工作总结与计划（1 月 21 日到 2 月 6 日）

代驾

dg310012@mail.ustc.edu.cn

2022-2-6

目录

文档简介

本文档主要分为两个部分。

第一部分是对最近的工作进行总结，最近我阅读了 DI-SPA，csodiaq 和 RI-FIGS 三篇论文，并运行测试了 csodiaq 论文中提供的图形化软件。唐师兄的代码也已经在服务器上配置好了环境。

第二部分是对未来工作计划的一个安排。

工作总结

论文阅读

首先是 DI-SPA 论文，这篇论文之前我已经看过了，这段时间又仔细看了一遍，有了更多的理解和体会。

这篇论文的标题是直接输注法定量鸟枪蛋白质组学分析。液相色谱-串联质谱技术为蛋白质组学提供了敏感的肽分析，但是需要大量分析时间，降低了吞吐量。论文中证明了气相肽分离技术可以代替液相色谱技术实现快速蛋白质组学分析。通过直接输注-鸟枪蛋白质组分析（DI-SPA）可以实现快速无偏的蛋白质组量化，提高蛋白质定量的吞吐量。

在最近的 MS 进展中，离子迁移率使得气相肽阳离子分离的另一个维度成为可能，离子迁移率分离根据电荷和形状对气相肽离子进行分类。高场非对称波形离子迁移谱仪（FAIMS）可以通过放置在质谱仪的电喷雾发射器和大气压入口之间的装置实现非常快速的气相分离。FAIMS 根据离子在高低不对称场中的不同迁移率，通过内外电极过滤离子。通过 FAIMS 和其他离子迁移率方法分离分析物可能会改进无 LC 复杂肽混合物的分析。

DI-SPA 的策略是气相分离可以替代液相色谱快速分析蛋白质组中的复杂肽混合物。肽样品通过电喷雾直接注入和电离，在 DIA 使用高分辨率 MS/MS 检测之前，产生的肽阳离子在气相中分离。论文中探索了 DI-SPA 数据收集参数，发现气相分离的程度与可观察蛋白质组覆盖的深度呈正相关。并通过实验验证了 DI-SPA 能够在不进行 LC 分离的情况下进行快速蛋白质组分析。

DI-SPA 采用三种主要技术在气相中分离肽：（1）离子迁移率（FAIMS）；（2）基于 m/z 的四极质谱过滤器分离；（3）离子离解。其中离子迁移率是成功的关键决定因素。DI-SPA 通过完全省略 LC，将缩短 LC 分离的概念发挥到了逻辑极限。为了实现 DI-SPA，需要几种解决方案的组合：（1）通过离子迁移率实现额外的分离维度，（2）通过 DIA 进行数据收集，（3）使用投影光谱概念进行肽鉴定，（4）共分离重标记标准肽，以便能够从片段中进行量化。与最近的研究相比，DI-SPA 侧重于通过更短的液相色谱分离进行更快的分析。DI-SPA 第一次迭代的一些缺点是，它还不适合执行无标记定量，并且还没有应用于生物流体中蛋白质的高通量定量。

之后又看了 csodiaq 论文。直接注射鸟枪蛋白质组分析（DIS-PA）是一种基于快速质谱的蛋白质组学新方式，但原始的数据分析工作流程繁重。CsoDIAq 是一个用户友好的软件包，它可以从 DISPA 数据中识别和量化肽和蛋白质。除了通过图形用户界面建立完整的自动化分析工作流程外，CsoDIAq 还引入了算法概念，以提高肽识别速度和灵敏度。其中包括减少搜索时间复杂性的频谱池，以及一种称为匹配计数和余弦（match count and cosine，MaCC）的新频谱匹配分数，该分数提高了目标诱饵分析中的目标识别能力。论文中说明了片段质量耐受性校正后的再分析增加了肽鉴定的数量，在将 CsoDIAq 应用于标准 LC-MS DIA 后，说明了其性能优于其他频谱匹配软件。

几乎所有的蛋白质组学实验都依赖于 LC 在电离和质谱分析之前分离肽。蛋白质组学领域正在经历一种缩短 LC 梯度的趋势。逻辑的极端是完全删除 LC，DISPA 不使用 LC 分离，而是依赖于额外的气相离子迁移分馏。由于直接输注数据缺乏随时间推移的肽片段共洗脱，DISPA 的原始报告依赖于 MSPLIT-DIA 的预测余弦评分来识别肽和蛋白质。然而，由于 MSPLIT-DIA 不是针对 DISPA 数据定制的，也不是天生识别蛋白质的，因此需要多个定制 python 和 R 脚本来实现 FDR 计算、蛋白质识别和量化。

论文中描述了 CsoDIAq(用于 DIA 定性和定量分析的余弦相似性优化),这是一个 python 软件包，旨在增强 MSPLIT-DIA 最初使用的投影光谱概念的可用性和敏感性。CsoDIAq 引入了一些算法改进，包括汇集光谱峰值以缩短时间复杂性和一个新的光谱评分功能，提高了目标和诱饵肽的区分。结合图形用户界面（GUI），CsoDIAq 既有效又友好，可以分析来自 DISPA 和 LC-MS 的 DIA 数据。下面是论文中提到的一些具体方法

