工作总结与计划（7 月 26 日到 8 月 8 日）

代驾

dg310012@mail.ustc.edu.cn

2021-08-08

目录

文档简介

本文档主要分为两个部分。

第一部分是对最近两周的工作进行总结，最近两周我主要做了以下几个工作：

1、阅读了唐师兄最近整理的蛋白质组学总结文档。

2、完成了 FIGS 投稿论文剩余的实验复现。

第二部分是对未来工作计划的一个安排。

工作总结

阅读论文

最近这段时间阅读了一下之前唐师兄发给我的蛋白质组学总结文档，文档中详细介绍了有关蛋白质定性定量实验的预备知识和主要内容。

首先学习了一些基础生物知识，在使用蛋白酶对蛋白质水解后，生成的产物被称作肽段，带电的多肽被称为母离子，母离子进一步碎裂后就生成碎片离子，根据不同的碎裂方式会生成不同类型的子离子，这些子离子一般分为 a,b,c,x,y,z 六种类型。

之后学习了有关数据采集仪器的知识，蛋白质组学的研究离不开质谱仪的使用，质谱仪是一个用于测定离子质荷比的仪器，它将带电离子根据其荷质比将其分离，以便于记录各种离子的质荷比和丰度信息。实验中还经常将相同或者不同的质谱仪串联起来，实现串联或者并联工作，来实现不同质谱仪性能之间的优势互补。除了使用质谱仪来分离混合物外，色谱分离技术也是一种常用手段，由于不同物质的理化性质存在差异，它们在色谱仪的保留时间也不相同，根据不同的保留时间可以对混合样品进行分类。在蛋白质组学研究中，往往将色谱仪与质谱仪进行联机使用，混合物先经过色谱仪分离后，再进入到质谱仪中来分析物质的结构。目前质谱数据的数据格式一般是 mzML 格式。

之后又学习了仪器对样品的各种采集方式，包括全扫描模式，DDA，DIA 和靶向采集方式。其中最常用的是 DDA 和 DIA，DDA 是使用质谱仪器进行全扫描，对从全扫描质谱中选择的母离子进行二级扫描，这种方法需要对目标母离子进行筛选，因此采集信息覆盖率较低，采样不足的情况。而 DIA 是将质谱仪的全扫描范围分为若干个窗口，高速、循环的对每个窗口中的所有离子进行检测，因此可以利用到所有的碎片信息，数据利用率大大提升。

之后学习了数据分析的流程，主要包括预处理/校正，搜库打分，质量控制和蛋白质肽段定量。这个流程我在具体实验中已经复现过了，有关质谱数据的校正主要是关于 mass error 的校正，mass error 一般指实验测得的峰和理论计算的峰之间的荷质比差异。搜库打分是指根据实验图谱和库图谱的相似性来确定对应肽段是否存在，根据分数的高低来确定肽段。由于打分函数具有一定的主观性，因此需要评估定性的肽段，对定性结果的错误数量进行有效的控制，这被称作是质量控制，常用的质量控制方法有 TDA，经验贝叶斯方法和 Entrapment。蛋白质肽段定量包括相对定量和绝对定量，相对定量是研究蛋白质组在不同情况下表达的差异，而绝对定量是获得蛋白质的特定表达水平，可以利用已知含量的参照物和

相对定量值得到绝对定量值。

最后学习了一些统计学相关的知识，主要包括 P-value，E-value，q-value，皮尔森相关系数，斯皮尔曼相关系数和肯德尔相关系数等概念的学习。

实验运行

首先对常规 DIA 实验中的混合样品实验进行了复现：

为了评估 FIGS 在混合样品的常规 DIA 质谱数据上的性能，实验选取 HYE124 质谱数据集进行验证。数据集是在 2 种蛋白质样品 A 和 B 上各进行 3 次重复 SWATH-DIA 实验得到，它们是由人类、酵母和大肠杆菌蛋白质的胰蛋白酶消化物用特定的比例混合而成的。尽管单个肽段和蛋白质的绝对量未知，但是由于 A 和 B 样品的混合比例都是已知的，所以各个肽段在 A 和 B 样品中的丰度比率是一定的。

