州南京大学环保产业研究院，泉州，362008）

摘 要 化学工业的发展使得环境介质中未知化合物数目巨大，对其进行识别是认识其环境风险进而开发削减策略的关键. 液相色谱串联高分辨质谱是化合物识别的常用技术，该技术采集的数据一般较复杂，需要适当的数据解析手段方能呈现复杂环境样品中的化合物信息，化学信息学在高分辨质谱非靶向筛查中的发展为化合物结构解析提供了可能. 本文综述了化学信息学在非靶向筛查中的应用. 基于非靶向筛查流程中的峰提取、去冗余、特征峰筛选、注释与结构确定步骤，从涉及的算法、软件、化合物数据库、谱图数据库等进行了阐述. 在此基础上，对算法和软件工具的参数优化和数据处理一致性进行了阐述. 本综述为更好的进行高分辨质谱数据非靶向处理提供了支撑.

关键词 高分辨质谱数据，化学信息学，非靶向筛查，算法.

Cheminformatics in untargeted screening of liquid chromatography coupled to mass spectrometry data

SHEN Shanqi[1] *LI Aimin*[1,2]** *HUANG Yajuan*[1] *SHI Peng*[1] *PAN Yang*[1]
LI Wentao[1] *ZHANG Huaicheng*[1] *WU Ji*[1]

（1. State Key Laboratory of Pollution Control and Resource Reuse, School of the Environment, Nanjing University, Nanjing, 210023, China；2. Quanzhou Institute for Environmental Protection Industry, Nanjing University, Quanzhou, 362008, China）

Abstract The development of the chemical industry has resulted in the exposure of a huge number of unknown compounds in environment matrices. Identifying the chemicals is the key of assessing environmental risks of them and further attenuating them in environmental matrices. Liquid chromatography coupled to mass spectrometry (LC/MS) is a common technique of identifying compounds. However, the data collected by LC/MS is generally complex, which requires appropriate data analysis to reveal the information unrevealed in environmental samples. The development of analytical chemistry in untargeted screening of high-resolution mass spectrometry provides the possibility for compound structure identification. In this paper, we reviewed the application of analytical chemistry in untargeted screening, focused on the algorithm, software, compound database, spectrometry database and other aspects of the workflow such as process of peak extraction, de-

2021 年 7 月 1 日收稿(Received：July 1，2021).

* 中央高校基本科研业务费 (14380179)，江苏省重点研发计划社会发展重点项目 BE2019708）和泉州市高层次人才项目（2018CT006）资助.

Supported by the Fundamental Research Funds for the Central Universities (14380179)，Key Social Development Projects of Key R & D Plans in Jiangsu Province (BE2019708) and High-level Talent Team Project of Quanzhou City (2018CT006).

** 通信联系人 **Corresponding author**，E-mail：liaimingroup@nju.edu.cn

redundancy, prioritization, annotation and structure determination. In addition, the parameter optimization and data processing consistency of algorithms and software tools are discussed. This review provides a better support of untargeted processing of high-resolution mass spectrometry data.

Keywords High-resolution mass spectrometry data, cheminformatics, untargeted screening, algorithm.

化学工业的发展在给人们生活带来便利的同时，也造成了环境介质中大量的化学品残留[1]. 据统计，当前化学文摘（Chemical Abstract Service, CAS）数据库收录的化学品已达到 1.25 亿种[2]. 暴露在环境中的化学品不仅给生态系统造成威胁，也会通过各种途径进入人体进而损害人体健康. 如 Stehle 和 Schulz [3] 指出，杀虫剂的广泛使用造成大型无脊椎动物数量削减 30%. 通过食物链富集，暴露在环境中的双氯芬酸已造成印第安秃鹰数量的削减[4]. 因此，控制环境介质中化学品的含量对降低其对生态系统乃至人类的危害至关重要.

识别环境介质中的化学品是对其进行削减的前提，液相色谱串联高分辨质谱（liquid chromatography coupled to mass spectrometry, LC/MS)）是有机化学品检测与识别的最常用手段[2, 5]. 由于环境介质较为复杂，环境样品的质谱数据极为复杂，对其进行有效的解析需要一定的策略[6]. 非靶向筛查在当前环境样品的液相色谱质谱检测数据中广泛使用[7－8]. 非靶向筛查是指在没有标准品及样品先验信息的前提下，仅根据质谱数据识别样品中未知化合物信息的流程及方法[9]. 非靶向筛查的流程包括峰提取、去冗余、特征峰筛选、结构注释与鉴定等步骤[10]. 化学信息学在每个步骤的数据处理过程中扮演重要的角色. 如不同的数据处理步骤会涉及不同的算法[11]. 不同的数据处理软件在同一步骤的数据处理过程中使用不同的算法[12]. 对于同一步骤，不同的算法在处理数据时会产生一定的差异[13]. 因此，对非靶向筛查过程中化学信息学的阐述有利于科研工作者更好的使用不同的非靶向筛查数据处理软件，从而产生更可信数据分析结果.

基于此，针对基于液相色谱串联质谱数据的非靶向筛查流程，本文综述了如图 1 的分析流程，以及该流程中每个数据处理步骤所涉及的化学信息学知识，对不同非靶向数据处理软件在该过程中使用的算法及该数据处理过程中算法的发展及优劣进行了综述和比较. 在此基础上，对非靶向数据处理过程中所使用的化学信息学知识的未来发展方向进行了展望.