1、光谱池技术

CsoDIAq 引入了一种称为“谱池”的库查询峰值比较方法，该方法将时间复杂度降低了一个指数因子。在 DIA 分析的任何给定 m/z 窗口中，四个变量主要影响算法的速度，即对应于该窗口的库谱数（nLS）；nLS 库光谱中碎片离子峰的总数（pLS）；查询光谱的数量（NQ）；以及 nQS 查询光谱（pQS）中碎片离子峰的总数。MSPLIT-DIA26 迭代地将每个库谱与每个查询谱进行比较，假设库谱表示的肽的前体质量在查询谱捕获的 m/z 窗口内。如果将上述变量分别分配给 nLS、pLS、nQS 和 pQS 的字母值，则该方法的时间复杂度为：nQS * pLS + nLS * pQS。这些因素的变化会显著影响完成算法所需的时间长度。在大 O 表示法中，上述方程的时间复杂度为 O（n*m）。光谱池减少了峰值比较中不必要的重复，显著提高了速度，而不影响准确性。MSPLIT-DIA 分别将查询频谱与每个相关库频谱进行比较，因此，在给定的 m/z

查询窗口内，每一个具有前体 m/z 的频谱引用一次来自一种频谱类型的相同峰值。光谱合并是指除了固有的质量和强度值之外，还为每个碎片离子分配一个光谱标签，这允许将多个光谱合并为一个光谱进行比较。在使用光谱标签计算单独的匹配分数进行匹配后，可以分离合并光谱中片段的匹配。因此，通过比较池查询频谱和池库频谱，任何峰值都只会被引用一次。这大大降低了上述传统方法的时间复杂度实验设计。在大 O 表示法中，上述方程的时间复杂度为 O（n+m）。对于每个 FAIMS 补偿电压设置，DISPA 侦察实验至少在同一 m/z 查询窗口上迭代一次。根据上述方程式，nQS（查询光谱的数量）通常等于实验中运行的补偿电压设置数。论文中通过设计对照实验说明了使用频谱池可以降低运行的时间。

2、评分方法

CsoDIAq 采用新的光谱-光谱匹配 (SSM) 评分功能，改进目标和诱饵肽分布的分离，以优化低于标准错误发现率 (FDR) 截止值的肽命中数。CsoDIAq 首先取光谱库和实验光谱中碎片离子峰强度的平方根，以标准化碎片离子强度的贡献。接下来，对于每个实验光谱，将碎片离子与所有可能匹配的合并库光谱进行比较。与 MSPLIT-DIA 一样，CsoDIAq 使用“投影频谱”概念；只有在库光谱中碎片离子的定义质量容差范围内发现的实验碎片离子才用于计算分数。片段比较使用百万分之几 (PPM) 而不是绝对 m/z 差异。记录所有匹配的碎片离子，然后用于计算池库中所有可能肽段的 SSM 分数。CsoDIAq 通过将匹配碎片数的五次方根乘以余弦分数来计算 SSM 分数。由于峰匹配对 SSM 分数的重要性和影响，CsoDIAq 将三个碎片离子匹配的最小值强加到库光谱中，没有最大值。

3、碎片质量的误差修正过程

CsoDIAq 采用双重搜索策略进行碎片离子质量校正。当比较库峰值和查询峰值时，实际对应片段的 m/z 值预计不会精确匹配。除了质谱仪的自然变化导致的查询光谱的一般误差范围外，质量校准中的漂移可能会导致系统质量值偏移。为了对此进行调整，CsoDIAq 使用 0 PPM 的通用偏移量和 30 PPM 的默认用户可调公差，对数据进行初始的、未修正的分析。这些数字基于之前的实验，该实验表明，除了计算优化公差所需的足够数据外，30ppm（约 0ppm）的总窗口还可以捕获真实偏移量。在使用前面描述的评分方法确定感兴趣的肽后，csoDIAq 根据这些点击的 PPM 差异确定新的偏移量和耐受性。默认情况下，偏移量和公差根据 PPM 分布进行自定义。偏移量是给定直方图的最高仓位值，公差是距离偏移量最远的仓位，其值大约等于周围的噪声。用户也可以选择分别使用偏移和公差的中值和标准偏差。然后，CsoDIAq 排除所选 PPM 偏移量和公差之外的所有峰值匹配，从而得到在唯一标识数量上优于