首先分析了 FIGS 定量的高质量肽段在 A 和 B 样品的相对丰度对数比率分布。

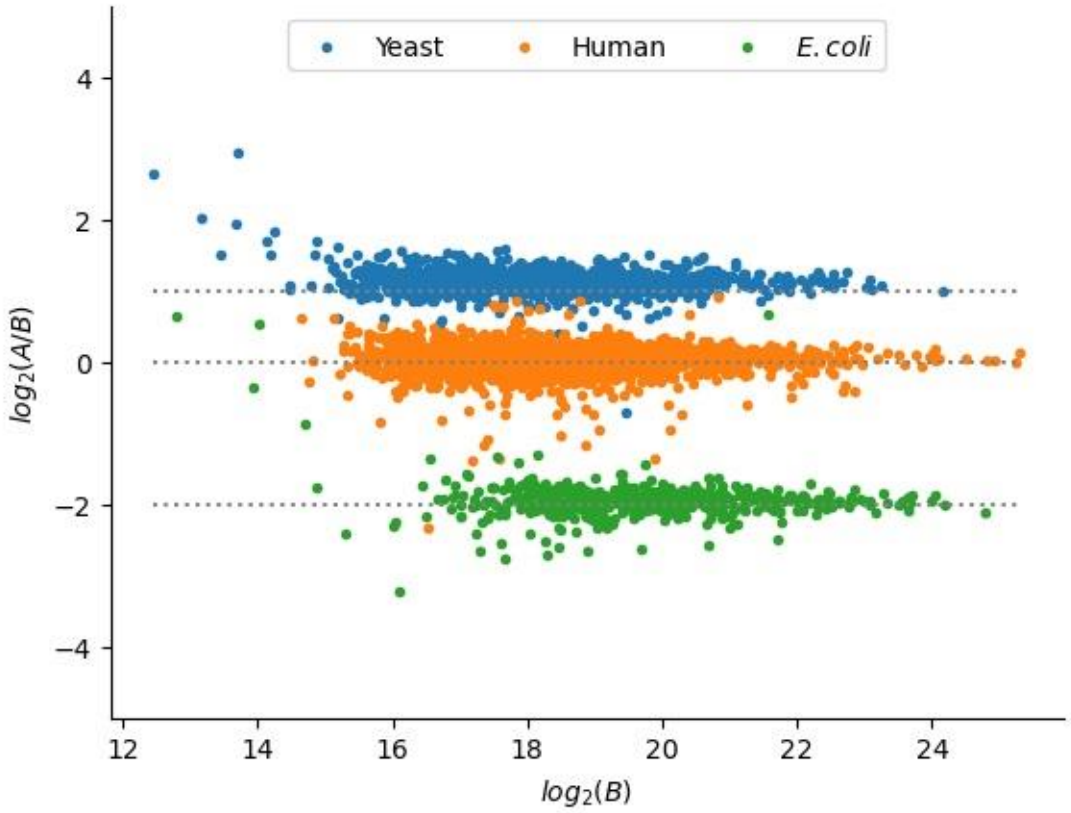

图 1：FIGS A/B 肽段丰度比率分布

从图中可以看出每个物种的肽段都比较集中的分布在理论丰度比率线附近，可见 FIGS 对于混合样品肽段的定量分析符合预期的相对关系。

为了进一步评估 FIGS 对混合样品的定量准确性，**还使用了其他六种 DIA 分析方法（Skyline、OpenSWATH、Spectronaut、DIA-Umpire、PeakView 和 Specter）对 HYE124 质谱数据集进行定性和定量分析。**