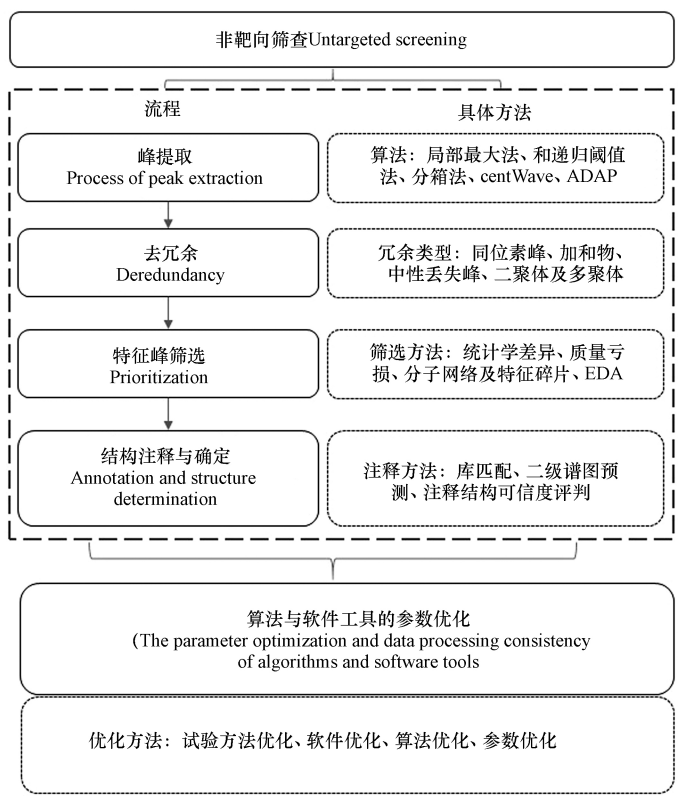

图 1 非靶向筛查数据处理流程

Fig.1 The workflow of untargeted screening of mass spectrometric data

1　峰提取 (Process of peak extraction)

在 LC/MS 原始数据中确定二维信号的边界、中心和强度称为峰提取[14]. 峰提取是 LC/MS 数据处理的第一步, 对后续的数据处理质量好坏有至关重要的影响. 峰提取算法经历了从局部最大法、递归阈值法[15]、分箱法[11]、centWave 算法[14] 到 ADAP 算法[13] 的发展, 这些峰提取算法已在不同时期的非靶向数据处理软件中使用.

局部最大法和递归阈值法类似, 均是对用户定义的 m/z 分箱范围内光谱数据进行离子色谱提取(Extracted ion chromatogram, XIC). 其主要步骤在于(1)在光谱数据维度提取色谱峰; (2)基于谱峰的强度值过滤掉强度较小的谱峰; (3)将不同分箱内的谱峰连接起来. 两种峰提取算法的不同之处在于, 局部最大法将光谱强度最大值视为最大的谱峰, 而递归阈值法将与噪音的尖峰宽度差别最大的峰宽定义为最大值进行数据处理. 最后, 对提取的峰在色谱维度和质谱(m/z)维度对强度较小的噪音进行过滤, 进而产生峰形较好的色谱峰[15]. 局部最大法和递归阈值法最初被用于 mzMine2 软件中. 这两种算法也被纳入第二代 mzMine 2 软件的峰提取过程中 [16] 局部最大法和递归阈值法计算速度较快, 对大数据集可以很快的计算出峰提取结果. 但是, 环境样品基质复杂, 目标化合物的色谱峰可能由于基质的干扰, 并不能达到优美的峰形, 这使得局部最大法的峰脊线的确定比较困难. 同样, 递归阈值法中其峰确定化合物峰和假阳性峰的阈值也比较难. centWave 算法较好的解决了上述两个问题, 对环境样品由于基质复杂导致的峰形较差过程中, 假阳性峰的鉴别效果有所提高.

分箱法是指根据高分辨质谱的分辨率, 在质谱维度根据合适的质量单位(如 0.1 m/z)进行分箱, 在每个分箱内, 取信号最强 m/z 的强度作为该分箱内的强度值, 然后将不同时间点该分箱的强度值相连即为提取的色谱峰[11]. 随后, 对提取的色谱峰进行模拟高斯峰匹配过滤, 从而改善峰形, 进一步的, 利用信噪比过滤的方式将噪音峰过滤以减少假阳性结果. 通过大量的研究结果表明, 信噪比 10 是比较合适的过滤值. 分箱法在最初基于 R 语言程序操作版本的 XCMS 中被使用. 利用该算法, Smith 等[11] 对 476 个血浆样品进行了峰提取, 每个样品中提取的峰平均数目为 3899 个. 在后续发展的基于网络平台的 XCMS 中, 分箱法的峰提取算法也被嵌入该非靶向数据处理流程中[17]. 虽然分箱法在对质谱数据处理时速度较快, 然而合适的分箱值可能难以获得. 如果分箱值过小, 色谱峰将在不同的箱之间存在, 这会导致在每个箱中检出的色谱峰部分分型损失. 如果分箱值过大, 每个箱内可能存在不同的色谱峰叠加, 这会使得强度较小的色谱峰被噪音掩盖[14, 18 − 19].

centWave 算法的原理在于, 具有化合物检出的质谱数据区域其强度更高(较噪音区域的强度信号), 可以先找出该轻度更高的区域进行峰检出. 在此基础上, 采用连续小波变化进行峰匹配和过滤. 这种过滤方式的好处在于可以产生预样品质谱数据峰宽等参数相匹配的模式峰, 得到可信度更高的峰提取结果[14]. 当前 centWave 算法在各类非靶向筛查数据处理软件中广泛使用.