未修正分析的修正分析。

4、肽和蛋白质的鉴定过程

CsoDIAq 为每个质谱文件生成三个输出文件，报告过滤掉 FDR<1%的光谱、肽和蛋白质。在每种情况下，CsoDIAq 根据上述分数对肽识别进行排序，使用目标诱饵方法的修改计算每个识别的 FDR，其中分数 S 处的 FDR= #诱饵 / #目标，并移除低于 0.01 FDR 阈值的 SSM。返回光谱报告时不使用独特的肽过滤。肽 FDR 计算仅使用每个肽的所有 SSM 中得分最高的实例。CsoDIAq 使用 Idicker 算法从发现的肽列表中识别蛋白质组，并将其作为附加列添加到输出中。TraML 光谱库中的蛋白质组用于蛋白质推断，而不是将肽匹配回 FASTA 文件中的蛋白质条目。论文中对 Idicker 算法的实现优先识别在肽减少步骤后具有更多肽连接的蛋白质。当出现连接时，该算法将使用每个蛋白质最初的肽连接数。蛋白质的蛋白质 FDR 计算使用得分最高的肽作为蛋白质组得分，尽管肽 FDR 输出中与这些蛋白质连接的所有肽都重新包含在蛋白质报告中以供参考。

5、蛋白质的定量方法

准确的蛋白质定量需要第二次 DIA 数据收集，以 m/z 和补偿电压（CV）值为目标，对应于已识别蛋白质的最佳肽靶。CsoDIAq 使用两个标准来选择每个蛋白质的代表性肽。首先，不考虑特定蛋白质所特有的肽。接下来，CsoDIAq 根据匹配片段离子的片段离子强度总和对每个蛋白质中的肽进行排序。最后，该软件允许用户从每个蛋白质中输入所需的最大代表性肽数。例如：CsoDIAq 使用默认的初始公差 30ppm，然后选择性地应用前面讨论的相同质量校正算法，以确定特定于 DISPA 运行的偏移量和公差。在确定匹配的峰（默认值：前三个最密集峰中的至少一个）后，CsoDIAq 根据确定的峰（默认值：中值比值）计算每个肽的 SILAC 比率。用户可以输入（1）用于确定校正过程公差的标准偏差，（2）计算肽的 SILAC 比率所需的最小匹配数，以及（3）比率选择模式。其中用于目标重新分析的文件不会包含蛋白质 FDR 文件中的所有前导蛋白质。这是因为诱饵将被移除，而且 Idicker 算法识别的一些蛋白质组不会有独特的肽靶点可供使用。

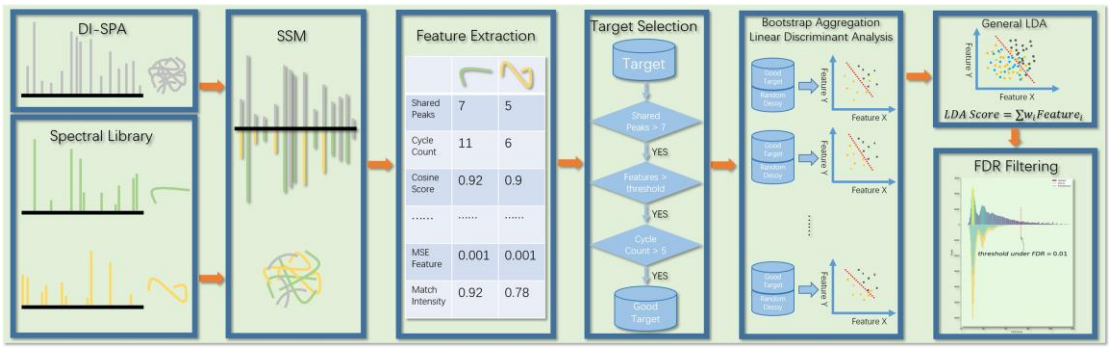
之后阅读了唐师兄的论文 RI-FIGS。

DI-SPA 与 DIA-MS 的结合实现了快速的无色谱蛋白质定性与定量，但是它无法处理无标记的 DI-SPA 数据。在没有色谱法的情况下，可以通过扩展和最大限度地利用特征“pure”重复特征，并结合核心功能的改善，来促进对 DI-SPA 的识别。这里提出一种基于 FIGS 的重复定量方法 RI-FIGS，来进行 DI-SPA 数据的定性与定量，并且具备较高的肽识别度和重复性。

利用 DI-SPA 可以稳定快速的进行蛋白质定性定量，但是目前用于分析 DI-SPA 数据的软件工具在蛋白质组学的研究中受到了限制。例如，通过投影光谱，MSPLIT-DIA 可以提供有限的肽识别，而无需定量信息。尽管 csodiaq 能够以更高的准确度鉴定肽段，但其定量只能用于重标签样品和严格的肽段，因此需要设计出一个可以用于 DI-SPA 数据特别是无标记样本的软件工具。

在没有色谱的情况下，DI-SPA 数据中”pure”这一重复性质可以通过循环扩展得到充分利用，这里提出了一种称为 RI-FIGS 的方法可以用于无标记肽段定量实验。RI-FIGS 可以使用一种线性判别分析方法来自动校正肽谱匹配（SSM），并根据混合实验光谱动态地获得肽等式。为了实现不同类型数据采集（DDA、DIA 和 DI-SPA）的通用性，我们充分利用了 DI-SPA 的特性，并在不存在保留时间信息的情况下通过 RI-FIGS 设计了实验流程，可以快速获取海量光谱数据，并进行识别和量化。

