工作总结与计划（9 月 29 日到 10 月 26 日）

代驾

dg310012@mail.ustc.edu.cn

2021-10-26

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，最近我阅读了有关 entrapment 的几篇论文，之后尝试对图谱分解过程中的特征离子峰进行筛选以提升特征峰匹配的线性相关度和最终定量的准确性。

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

根据唐师兄的安排，这段时间我看了两篇关于 entrapment 的论文，entrapment 是 target-decoy 方法的一个补充，主要用在对肽段进行质量控制这一步上。

在质量控制模块中，碎片光谱与源自目标数据库的理论光谱相匹配，目标数据库由来自分析生物的肽组成。实验和理论光谱之间的匹配结果表示为肽谱匹配 (PSM)，评分函数为每个 PSM 分配一个分数，表明匹配的质量。观察到的光谱与肽序列之间匹配的质量通过评分函数来量化，该函数极大地影响了蛋白质组学分析的最终结果。因此，用于评估给定评分函数质量的有效统计方法极为重要。

一个好的评分函数首先应该具有判别性，这意味着它可以成功地将正确的 PSM 与不正确的 PSM 分开。一般而言，搜索引擎会识别出最能解释每个观察到的光谱的单个目标肽，即得分最高的 PSM。然而，其中一些假设是不正确的，通常是因为给定的谱图并非来自数据库中的肽段。给定大量光谱，高度区分的评分函数将为正确的 PSM 分配比不正确的 PSM 更高的分数。

其次，评分函数应该可以被很好地校准，这意味着评分具有明确的定义。校准良好的评分函数允许研究人员设计后续实验，准确估计假阳性识别的概率。相反，校准不当的分数可能会导致过度乐观或保守的结论。

目标诱饵搜索策略已成为控制肽和蛋白质识别中错误识别的最流行策略，该方法基于目标数据库中的错误识别数量等于诱饵数据库中的错误识别数量这一假设。虽然此策略可以估计数据集中的错误发现率 (FDR)，但它不能直接评估目标识别中的误报匹配。

Entrapment 论文专注于对评分函数进行校准和评估，它将诱捕序列引入到目标诱饵搜索策略。诱捕序列代表那些极不可能在样品中发现的蛋白质，例如那些从进化上遥远的生物体或样品序列的改组版本获得的蛋白质。将大量的标记诱捕序列与少量的已知样本序列结合起来构建目标数据库进行搜索，诱捕命中可被视为假阳性结果，因此可以使用诱捕命中计算错误(FMR) 来评估结果的质量，并将结果与我们想要测试校准的报告统计分数进行比较，从而给定评分函数进行评估和校准。

实验设计

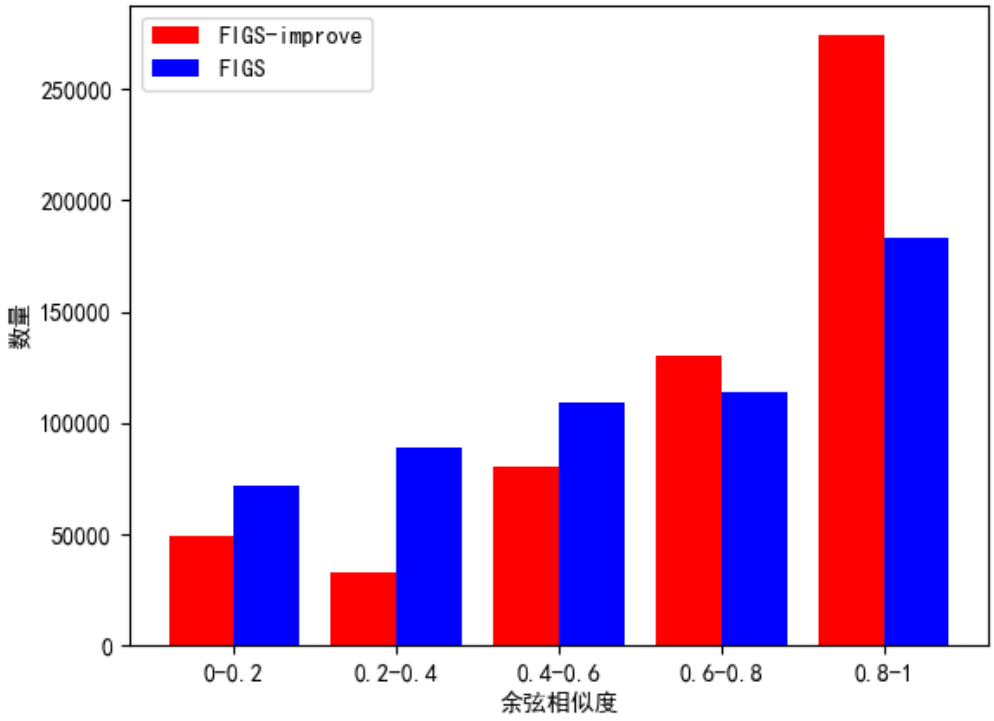
近期，我尝试对图谱分解过程中的特征离子峰进行了筛选以提升特征峰匹配的线性相关度。