基于 LC/MS 的峰提取 ADAP 算法是在基于气相色谱质谱 ADAP 算法基础上发展起来的, 主要用于解决 centWave 峰提取假阳性结果较多的问题[13, 20]. ADAP 算法首先将质谱第一次扫描得到的 m/z 按强度大小依次排列, 随后对第二次、第三次扫描中的 m/z 值也依次排列. 进而对不同扫描次序中出现的同一 m/z 在时间维度上将其相连, 形成色谱峰. ADAP 峰提取算法打破了如分箱算法及 centWave 算法需要认为设定封箱大小或 m/z 大小的做法, 从而避免了认为设定该参数可能造成的数据处理混乱[13, 21 − 22]. Myers 等 [13] 指出, 利用 ADAP 峰提取算法得到的色谱峰与基于 XCMS 和 mzMine 2 得到的色谱峰重合度为 83%, 然而 XCMS 和 mzMine 的 2 两种软件的峰提取算法提取的色谱峰重合度仅仅为 68%, 这表明 ADAP 可以在一定程度上减少假阳性峰的检出.

在峰提取的过程中, 平衡假阳性和假阴性峰的提取结果至关重要. 为了更好的覆盖所有可能的色谱峰, 相应的峰提取参数门槛应该被尽可能设置的更小[23], 然而这可能会导致更多的噪音峰被检出, 提高假阳性的峰提取结果. 相反, 又会增加假阴性峰的数目. 当前已有不同的尝试去解决这个问题. 如 Ju 等 [24] 利用 XCMS, MZmine 2 和 SIEVE 软件同时对样品数据进行峰提取, 通过叠加不同数据处理软件的峰提取结果, 从而较大提高的提取的色谱峰数目(1619 个、1103 个 XCMS、1500 个 MZmine2、387 个 SIEVE). Hu 等[23] 通过将 XCMS, MZmine2 及 MSDIAL 等软件提取的色谱峰数据分为三类: (1)好的一级峰, 好的二级谱图; (2)差的一级峰, 好的二级谱图; (3)差的一级峰, 好的二级谱图; (4)差

的一级峰，差的二级谱图或无二级谱图. 进而将第三类峰也纳入后续的化合物注释及结构鉴定，从而提高了化合物注释鉴定的结果.

2 去冗余(De-redundancy)

一般来说，化合物在液相色谱串联质谱检测过程中可能发生源内裂解、二聚体及多聚体的出现，以及同位素峰及加和物的出现，并非提取的每个色谱峰对应一个化合物信息[25－26]. 因此，对峰提取过程中的所有色谱峰进一步去冗余可以减少后续数据分析的复杂程度，增加数据分析的可信度[27]. 当前已有不同的算法或软件用于去除质谱中的冗余数据. Zeng 等[28] 提出离子融合的概念，用以将同位素峰、加和物、中性丢失峰聚到一起，从而达到理想的一个质谱特征峰指代一个化合物信息. 离子融合的原理在于同位素、加和物及中性丢失离子与前体离子间具有一定 m/z 差异的关联. 如同位素峰一般是由于 C_{12} 或者 C_{13} 造成的 m/z 为 1.0034 Da 的差异，加和物峰一般出现的是常见类型的加和物所造成的 m/z 差异. 根据这种关系构建算法，将血浆样品的正离子模式下，峰从 609 聚合到 106 个，负离子模式下的峰从 1084 聚合到 169 个，大大减少了冗余结果. DeFelice 等[29] 利用峰形相似性和这些离子对间的质量偏差进行皮尔逊和峰高相似性计算从而聚类同位素峰、加和物峰，并去除仪器等空白噪音. 在此基础上开发了 MS-FLO 软件，可以减少 7.8% 的冗余信息. Senan 等[30] 利用同位素峰及加和峰之间峰形及 m/z 差值之间的相似性，构建相似性网络，从而聚类同位素及加和物峰信息，并开发了 CliqueMS 软件，从而减少冗余峰信息. 根据同位素峰、源内裂解峰和加和物 (表 1) 峰形间的类似性(峰顶时间、左右半峰宽处的时间、以上 3 个时间之和)，通过 30 种标准品判断相似性大小从而聚类环境样品中冗余峰的信息. 利用这种方法，对尼罗河水样进行检测，从而实现对每个样品中平均 46% 冗余峰信息的扣除[31]. 利用这些同位素峰、源内裂解化合物及加和物间峰形的类似性，其他一些软件或算法也已经大量被开发或使用[32－36].

表 1 正负离子模式下常见的加和物

Table 1 The common adduct ion in positive and negative ion mode

ESI (+)	ESI (−)
M+H, M+Na, M+K	M-H, M+Na-2H, M+K-2H
$M+NH_4$, M+ACN+H	M+Cl, M+Br, M+FA-H
M+2ACN+H, M+ACN+Na, $M+CH_3OH+H$	M+Hac-H, M+TFA-H, $M-H_2O-H$

3 特征峰筛选 (Prioritization)

扣除冗余的色谱峰信息后，一般来说仍有大量的色谱峰信息. 这些信息可能并不全是我们需要关注的色谱峰信息，因此合适的特征峰筛选手段十分必要[36－37].