其算法流程如下图所示：

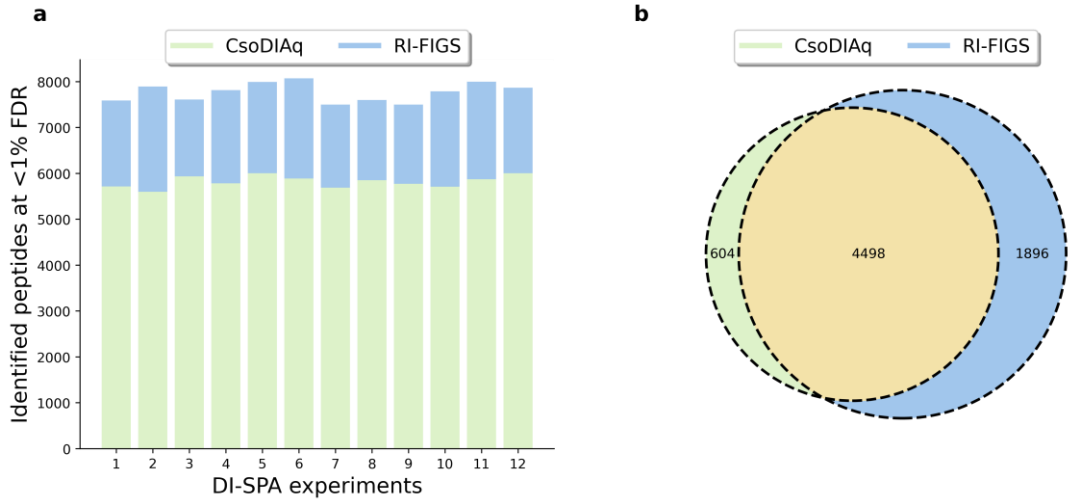


首先根据质谱数据和图谱库数据进行肽谱匹配（SSM），之后对匹配上的图谱进行特征提取，从而选择出那些匹配程度较好的图谱，然后利用这些较优图谱对 LDA 分类器进行训练，最后根据训练好的 LDA 分类器对应 FDR 控制，将匹配较差的图谱过滤掉。

下面介绍论文的实验结果：

1、RI-FIGS 加速了肽段的定性

用 RI-FIGS 对 DI-SPA 数据集进行了测试（12 次重复实验），保证与 csodiaq 使用相同的数据集与处理标准（包括使用相同的图谱库，相同的 FDR 控制方法，相同的反库）。相比于 csodiaq，RI-FIGS 鉴定出了更多的肽段（提升了 25.3%到 33.5%），并且可以找到更多的独特肽段。除此之外超过 82.3%的肽段可以被重复定量（12 次重复实验被定性出 8 次）。具体结果如下图所示。a 图是 12 次重复实验 csodiaq 与 RI-FIGS 的肽段定性结果数量图，b 图是那些被重复定量出的肽段的唯一分布情况。