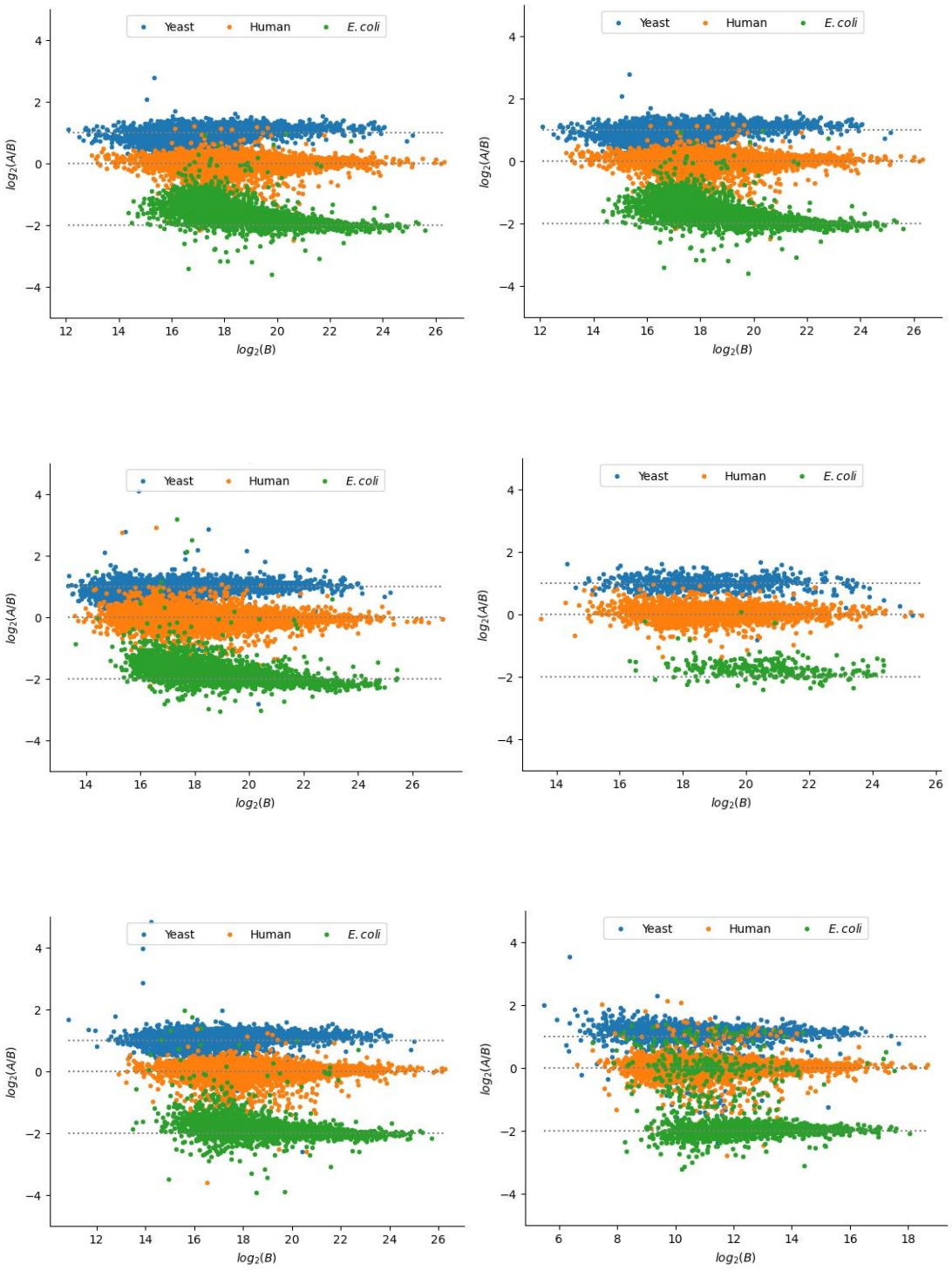

图 2：Skyline、OpenSWATH、Spectronaut、DIA-Umpire、PeakView 和 Specter A/B 肽段丰度比率分布（对应上图顺序为从左到右、从上到下）