根据统计学知识进行检验是常用的特征峰筛选手段之一. 这种差异可以归类为时间前后的差异及空间上的差异. 利用时间上前后的差异，Gornik 等[38] 通过比较生物转化前后样品质谱检测数据提取到的峰强度差异(变化达到 10 倍以上，P 值小于 0.05)视为可能的转化产物峰，从而鉴定出 10 种舍曲林的生物转化产物. 采用类似的方法，Weizel 等 [39] 识别出污水生物处理过程中糖皮质激素的 41 种转化产物. Purschke 等 [40] 采用 PCA 及分组 PCA 分析了污水处理过程污染物变化趋势，并筛选出特征污染物氮-甲基吡咯烷酮. 利用空间上点的差异，Hohrenk 等[41] 采用主成分分析(PCA)及多元曲线分辨交替最小二乘法(MCR-ALS)对污水厂进水、出水及下游河流中的特征污染物进筛选，发现经过臭氧处理后，24 种污染物被削减至检测限以下.

根据同系物之间的质量亏损值是发现同系物特征峰的有效手段. 根据全氟化合物间同系物结构-CF2 结构的共性差异，利用质量亏损开发的算法识别出了一系列新的全氟化合物[42－44]. 根据这一特征，Koelmel 等[45] 开发了全流程自动化识别全氟化合物的软件，提高了全氟化合物的识别通量，减少了其识别的工作量和时间.

根据分子结构类似性的分子网络发现类似化合物是一种常用手段. Fu 等[46] 通过抗生素类化合物的特征碎片对食品中的抗生素类化合物及其产物进行筛选，从而发现 6 种抗生素类化合物. Zhan 等 [47]

通过已知神经毒气类化合物特征碎片离子筛选实际样品中的神经毒气类化合物, 从而发现一系列新的神经毒气类化合物. Esposito 等[48] 根据结构类似的化合物具有类似的二级谱图, 进而根据二级谱图类似原则构建分子网络, 发现了新的藻毒素类化合物[49]. 根据类似的原理, Le Dare 等[50] 将分子网络用于转化产物的鉴定, 从而筛选出两种奎硫平在人体内的代谢产物.

根据毒性值来识别是基于毒性效应特征峰的筛选手段[51]. Pochiraju 等[52] 利用污水水样的极性程度差别进行分馏, 针对分馏出各部分水样的不同毒性效应进行关键致毒化合物的识别, 最终识别出污水中具有雌激素效应的化合物(邻苯二甲酸二异丁酯、邻苯二甲酸二乙酯和苯甲酮).

4　注释与结构确定 (Annotation and structure determination)

对筛选出的特征峰进一步的注释和结构鉴定对下一步的生物学或毒理学意义解释至关重要. 化合物结构注释需要将样品的质谱数据与数据库中的谱图进行比对打分, 这其中涉及到的打分算法对匹配结果的可靠性至关重要[53－55]. 另外, 谱库中一般无法涵盖所有感兴趣化合物的二级谱图, 二级谱图预测软件在一定程度上解决上述问题[56]. 另外上述流程目前已经被整合到不同的数据处理软件中, 各个数据处理软件算法及所选择的谱库的差异也会影响最后的分析结果.

化合物注释是指将样品谱图信息与谱库中已知化合物信息进行匹配, 进而确定化合物的结构[37]. 匹配样品质谱数据和谱库质谱数据的过程需要合适的匹配算法, 当前 DP(dot-product)算法使用最为广泛[54]. DP 算法是指将质谱数据的 m/z 数值和峰强度数值看做二维向量, 从而对样品质谱数据和谱库中对应的质谱数据进行余弦相似性计算, 计算所得的余弦值大小即可反映两谱图类似性大小. 进而, 根据规定的余弦类似性阈值确定谱库中该谱图对应的化合物是否是样品谱图对应的化合物[53]. 利用该算法, 目前已有各种不同的非靶向处理软件或平台实现了自动化的谱图匹配、打分, 最终筛选出了最可能的对应结构的化合物, 从而达到样品谱图数据自动注释的功能[57－61]. 如 MS-DIAL 在对原始样品的质谱数据进行峰提取、去冗余等步骤后, 将样品质谱数据的 MS 和 MS/MS 信息与载入的谱图库进行对比, 进而根据余弦类似性大小判断识别出的结构的可信性[62-63]. 在对市政污水厂污水样品进行分析时, Qian 等[64] 利用 MS-DIAL 对样品数据进行处理和结构注释, 从而识别出 568 种化合物.