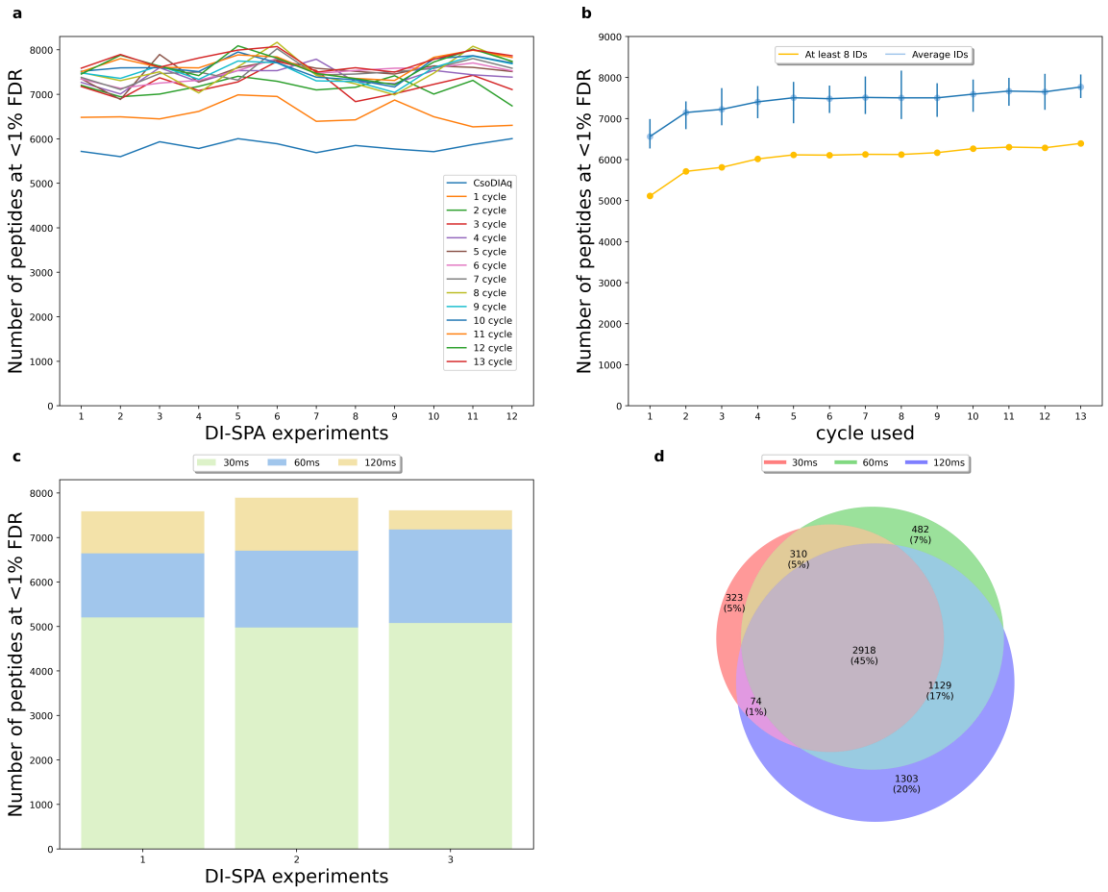
2、重复 DI-SPA 实验可以提升定性结果

循环特征和机器学习算法的使用可以增加肽的定性数量，为了进一步探索最佳 DI-SPAdata 的采集参数，以便在可接受的时间内以较高的置信度进行大规模识别，检查了可能影响总运行时间的光谱仪设置，即用于数据采集的循环计数和用于实现隔离窗口的注射时间。通常情况下，随着周期和注射时间的增加，肽识别的数量增加。在 120ms 的喷射时间下，首先检测了被识别的肽的数量。在最开始的四个周期中，已识别的肽增加，但是在第五个周期以后提升情况会减弱。但是总的来说，RI-FIGS 在所有实验中都优于 csodiaq。这一点可以通过反库的肽谱匹配来说明。来自反库和噪声峰的肽谱匹配不太可能重复发生，所以 SSM 中出现的重复特征将会是肽段定性的重要判断依据。

之后比较了数据采集过程中不同的注入时间。窗口的注入时间越短，获取时间越短，光谱质量越差，因此存在一个平衡点，使得注入时间和光谱质量都在一个较为理想的范围内。实验表明，将注入时间从 30ms 提升到 60ms 可以定性出更多的肽段。关于注入时间这一参数，利用 DI-SPA 数据进行了多次试验。与 120ms 数据相比，60ms 数据定性出的肽段数量是其定性肽段数量的 88.9%，30ms 数据定性出的肽段数量是其定性肽段数量的 66.1%。最终选择 60ms

的注入时间和 4 轮采集这一参数选项，这种情况的采集时间为 7-8 分钟。

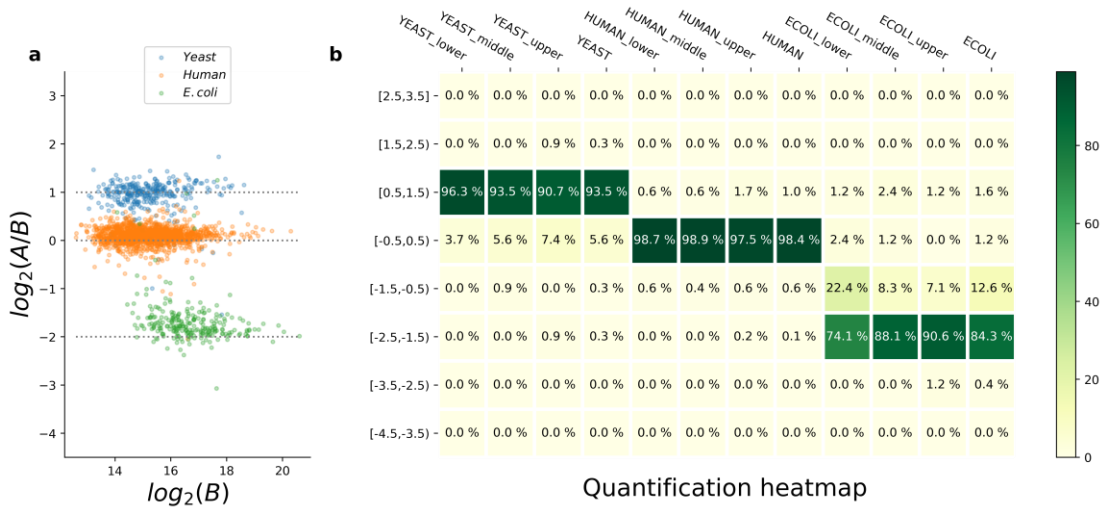
这一部分的实验结果如下图所示。a 图是不同 cycle 下 DI-SPA 数据的肽段定性数量，b 图说明了 cycle 与定性数量之间的变化关系，c 图是不同注入时间的肽段定性数量，d 图是不同注入时间下独特肽段定性情况。