对应 FIGS 的 A/B 肽段丰度比率分布图可以粗略看出，FIGS 对混合样品的定性及定量更加准确。

为了量化所得肽段丰度与预期比率的整体差距以及各个物种肽段在 A 和 B 样品中的相关

性，将每种方法分析得到的高质量肽段按照肽段物种以及在 B 样品中相对丰度的三分位数分为 9 组，**统计了各组肽段定量绝对中位差以及相关系数。**

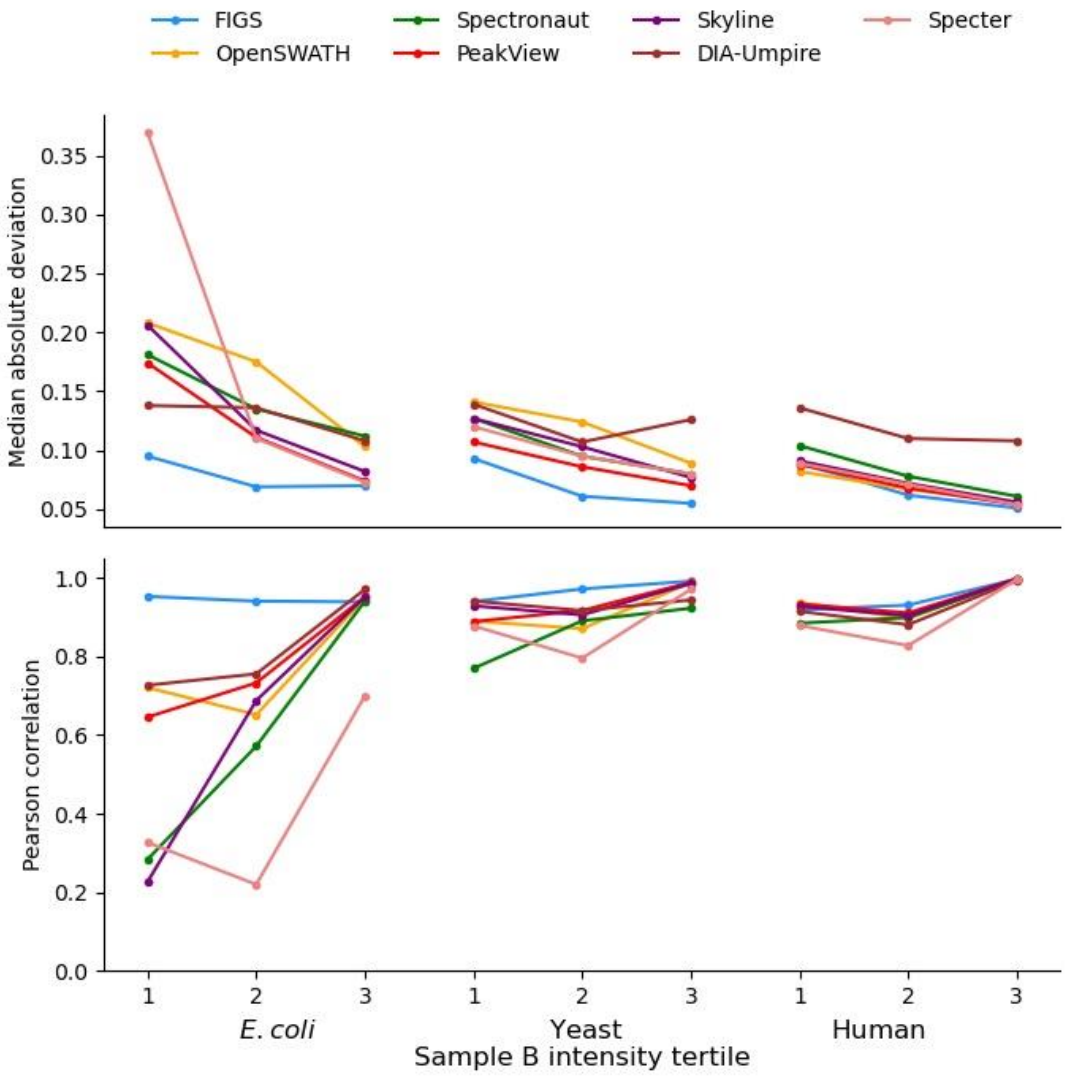

图 3：常规 DIA 实验各组肽段定量绝对中位差及相关系数

可以看出 FIGS 实验结果的绝对中位差相比于其他六组普遍较低，并且 FIGS 实验的相关系数始终保持在 0.9 以上，可见 FIGS 在混合样品实验的定性定量性能更优。

