工作总结与计划（8 月 9 日到 9 月 5 日）

代驾

dg310012@mail.ustc.edu.cn

2021-09-05

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，最近我主要做了以下几个工作：

1、结合之前阅读的论文和工作思考了有关 FIGS 的两个问题。

2、阅读了 Dia-NN 论文和 SWATH-MS 教程论文。

3、学习了如何使用 Dia-NN 软件。

第二部分是对未来工作计划的一个安排。

工作总结

思考题

1、FIGS 为什么在处理低丰度问题上比较好？

这个问题方师兄在毕业论文中给出过一个解释：在混合样品实验中，对比多个 DIA 分析方法在 HYE124 质谱数据集的实验结果， FIGS 对低丰度肽段的鉴定和定量具有更大优势。究其原因，这些 DIA 分析方法均面向经过色谱洗脱的质谱数据，色谱洗脱提供的保留时间信息对其至关重要，通常需要提取碎片离子或者肽段母离子的色谱曲线，而低丰度肽段母离子的性质可能导致其部分信号丢失和色谱峰不完整等问题，从而使得丰度比率与预期值存在显著偏差。相反， FIGS 侧重于利用质谱数据的质荷比维度进行肽段定量分析，这些问题产生的影响十分有限，因此 FIGS 在不同分位组中的定量结果均具有更高的准确度和可重复性。

关于这个问题我自己也有一些理解，对比 FIGS 和 Specter 的实验结果可以看出，FIGS 在处理肽段定量问题特别是低丰度肽段定量问题上是要优于 Specter 的。这两者的区别在于，FIGS 在计算质谱数据与已匹配图谱库数据的系数 coeff 时，是将该图谱的特征峰和质谱数据中相匹配的离子峰提取出来，按质荷比对齐后利用线性回归得到相应肽段母离子的系数，不同的肽段母离子系数的计算过程都是相互独立的。而 Specter 是利用参考图谱线性组合与 DIA 图谱的整体拟合误差最小化来计算 coeff，每个肽段母离子的系数会受到其他母离子的影响，因此容易出现某些肽段母离子系数绝对误差显著高于其他肽段母离子系数绝对误差的情况，在绝对误差相等的情况下，低丰度肽段的相对误差显然是要大于高丰度肽段的，所以就很有可能出现某些低丰度肽段定量值测量非常不准的情况。而对于 FIGS 来说，某个肽段母离子的系数仅仅和该图谱特征峰与质谱数据对应峰有关，该肽段母离子系数的误差只会来源于特征峰处的噪声，不会受到其他位置噪声的影响，考虑到噪声的出现具有随机性，所以所有肽段的绝对误差应当是相差不多的，不会出现某些肽段绝对误差非常大的情况，因此避免了低丰度肽段测量非常不准的情况。

2、FIGS 的缺陷

在方师兄的论文中，分析了肽段母离子的特征峰与 DIA 图谱中对应离子峰的线性关系，具体分析方式是通过计算两者的余弦相似度来评估两者的线性相关度。以 6.75ng 样品首次

DIA 重复实验为例，在有效特征峰匹配中，余弦相似度高于 0.8 的特征峰匹配比例超过35%，而低于 0.2 的特征峰匹配则有 5.3%。可以看出由于特征峰位置的噪声影响，并不是所有系数对应特征峰匹配的余弦相似度都能保持一个较高的数值。由于 FIGS 只利用特征峰来计算肽段母离子的系数，因此当特征峰匹配的余弦相似度较小时，所求系数的误差也会比较大。因此在求解混合图谱时，可以优先求解特征峰匹配较好的肽段母离子系数，对于那些特征峰匹配较差的肽段母离子可以考虑利用参考图谱线性组合与 DIA 图谱的整体拟合误差最小化来计算系数。