3、RI-FIGS 可以用于 DI-SPA 数据的 LFQ 实验

目前 csodiaq 只能应用到有标签样本上，对于 DI-SPA 数据而言，还没有对应的 LFQ 分析工具。这里利用 RI-FIGS 在两种包含三类蛋白质的混合样品上进行了测试，两样品为 A 样品和 B 样品，都由酵母菌蛋白质、人类蛋白质和大肠杆菌蛋白质组成。A、B 样品的酵母菌蛋白质、人类蛋白质、大肠杆菌蛋白质浓度之比的对数（log2）分别为 1,0，-2，每个样品进行 5 次重复实验。考察 5 次实验至少出现 4 次的肽段，A 样品中共定性出 3855 个肽段，B 样品中共定性出 3959 个肽段，其中有 2849 个肽段被两样品共享。为了对实验结果进行评估，将定性肽段按照 B 样品中丰度的大小分为三份进行比较分析，得到相应的结果。

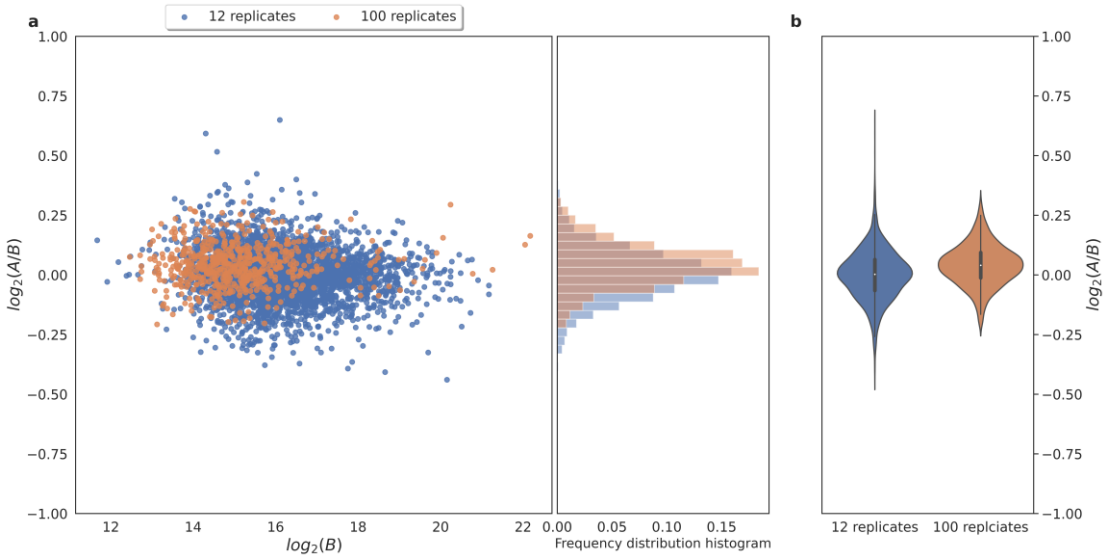
这一部分的实验结果如下所示，a 图是三类肽段 A 样品与 B 样品定量结果的比值的对数对应的散点图，b 图是不同物种和三分位数对应的量化热图。



4、RI-FIGS 具有稳健性和可生产性

对 RI-FIGS 进行重复实验，随机将数据集分为两个子集（人工样本 A 和 176 B），以计算出一个预期的均衡比率。从这两类子集中一共鉴定出 5072 和 1304 个“稳定”肽。对两个数据集进行了识别后发现定量结果的比值都分布在预期范围内，这说明 RI-FIGS 可以在不影响肽识别的情况下准确的执行。