为了验证 FIGS 在图谱库不完备的情况下的鲁棒性，**将 HYE124 图谱库的规模随机减少到原来的 50%，进行重复试验。**

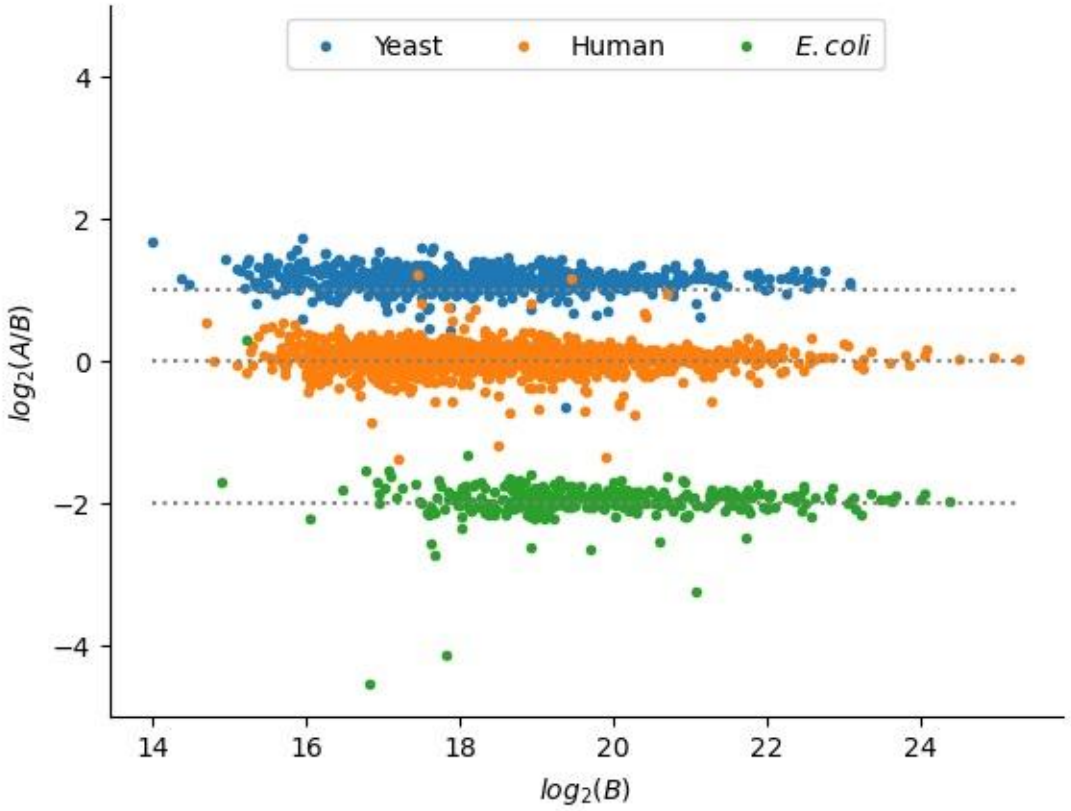

图 4：FIGS A/B 肽段丰度比率分布（参考图谱去除 50%）

从图中可以看出和与使用原始图谱库定量的结果相当，这说明 FIGS 在图谱库不完备的情况下依然具有较强的鲁棒性。

之后进行了质量控制分析实验：

在单一样品实验中，通过双物种图谱方法分别评估了 Specter 和 FIGS 的错误发现性能，用 HEK293T 图谱库作为正库，大肠杆菌图谱库作为反库，**分别统计正库和反库鉴定出的母离子数量：**