阅读论文

首先阅读了 DIA-NN 的论文。

DIA-NN 是一个易于使用的集成软件套件，它利用深度神经网络来处理 DIA 蛋白质组学实验，并且可以与快速色谱方法结合使用。DIA-NN 工作流程为：提取每个母离子及其碎片离子的色谱图；对假定的洗脱峰进行评分并选出最佳侯选峰；检测并去除潜在的干扰肽；利用 DNN 来计算 q-value；从碎片离子洗脱曲线和定量中去除干扰。

DIA-NN 在每个肽段母离子的假定保留时间附近识别色谱洗脱峰，之后通过一组反应峰特征的分数描述每个洗脱峰，然后使用线性分类器的迭代训练为每个肽段母离子选择最佳侯选峰。由于可能有多个前体共享一个或者多个色谱洗脱峰，所以需要利用以谱为中心的定性方法，为每个频谱选择最佳匹配前体。

在 DIA-NN 中，使用一个前馈、完全连接的 DNN（具有 5 个 tanh 激活层和一个 softmax 输出层），利用交叉熵作为损失函数来区分目标前体和诱饵前体。对于每个前体，将对应于各自洗脱峰的一组分数用作神经网络的输入，神经网络则会输出当前洗脱峰来自于目标前体的可能性，然后对每个前体对神经网络的预测进行平均，从而得到用于 q 值计算的最终分数集。当提供一组峰值分数时，如果连接权重发生变化，输出值也会发生变化，因此可以利用梯度下降型优化算法调整权重使网络成为更好的预测器，更好的区分目标和诱饵。

此外，DIA-NN 包括一种用于检测和去除质谱图干扰的算法。对于每个假定的洗脱峰，DIA-NN 选择受干扰影响最小的片段，并将其洗脱曲线视为肽的真实洗脱曲线的代表。

之后又阅读了一个有关 SWATH-MS 的教程论文。

SWATH-MS 是数据独立采集方法的一种特殊方法，是将深度蛋白质组覆盖能力与定量一致性和准确性相结合的技术。SWATH-MS 将给定样品中落在指定质量范围内的所有多肽以系统和无偏置的方法破碎。为了分析 SWATH-MS 数据，需要建立一种基于肽中心评分的策略。论文中介绍了如何设置 SWATH-MS 实验、如何执行质谱测量以及如何使用肽中心评分分析 SWATH-MS 数据。此外，还讨论了如何改进 SWATH-MS 数据采集的概念、参数设置的潜在权衡和替代的数据分析策略。

SWATH-MS 是一种新兴的策略，对于 SWATH-MS 测量，使用用胰蛋白酶对非标记蛋白质样品进行水解,产生的多肽通过与串联质谱仪在 DIA 模式下进行液相色谱分析。在这种模式下，特定质量范围内的给定样品的所有电离化合物都以无偏置的方式破碎,这种获取方案将导致在前体离子窗口中有许多共洗脱的片段，形成高度复杂的离子光谱，为了处理这种复杂性，使用了一种基于肽中心评分的新的数据分析策略,该策略依赖于以肽查询参数(PQPs)的形式查询感兴趣的蛋白质和肽的色谱和质谱坐标。

论文的主要内容如下：

一、建立和计划 SWATH-MS 实验

SWATH- MS 背后的基本概念是,通过经验推导出的有关肽的质谱和色谱行为的先验知识，可以用于选择性地从高度卷积的 SWATH 数据中以目标方式提取肽特定信息这种需要的先验知识被称为“肽查询参数”(PQPs)。PQPs 通常可以从光谱库中获得，并以表格式存储。PQPs 的信息包括如下：

(1)对给定蛋白质进行监控的肽序列。

(2)肽的主要前体离子以及由此产生的电荷态分布。

(3)在应用片段条件下肽的 4 到 6 个最强烈片段离子。

(4)在应用条件下预期片段模式的信息，即相对片段离子强度。

(5)肽的预期保留时间。

迄今为止，SWATH-MS 研究通过 DDA 分析对感兴趣的样品进行并排表征，以生成光谱库。这些库通常包括 DDA 分析前的样品分离步骤。由于单次 DDA 分析的灵敏度和覆盖率往往低于 SWATH-MS 数据，因此单次 DDA 谱库生成策略不会完全覆盖 SWATH-MS 数据。