这一部分的实验结果如下所示，下图表示了 A 样品和 B 样品的虚拟比率分布，a 图和 b 图分别是散点图和小提琴图。

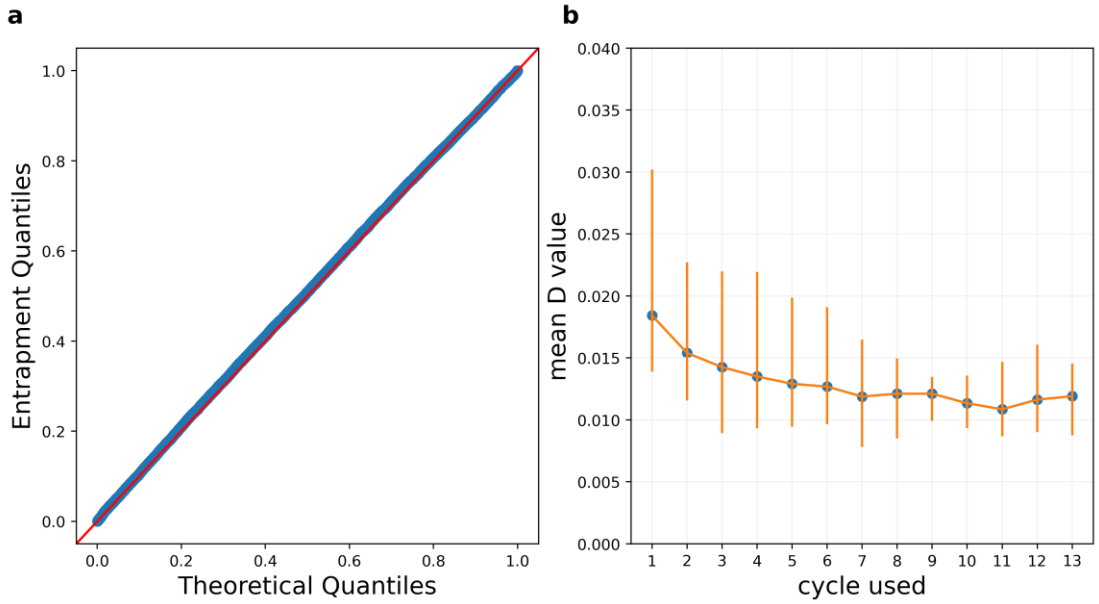


5、RI-FIGS 可以进行准确的 FDR 控制

通过比较诱饵库肽谱匹配 p-value 分布，可以判断出评分函数的质量（是否具有无偏

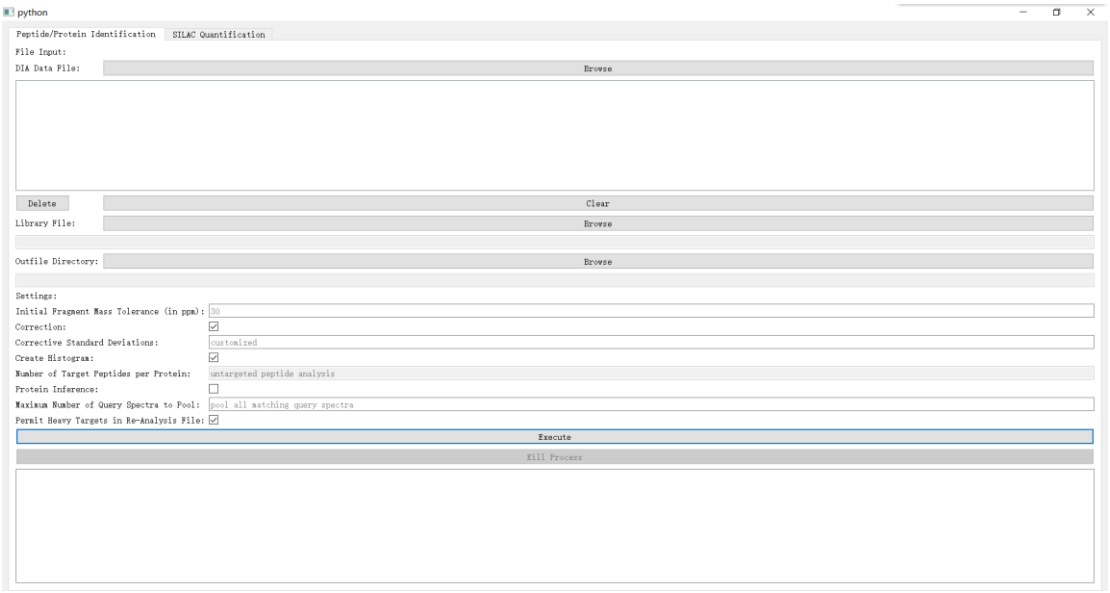
性）。在所有识别结果的基础上，计算了最终的识别值（D value），并对其取平均数。基于 K-S test 的 D value 值，错误的 SSM 和对应的 FDR 有一个中间值描述，从而帮助进行 FDR 控制。