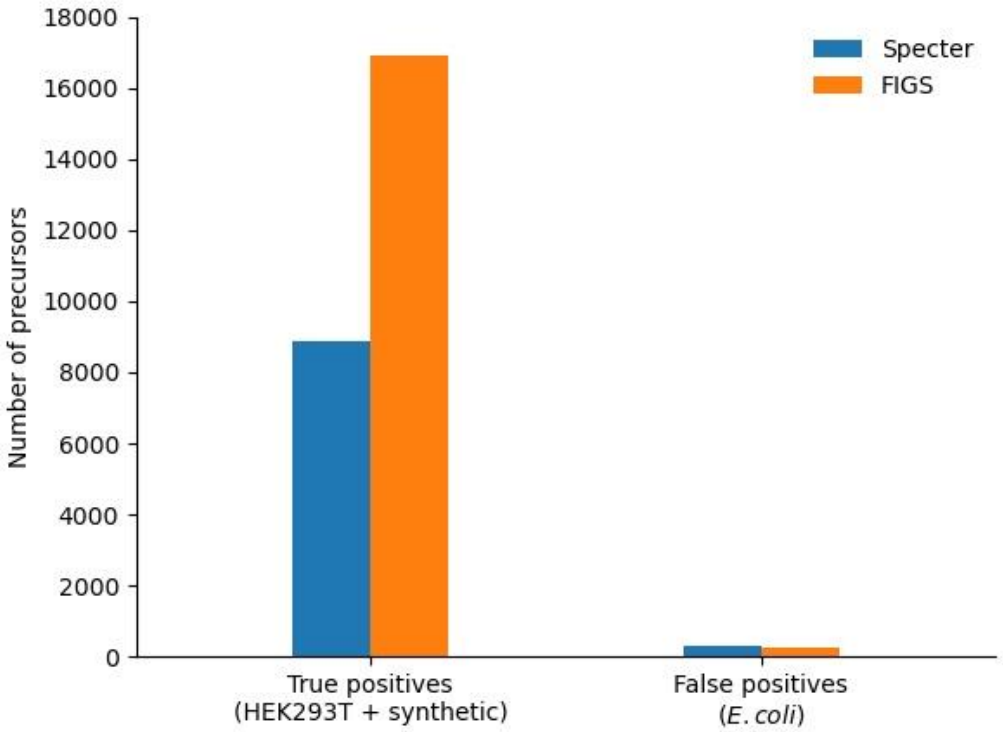

图 5：Specter 和 FIGS 正反库定量结果

可以看出 FIGS 相较于 Specter 可以定量出更多的正库母离子，而作为反库的大肠杆菌母离子在两种方法中鉴定数量都比较低。

利用线性判别方法可以排除正库鉴定的低分段肽段母离子数量来进一步降低错误发现率，**使用 LDA 模型应用到 FIGS 实验生成的正反库结果中：**

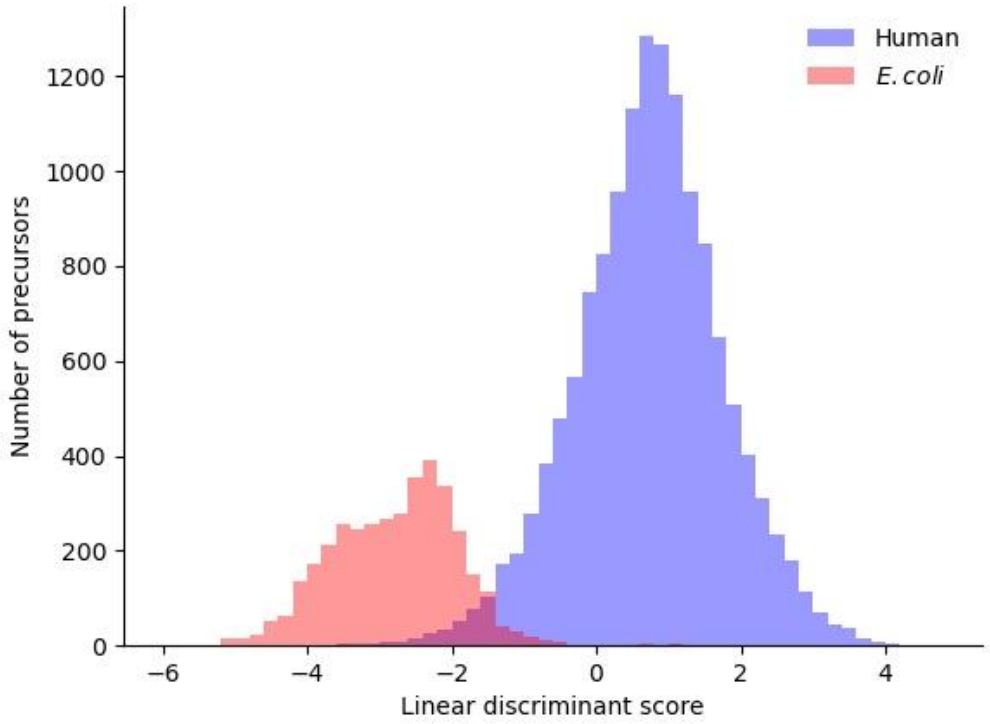

图 6：FIGS 正反库定量肽段打分情况

LDA 模型可以显著区分正库和反库样本，只需要排除线性判别分析分数最低的一部分正库母离子就可以将错误发现率降低到 1%以下。