在 FIGS 图谱分解模型中，对于每个从给定 DIA 图谱中鉴定的母离子，通过分析其所有特征离子峰强度以及 DIA 图谱中相应的离子峰强度得到其系数。但是特征离子峰对应的 DIA 图谱离子峰强度存在误差，使得特征峰匹配的整体线性相关度仍然存在较大的提升空间，误差来源可以分为两个部分：第一部分是特征离子峰对应的 DIA 图谱离子峰强度存在随机误差，这一误差可能会导致定量的系数偏大，也可能会导致其偏小；第二部分是由于图谱库的不完备性，使得某些离子峰被错误地认定为是某母离子 A 的特征离子峰，而实际上这些离子峰是由母离子 A 和母离子 B 共享的，只是母离子 B 并不在图谱库中。因此母离子 B 对这些离子峰的强度贡献将被划分到母离子 A 中，这会导致测定的定量系数偏大。近期，我主要针对第二部分误差尝试进行了修正，使得优化后的 FIGS 特征峰匹配的线性相关度得到了提升，同时减小了图谱库不完备性对定量结果的影响。

对于每次解谱过程中鉴定出的母离子，可以得到该母离子的所有特征峰离子强度（M1,M2,M3,...）和 DIA 图谱中对应的离子峰强度（N1,N2,N3,...），在理想情况下 N1/M1=N2/M2=N3/M3=...=coeff，这表明这些离子峰是由该母离子独享的，且 DIA 图谱特征离子峰强度不存在误差。但是，在实际情况下，由于图谱的不完备性，某些 DIA 图谱的离子峰并不是该母离子独享的，而是由该母离子和另一母离子（该母离子并不在图谱库中）混合得到的，这会导致该离子峰的 N/M 值明显大于其他离子峰的 N/M 值。在具体代码实现上，可以对所有特征离子峰的 N/M 值进行排序，计算该组数据的中位数来代表该组特征峰的系数值，考察每个特征离子峰的 N/M 值与中位数的比值，设定一个阈值，当比值大于这个阈值时，认定该离子峰是该母离子与其他母离子共享的离子峰，之后再将这样的离子峰排除出去并计算该母离子的系数。

实验结果

首先考察图谱分解过程中特征峰匹配的余弦相似度分布，以 6.75ng 样品首次实验为例，其余弦相似度各个分布区间的数量分布图如下所示：



从图中可以看出在经过特征峰筛选后，特征峰匹配的余弦相似度得到了明显提升。原 FIGS 特征峰匹配的余弦相似度的均值为 0.59，而改进后的 FIGS 特征峰匹配的余弦相似度的均值为 0.70，余弦相似度高于 0.8 的匹配比例也从 32%提升到了 48%。

之后考察了修正后的 FIGS 在单一样品的常规 DIA 质谱数据上的性能，修正后的 FIGS 对 HEK293T 质谱数据集的 15 次实验可定量 13025 至 14222 个母离子，平均每次可定量 13517 个母离子；而原始 FIGS 程序对 HEK293T 质谱数据集的 15 次实验可定量 13021 至 14229 个母离子，平均每次可定量 13519 个母离子。从母离子定量数量这一指标来看，修正后的 FIGS 与原始 FIGS 程序基本持平，这是因为改进后的 FIGS 只是对每一个图谱中母离子定量的系数进行了修正，并没有直接改变母离子的定性结果。

为了验证修正后的 FIGS 减弱了图谱不完备对定量结果带来的影响，将图谱库随机减小到原来的一半，重复单一样品实验。由于此时图谱库是不完备的，因此在解谱过程中，必定存在着多个母离子的共享离子峰被错误地认定为了是某母离子的特征峰，从而使得该母离子

定量结果明显偏大。考察图谱库减小前后同一母离子的定量值的比值，修正后的 FIGS 对 HEK293T 质谱数据集的 15 次前后实验的比值为 0.721 到 0.729，平均值为 0.726，而原始 FIGS 对 HEK293T 质谱数据集的 15 次前后实验的比值为 0.707 到 0.717，平均值为 0.714。这说明修正后的 FIGS 在图谱不完备问题上的鲁棒性得到了提升，图谱不完备性对定量结果带来的影响被减弱，定量结果也会和完备图谱下的定量结果更加接近。

工作计划

有关特征峰筛选这一问题，我觉得应该还有更好的算法去优化。但是在和唐师兄讨论之后，我想先把方师兄论文中有关 RTfree 的实验给复现一下。因为 FIGS 程序本身没有用到色谱信息，所以它在处理 RTfree 数据上应当更具优势，后续可能还会针对 RTfree 数据对 FIGS 做进一步的修正和改进。