这一部分的实验结果如下所示，a 图是诱饵库对应的定量结果，b 图是基于 K-S test 的平均 D value。

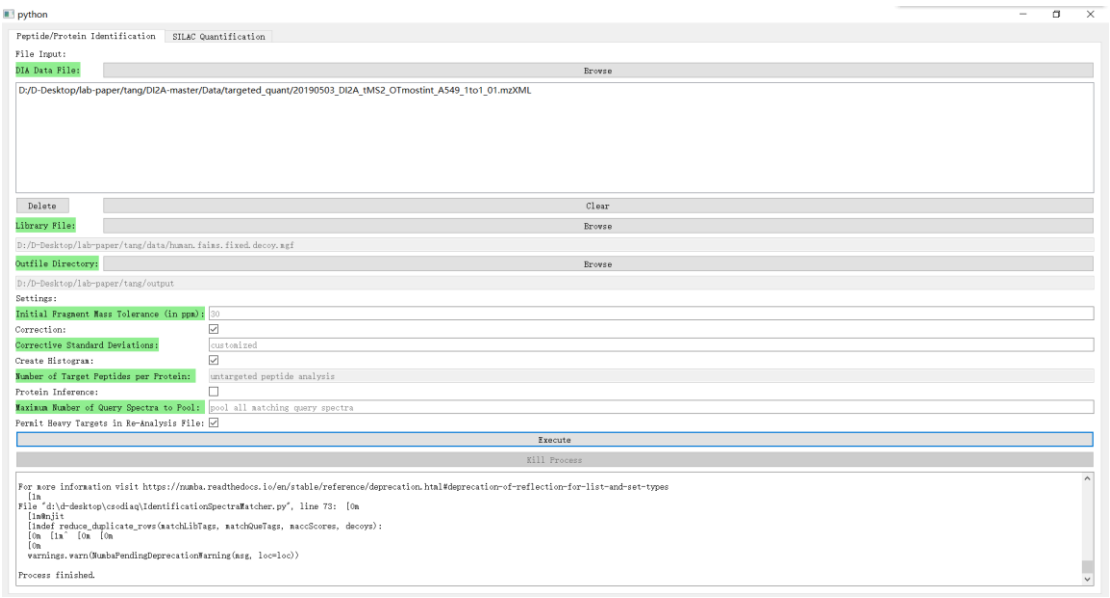


软件运行

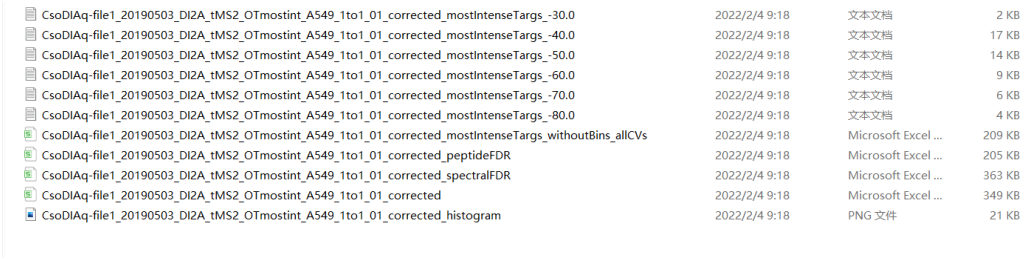
由于 csodiaq 是对 DI-SPA 的一个集成和优化，所以我只对 csodiaq 进行了复现运行。根据 csodiaq 论文中提供的 github 地址，下载对应的源码并配置环境，在命令行中输入 csodiaq gui 进入图形化软件界面。



软件共有两个功能，一是对肽段进行定性，二是对有标记数据进行定量。首先利用论文中提供的有标记数据进行肽段定性。

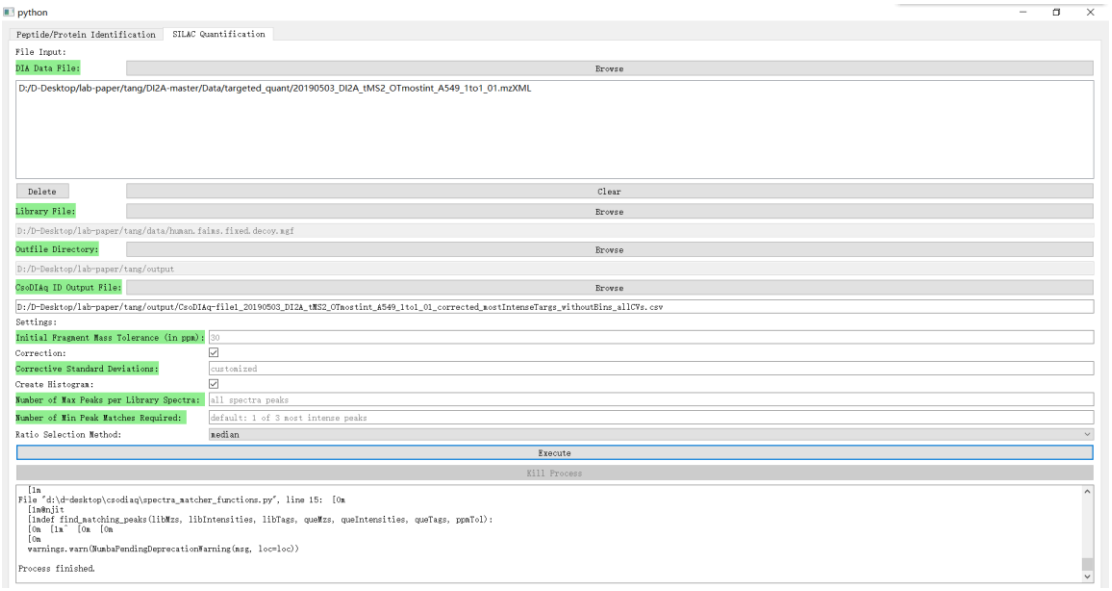


得到运行结果如下

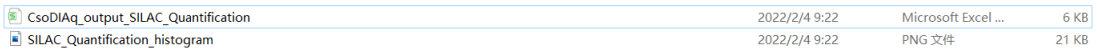


其中 peptideFDR 结尾的文件代表的是定性结果。

之后进行肽段的定量。



产生结果如下：



其中 Quantification 结尾的文件是定量结果。

之后又尝试使用 csodiaq 执行无标记数据，发现只能运行肽段定性这一部分的程序，可以得到肽段定性结果，但是在执行肽段定量部分程序时，软件会出现错误，可知 csodiaq 只可以用于处理有标记样本的定量实验。

在熟悉了 csodiaq 软件的使用方法后，我在服务器上下载了唐师兄 RI-FIGS 的代码，并配置好了实验环境，但是目前还没来得及对唐师兄的实验结果进行复现，这一部分工作会在之后进行。

工作计划

之后我会对唐师兄的代码进行仔细阅读，掌握其中所用到的方法，并结合唐师兄的论文对唐师兄的实验进行复现。