最后进行了一些补充实验：

在单一样品实验中，对特征离子和图谱库所有离子的归一化强度分布进行了分析：

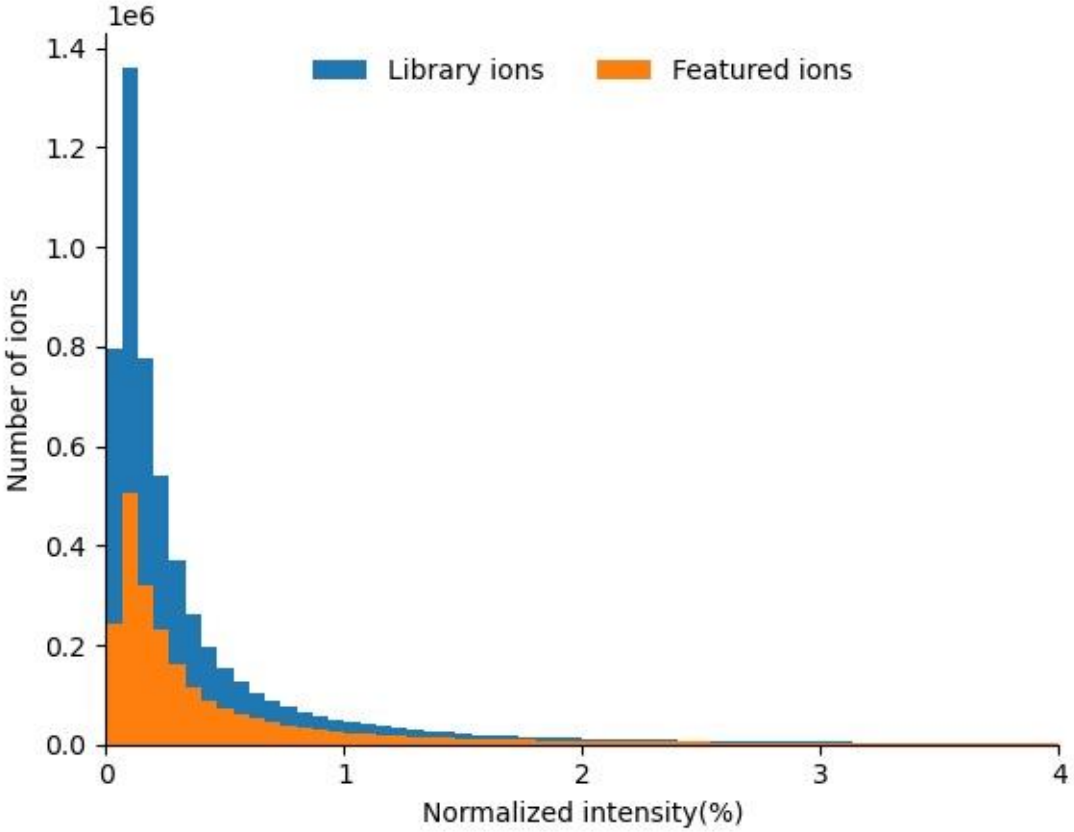

图 7：特征离子与图谱库离子强度分布

从图中可以看出特征离子与图谱库离子在强度分布上是一致的，具有无偏向性。

之后统计了特征离子的荷质比分布：

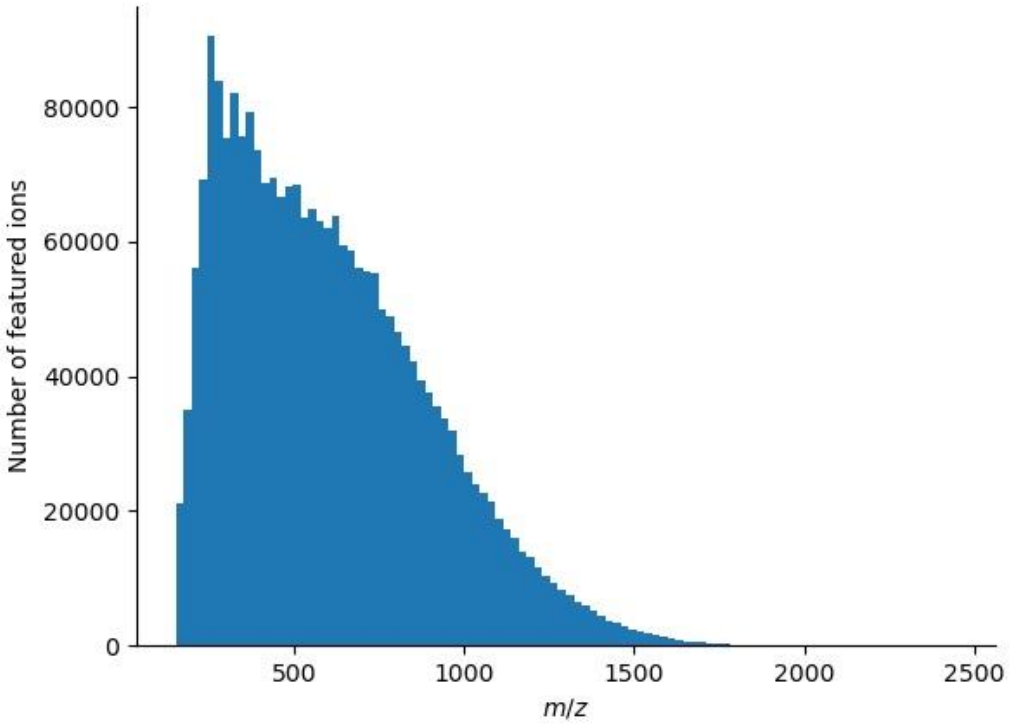

图 8：特征离子质荷比分布

之后分析了用于计算每个系数的特征离子数量的分布：

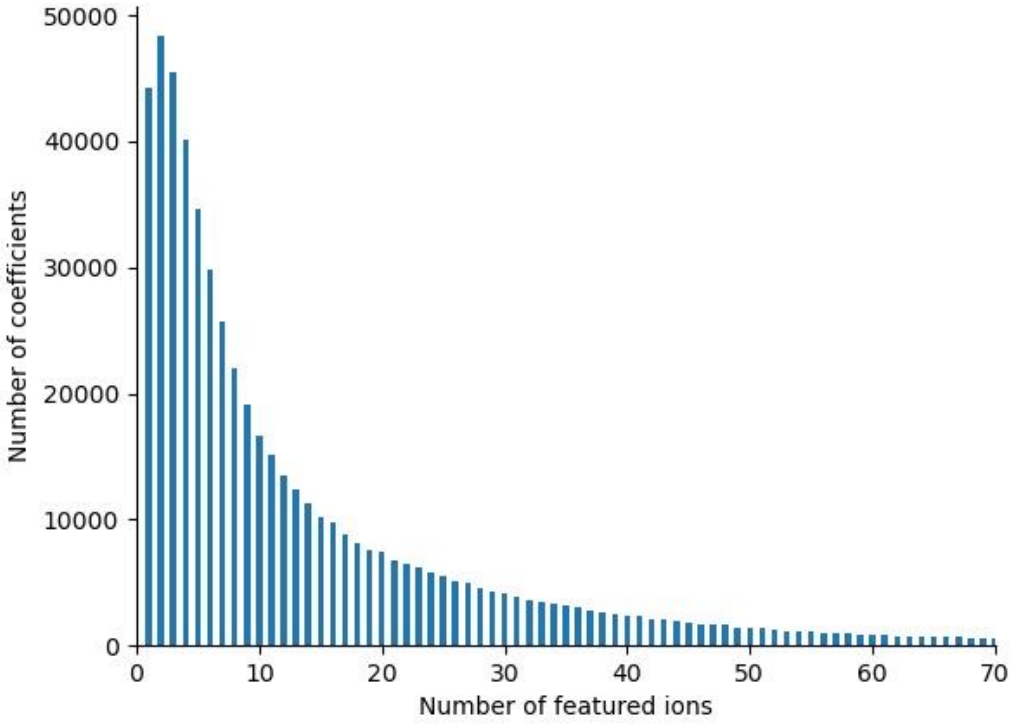