然而, 由于当前质谱谱图数据库中所含的化合物数据有限, 难以完全注释样品质谱数据中的特征峰, 谱图库中不存在的化合物结构注释困难. 利用二级谱图预测软件预测二级谱图从而对化合物结构进行注释或验证, 或者直接利用二级谱图预测软件预测的谱图数据扩充谱图库进而进行化合物结构注释可以解决上述难题[65]. 利用化合物在质谱仪中碎裂规则进行二级谱图预测的软件如 Metfrag[66] 和利用穷举法进行二级谱图的软件如 Mass Frontier[67] 均已有所应用. 为了扩大对环境样品中全氟化合物及其类似物的发现, Getzinger 等[68] 利用 CFM-ID 二级谱图预测软件[69] 扩充了全氟化合物的谱图库. 并在 DP 算法的基础上构建了自动匹配及注释的流程, 从而大大增加了环境样品中全氟化合物的识别数目. 对化合物结构进行注释后, 进一步的结构鉴定是确定所注释结构可信度的基础. 目前, Schymanski 等 [70] 提出的化合物结构鉴定可信度标准已被广泛接受.

5　算法与软件工具的参数优化 (The parameter optimization and data processing consistency of algorithms and software tools)

由于非靶向筛查的流程较长, 不同的算法与软件共同存在, 这会造成科研工作者在非靶向筛查质谱数据解析时面临困惑. 如 Hohrenk, Itzel, Baetz, Tuerk, Vosough 和 Schmidt [12] 比较了 MZmine2, enviMass, Compound Discoverer 和 XCMS 等 4 种常用的非靶向数据处理软件对同一批数据处理之间的一致性, 发现对同一批样品, 4 个软件间共同识别出的的峰仅占 10%, 不同软件的数据处理一致性问题较大. 作者指出各个软件之间算法的不同是造成之一差异的主要原因. Li 等[70] 从定量准确度、定性覆盖度等角度比较了 MS-Dial, MZmine 2, XCMS, MarkerView 及 Compound Discoverer 等 5 个软件的数据处理结果, 发现 5 个软件在非靶向定量结果上与靶向定量的结果差异较大, 其中 MZmine 2 是非靶向定量准确度最高的软件. 因此针对不同非靶向数据处理软件和算法间的参数优化以及结合不同软件在数据处理方面的优势可能对解决数据处理一致性问题或更好的展示非靶处理结果有用[71].

非靶向数据处理常用软件的参数优化方面，目前已有一些解决策略. Libiseller 等[72] 开发了基于 XCMS 非靶向筛查不同步骤参数优化的软件 IPO. 根据样品中天然同位素 C_{13} 和 C_{12} 相对丰度的比值关系构建优化方程，利用梯度下降法对 XCMS 非靶向数据处理过程中的参数如最小宽、最大峰宽等参数进行优化，从而将真阳性结果提高了 146%—361%，减少了 3%—8% 的假阳性结果. Eliasson 等 [73] 通过将样品稀释成一定的浓度梯度进行检测，从而根据真实的峰在不同样品间存在这种浓度梯度信息，而仪器或背景噪音则不存在这种梯度信息，从而设计了相应的算法对 XCMS 数据处理过程中的参数进行优化. 为了缩减计算机处理时间，该算法被进一步优化，从而使得数据处理的可信度提高了 19.4%（标准混合物样品）和 54.7%（人类尿液样品）[74]. McLean 等[75] 通过构建机器学习算法，利用梯度下降法不断优化 XCMS 及 mzMine2 在数据处理过程中的参数，从而得到最好的非靶向数据处理参数组合，减少假阳性，并是的检出的峰数目最多.

目前也有一些研究将不同软件及算法的优势进行结合，进而提高数据分析质量. 为了提供给用户以最好的软件及算法的组合使用策略，Helmus 等[10] 将目前开源软件中的算法汇总，开发出了 patRoon 的 R 语言软件包，从而提高了数据处理的可信度.

6 结论与展望 (Conclusions and perspectives)

本文详细介绍了化学信息学在非靶向筛查数据处理流程中的应用，重点阐述了该过程中涉及到的算法、软件、谱库等工具. 基于目前产物识别方面存在的问题，提出以下展望：

（1）不同算法和软件之间数据处理一致性问题. 应进一步加强对不同软件、算法对质谱数据处理结果的比较、数据一致性的探讨. 从而使得不同软件处理的质谱数据可以进行比较.

（2）结构注释时谱库过小. 应基于标准品和二级谱图预测得到的谱图进一步扩充谱图库，从而增加化合物结构注释的数目.

（3）加强机器学习算法在非靶向数据处理过程中的应用. 如利用机器学习算法更好的解决峰提取过程中假阳性与覆盖度低的问题.

参考文献 (References)

[1] FANG W D, PENG Y, MUIR D, et al. A critical review of synthetic chemicals in surface waters of the US, the EU and China [J]. Environment International, 2019, 131: 104994.

[2] HERNÁNDEZ F, BAKKER J, BIJLSMA L, et al. The role of analytical chemistry in exposure science: Focus on the aquatic environment [J]. Chemosphere, 2019, 222: 564-583.

[3] STEHLE S, SCHULZ R. Agricultural insecticides threaten surface waters at the global scale [J]. PNAS, 2015, 112(18): 5750-5755.

[4] OAKS J L, GILBERT M, VIRANI M Z, et al. Diclofenac residues as the cause of vulture population decline in Pakistan [J]. Nature, 2004, 427(6975): 630-633.

[5] TANG Y, CRAVEN C B, WAWRYK N J P, et al. Advances in mass spectrometry-based omics analysis of trace organics in water. Trends in Analytical Chemistry[J], 2020, 128, 115918.

[6] ESCHER B I, STAPLETON H M, SCHYMANSKI E L. Tracking complex mixtures of chemicals in our changing environment [J]. Science, 2020, 367(6476): 388-392.

[7] SCHYMANSKI E L, JEON J, GULDE R, et al. Identifying small molecules via high resolution mass spectrometry: Communicating confidence [J]. Environmental Science & Technology, 2014, 48(4): 2097-2098.

[8] ALYGIZAKIS N A, GAGO-FERRERO P, HOLLENDER J, et al. Untargeted time-pattern analysis of LC-HRMS data to detect spills and compounds with high fluctuation in influent wastewater [J]. Journal of Hazardous Materials, 2019, 361: 19-29.

[9] WANG X B, YU N Y, YANG J P, et al. Suspect and non-target screening of pesticides and pharmaceuticals transformation products in wastewater using QTOF-MS [J]. Environment International, 2020, 137: 105599.

[10] HELMUS R, TER LAAK TL, van WEZEL AP, et al. patRoon: Open source software platform for environmental mass spectrometry based non-target screening[J].Cheminform, 2021 , 13(1):1 .

[11] SMITH C A, WANT E J, O'MAILLE G, et al. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification [J]. Analytical Chemistry, 2006, 78(3): 779-787.

[12] HOHRENK L L, ITZEL F, BAETZ N, et al. Comparison of software tools for liquid chromatography–high-resolution mass spectrometry data processing in nontarget screening of environmental samples [J]. Analytical Chemistry, 2020,

92(2): 1898-1907.

[13] MYERS O D, SUMNER S J, LI S Z, et al. One step forward for reducing false positive and false negative compound identifications from mass spectrometry metabolomics data: New algorithms for constructing extracted ion chromatograms and detecting chromatographic peaks [J]. Analytical Chemistry, 2017, 89(17): 8696-8703.

[14] TAUTENHAHN R, BÖTTCHER C, NEUMANN S. Highly sensitive feature detection for high resolution LC/MS [J]. BMC Bioinformatics, 2008, 9(1): 1-16.

[15] KATAJAMAA M, ORESIC M, Processing methods for differential analysis of LC/MS profile data[J]. BMC Bioinformatics, 2005, 6: 179.

[16] PLUSKAL T, CASTILLO S, VILLAR-BRIONES A, et al. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data [J]. BMC Bioinformatics, 2010, 11(1): 1-11.

[17] TAUTENHAHN R, PATTI G J, RINEHART D, et al. XCMS online: A web-based platform to process untargeted metabolomic data [J]. Analytical Chemistry, 2012, 84(11): 5035-5039.

[18] STOLT R, TORGRIP R J O, LINDBERG J, et al. Second-order peak detection for multicomponent high-resolution LC/MS data [J]. Analytical Chemistry, 2006, 78(4): 975-983.

[19] ÅBERG K M, TORGRIP R J O, KOLMERT J, et al. Feature detection and alignment of hyphenated chromatographic-mass spectrometric data: Extraction of pure ion chromatograms using Kalman tracking [J]. Journal of Chromatography A, 2008, 1192(1): 139-146.

[20] NI Y, SU M M, QIU Y P, et al. ADAP-GC 3.0: Improved peak detection and deconvolution of co-eluting metabolites from GC/TOF-MS data for metabolomics studies [J]. Analytical Chemistry, 2016, 88(17): 8802-8811.

[21] DU X, SMIRNOV A, PLUSKAL T, et al. Metabolomics data preprocessing using ADAP and MZmine 2[J]. Methods Mol Biol. 2020;2104:25-48.

[22] MYERS O D, SUMNER S J, LI S Z, et al. Detailed investigation and comparison of the XCMS and MZmine 2 chromatogram construction and chromatographic peak detection methods for preprocessing mass spectrometry metabolomics data [J]. Analytical Chemistry, 2017, 89(17): 8689-8695.

[23] HU Y X, CAI B, HUAN T. Enhancing metabolome coverage in data-dependent LC-MS/MS analysis through an integrated feature extraction strategy [J]. Analytical Chemistry, 2019, 91(22): 14433-14441.

[24] JU R, LIU X Y, ZHENG F J, et al. A graph density-based strategy for features fusion from different peak extract software to achieve more metabolites in metabolic profiling from high-resolution mass spectrometry [J]. Analytica Chimica Acta, 2020, 1139: 8-14.

[25] BAKER E S, PATTI G J. Perspectives on data analysis in metabolomics: Points of agreement and disagreement from the 2018 ASMS fall workshop [J]. Journal of the American Society for Mass Spectrometry, 2019, 30(10): 2031-2036.

[26] KUHL C, TAUTENHAHN R, BÖTTCHER C, et al. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets [J]. Analytical Chemistry, 2012, 84(1): 283-289.

[27] SINDELAR M, PATTI G J. Chemical discovery in the era of metabolomics [J]. Journal of the American Chemical Society, 2020, 142(20): 9097-9105.

[28] ZENG Z D, LIU X Y, DAI W D, et al. Ion fusion of high-resolution LC-MS-based metabolomics data to discover more reliable biomarkers [J]. Analytical Chemistry, 2014, 86(8): 3793-3800.

[29] DEFELICE B C, MEHTA S S, SAMRA S, et al. Mass spectral feature list optimizer (MS-FLO): A tool to minimize false positive peak reports in untargeted liquid chromatography-mass spectroscopy (LC-MS) data processing [J]. Analytical Chemistry, 2017, 89(6): 3250-3255.