图 9：系数与特征离子数量分布

在单一样品实验中，为了比较 Specter 和 FIGS 对相同 DIA 质谱数据进行图谱分解的准确度，**计算了每个特定肽段母离子数量下所有被分解图谱的平均相关系数。**

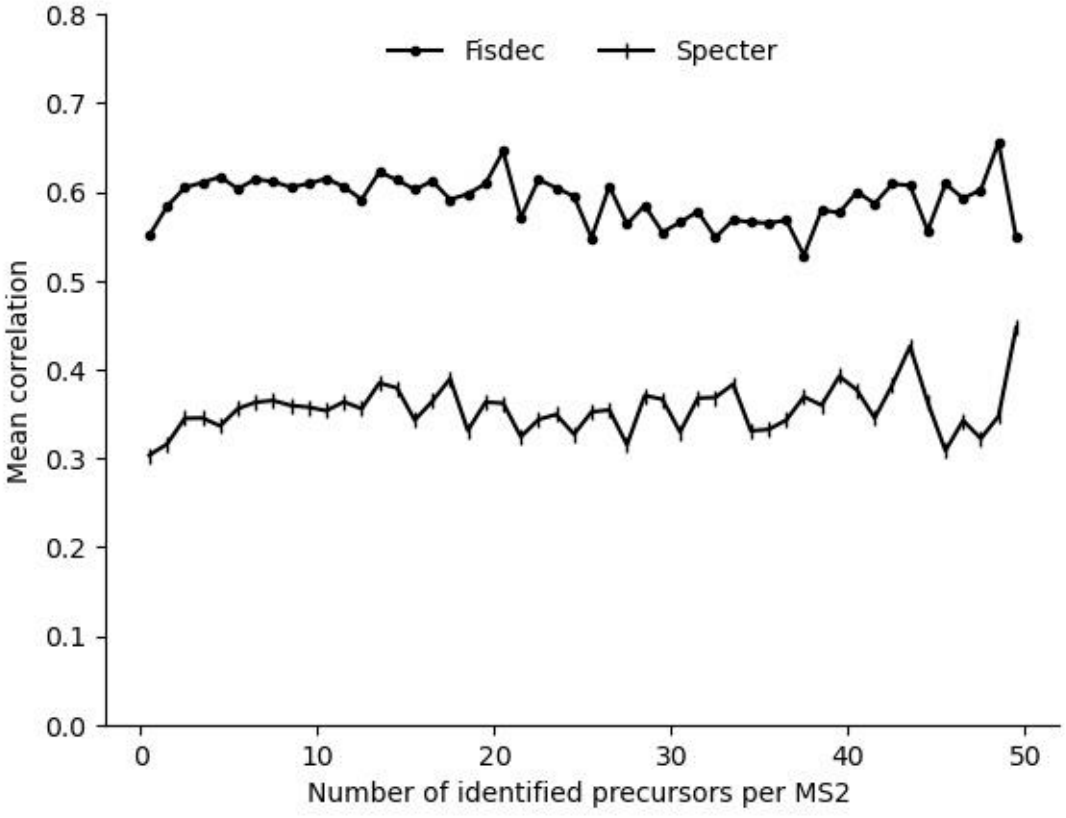

图 10：图谱分解准确率

可以看出，相较于 Specter，FIGS 显著提升了图谱分解的准确率。

工作计划

唐师兄近期又给我发了有关dia-nn的论文,dia-nn是目前开源的定性定量最好的软件，唐师兄让我学习一下这方面的内容并学会初步使用这个工具。我准备先做一下这个事情，做完之后再和唐师兄讨论下一步的工作内容。