[30] SENAN O, AGUILAR-MOGAS A, NAVARRO M, et al. CliqueMS: a computational tool for annotating in-source metabolite ions from LC-MS untargeted metabolomics data based on a coelution similarity network [J]. Bioinformatics, 2019, 35(20): 4089-4097.

[31] KÖPPE T, JEWELL K S, DIETRICH C, et al. Application of a non-target workflow for the identification of specific contaminants using the example of the Nidda river basin [J]. Water Research, 2020, 178: 115703.

[32] BROECKLING C D, AFSAR F A, NEUMANN S, et al. RAMClust: a novel feature clustering method enables spectral-matching-based annotation for metabolomics data [J]. Analytical Chemistry, 2014, 86(14): 6812-6817.

[33] JU R, LIU X Y, ZHENG F J, et al. Removal of false positive features to generate authentic peak table for high-resolution mass spectrometry-based metabolomics study [J]. Analytica Chimica Acta, 2019, 1067: 79-87.

[34] DALY R, ROGERS S, WANDY J, et al. MetAssign: probabilistic annotation of metabolites from LC-MS data using a Bayesian clustering approach [J]. Bioinformatics, 2014, 30(19): 2764-2771.

[35] KOUŘIL Š, de SOUSA J, VÁCLAVÍK J, et al. CROP: correlation-based reduction of feature multiplicities in untargeted metabolomic data [J]. Bioinformatics, 2020, 36(9): 2941-2942.

[36] FRAISIER-VANNIER O, CHERVIN J, CABANAC G, et al. MS-CleanR: A feature-filtering workflow for untargeted LC-MS based metabolomics [J]. Analytical Chemistry, 2020, 92(14): 9971-9981.

[37] LJONCHEVA M, STEPIŠNIK T, DŽEROSKI S, et al. Cheminformatics in MS-based environmental exposomics: Current achievements and future directions [J]. Trends in Environmental Analytical Chemistry, 2020, 28: e00099.
[38] GORNIK T, KOVACIC A, HEATH E, et al. Biotransformation study of antidepressant sertraline and its removal during biological wastewater treatment [J]. Water Research, 2020, 181: 115864.
[39] WEIZEL A, SCHLÜSENER M P, DIERKES G, et al. Analysis of the aerobic biodegradation of glucocorticoids: Elucidation of the kinetics and transformation reactions [J]. Water Research, 2020, 174: 115561.
[40] PURSCHKE K, VOSOUGH M, LEONHARDT J, et al. Evaluation of nontarget long-term LC–HRMS time series data using multivariate statistical approaches [J]. Analytical Chemistry, 2020, 92(18): 12273-12281.
[41] HOHRENK L L, VOSOUGH M, SCHMIDT T C. Implementation of chemometric tools to improve data mining and prioritization in LC-HRMS for nontarget screening of organic micropollutants in complex water matrixes [J]. Analytical Chemistry, 2019, 91(14): 9213-9220.
[42] WANG X B, YU N Y, QIAN Y L, et al. Non-target and suspect screening of per- and polyfluoroalkyl substances in Chinese municipal wastewater treatment plants [J]. Water Research, 2020, 183: 115989.
[43] LI Y Q, YU N Y, DU L T, et al. Transplacental transfer of per- and polyfluoroalkyl substances identified in paired maternal and cord sera using suspect and nontarget screening [J]. Environmental Science & Technology, 2020, 54(6): 3407-3416.
[44] WANG Y, YU N Y, ZHU X B, et al. Suspect and nontarget screening of per- and polyfluoroalkyl substances in wastewater from a fluorochemical manufacturing park [J]. Environmental Science & Technology, 2018, 52(19): 11007-11016.
[45] KOELMEL J P, PAIGE M K, ARISTIZABAL-HENAO J J, et al. Toward comprehensive per- and polyfluoroalkyl substances annotation using FluoroMatch software and intelligent high-resolution tandem mass spectrometry acquisition [J]. Analytical Chemistry, 2020, 92(16): 11186-11194.
[46] FU Y Q, ZHANG Y H, ZHOU Z H, et al. Screening and determination of potential risk substances based on liquid chromatography–high-resolution mass spectrometry [J]. Analytical Chemistry, 2018, 90(14): 8454-8461.
[47] ZHANG M J, LIU Y L, CHEN J, et al. Sensitive untargeted screening of nerve agents and their degradation products using liquid chromatography-high resolution mass spectrometry [J]. Analytical Chemistry, 2020, 92(15): 10578-10587.
[48] ESPOSITO G, TETA R, MARRONE R, et al. A fast detection strategy for cyanobacterial blooms and associated cyanotoxins (FDSCC) reveals the occurrence of lyngbyatoxin A in Campania (South Italy) [J]. Chemosphere, 2019, 225: 342-351.
[49] TETA R, DELLA SALA G, GLUKHOV E, et al. Combined LC–MS/MS and molecular networking approach reveals new cyanotoxins from the 2014 cyanobacterial bloom in green lake, Seattle [J]. Environmental Science & Technology, 2015, 49(24): 14301-14310.
[50] le DARÉ B, FERRON P J, ALLARD P M, et al. New insights into quetiapine metabolism using molecular networking [J]. Scientific Reports, 2020, 10(1): 19921.
[51] HOLLENDER J, SCHYMANSKI E L, SINGER H P, et al. Nontarget screening with high resolution mass spectrometry in the environment: Ready to go? [J]. Environmental Science & Technology, 2017, 51(20): 11505-11512.
[52] POCHIRAJU S S, LINDEN K, GU A Z, et al. Development of a separation framework for effects-based targeted and non-targeted toxicological screening of water and wastewater [J]. Water Research, 2020, 170: 115289.
[53] SCHEUBERT K, HUFSKY F, PETRAS D, et al. Significance estimation for large scale metabolomics annotations by spectral matching [J]. Nature Communications, 2017, 8: 1494.
[54] STEIN S E, SCOTT D R. Optimization and testing of mass spectral library search algorithms for compound identification [J]. Journal of the American Society for Mass Spectrometry, 1994, 5(9): 859-866.
[55] VINAIXA M, SCHYMANSKI E L, NEUMANN S, et al. Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects [J]. TrAC Trends in Analytical Chemistry, 2016, 78: 23-35.
[56] HUFSKY F, SCHEUBERT K, BÖCKER S. New kids on the block: Novel informatics methods for natural product discovery [J]. Natural Product Reports, 2014, 31(6): 807.
[57] XUE J C, GUIJAS C, BENTON H P, et al. METLIN MS2 molecular standards database: A broad chemical and biological resource [J]. Nature Methods, 2020, 17(10): 953-954.
[58] TSUGAWA H, CAJKA T, KIND T, et al. MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis [J]. Nature Methods, 2015, 12(6): 523-526.
[59] DÜHRKOP K, FLEISCHAUER M, LUDWIG M, et al. *SIRIUS* 4: A rapid tool for turning tandem mass spectra into metabolite structure information [J]. Nature Methods, 2019, 16(4): 299-302.
[60] RÖST H L, SACHSENBERG T, AICHE S, et al. OpenMS: A flexible open-source software platform for mass spectrometry data analysis [J]. Nature Methods, 2016, 13(9): 741-748.
[61] WANG M X, CARVER J J, PHELAN V V, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking [J]. Nature Biotechnology, 2016, 34(8): 828-837.

［62］ TSUGAWA H, IKEDA K, TAKAHASHI M, et al. A lipidome atlas in MS-DIAL 4［J］. Nature Biotechnology, 2020, 38(10): 1159-1163.

［63］ TSUGAWA H, NAKABAYASHI R, MORI T, et al. A cheminformatics approach to characterize metabolomes in stable-isotope-labeled organisms［J］. Nature Methods, 2019, 16(4): 295-298.

［64］ QIAN Y L, WANG X B, WU G, et al. Screening priority indicator pollutants in full-scale wastewater treatment plants by non-target analysis［J］. Journal of Hazardous Materials, 2021, 414: 125490.

［65］ ALLARD P M, GENTA-JOUVE G, WOLFENDER J L. Deep metabolome annotation in natural products research: Towards a virtuous cycle in metabolite identification［J］. Current Opinion in Chemical Biology, 2017, 36: 40-49.

［66］ RUTTKIES C, NEUMANN S, POSCH S. Improving MetFrag with statistical learning of fragment annotations［J］. BMC Bioinformatics, 2019, 20(1): 1-14.

［67］ SCHYMANSKI E L, GALLAMPOIS C M J, KRAUSS M, et al. Consensus structure elucidation combining GC/EI-MS, structure generation, and calculated properties［J］. Analytical Chemistry, 2012, 84(7): 3287-3295.

［68］ GETZINGER G J, HIGGINS C P, FERGUSON P L. Structure database and in silico spectral library for comprehensive suspect screening of per- and polyfluoroalkyl substances (PFASs) in environmental media by high-resolution mass spectrometry［J］. Analytical Chemistry, 2021, 93(5): 2820-2827.

［69］ DJOUMBOU-FEUNANG Y, PON A, KARU N, et al. CFM-ID 3.0: Significantly improved ESI-MS/MS prediction and compound identification［J］. Metabolites, 2019, 9(4): 72.

［70］ LI Z C, LU Y, GUO Y F, et al. Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection［J］. Analytica Chimica Acta, 2018, 1029: 50-57.

［71］ COBLE J B, FRAGA C G. Comparative evaluation of preprocessing freeware on chromatography/mass spectrometry data for signature discovery［J］. Journal of Chromatography A, 2014, 1358: 155-164.

［72］ LIBISELLER G, DVORZAK M, KLEB U, et al. IPO: a tool for automated optimization of XCMS parameters［J］. BMC Bioinformatics, 2015, 16: 118.

［73］ ELIASSON M, RÄNNAR S, MADSEN R, et al. Strategy for optimizing LC-MS data processing in metabolomics: A design of experiments approach［J］. Analytical Chemistry, 2012, 84(15): 6869-6876.

［74］ ZHENG H, CLAUSEN M R, DALSGAARD T K, et al. Time-saving design of experiment protocol for optimization of LC-MS data processing in metabolomic approaches［J］. Analytical Chemistry, 2013, 85(15): 7109-7116.

［75］ MCLEAN C, KUJAWINSKI E B. AutoTuner: High fidelity and robust parameter selection for metabolomics data processing［J］. Analytical Chemistry, 2020, 92(8): 5724-5732.