目前为蛋白质组学的一般目的而合成蛋白质组的努力已经扩展到非常大的规模,方法是每个蛋白质合成更多数量的肽，并包括 PTMs 和常见序列变异。该方法的一个有用的扩展是

通过重组方法或体外转录/翻译系统来生成全长蛋白。通过这种方式，每个蛋白质最适合分析的多肽可以通过经验确定。

当使用非常大规模的光谱库资源和在许多由 SWATH-MS 测量的样品中查询非常多的肽时，另一个需要考虑的重要问题是适当的错误率控制。从库中查询的肽的很大一部分实际上在可检测的水平上不存在于感兴趣的样品中。如果 SWATH 提取结果是在推断的蛋白水平上进行总结，错误率会因这种假阳性进一步增加。因此，需要在蛋白质水平控制错误发现率。

二、执行 SWATH-MS 测量

以肽为中心的 SWATH-MS 数据分析依赖于在一个肽峰的洗脱轮廓上收集的足够数量的数据点，以允许准确地重建各自的色谱峰。SWATH-MS 方法循环通过前体分离范围所需的 2-4 s 周期时间，可以容纳由更高性能的 LC 设置产生的 2-3 s 宽色谱峰。随着 MS 仪器扫描速度的提高，更窄的色谱峰可能越来越适应于加窗的 SWATH-MS 方法。论文中建议使用 2 小时的纳米高效液相色谱梯度来获取高质量的光谱库，而使用更短的纳米高效液相色谱梯度来获取 SWATH-MS 数据。

在 SWATH-MS 获取方案中前体离子(MS1)和碎片离子(MS2)扫描的最佳设置取决于一系列考虑因素，包括分析物的预期质量范围、潜在的样品复杂性、可用的质谱仪类型及其特定分辨率和扫描速度，LC 的设置及其预期的平均色谱峰宽，以及预期的测量选择性和灵敏度。具体说明如下：

1、SWATH-MS 测量所涵盖的前体质荷比范围由仪器在色谱分离过程中循环通过的相邻一组前体隔离窗口定义。理想情况下，这个质量范围应该尽可能完全地覆盖感兴趣的蛋白质或肽的空间。

2、前体分离窗宽度定义了在给定 MS2 扫描中共分离和共片段的肽前体质量的范围。因此，分离窗的宽度直接影响测量的选择性和动态范围，进而影响肽检测的灵敏度。使用更宽的窗口会导致更多的肽前体被共碎片从而产生更复杂的 MS2 光谱。相反，使用较窄的隔离窗口可以减少共碎片前体的数量和信号干扰，但限制了其他参数，如记录点的数量。因此可以使用最优可变窗口模式，将固定窗口修改为可变窗口。

3、质谱采集或积累时间定义了质谱仪在给定的质谱扫描中积累离子信号的时间。随着积累时间的增加，获得的频谱的信噪比增加。相反，积累时间越长，仪器通过 MS 扫描的周期就越长。因此，积累时间的选择应结合隔离窗口的数量的宽度。

4、色谱循环时间在 SWATH-MS 中对应于 MS1 和 MS2 扫描序列的累积时间之和，周期时间

决定了沿着肽色谱洗脱轮廓同一离子被记录的频率，需要根据平均肽洗脱峰宽度调整色谱循环时间。

5、对于包含 100 个样本的大规模蛋白质组学研究，应当采用每个样本一次注射的方式，以保持样本吞吐量和定量的一致性。然而，对于规模较小的项目，应当对每个样品多次注射。

三、对 SWATH-MS 数据进行肽中心评分

以肽为中心的分析的第一步是定义 PQP（肽查询参数）。根据样品的复杂性，可以为给定的查询肽找到几个高质量的峰组。为了定量描述这种现象的频率，使用了诱饵肽，它有望在色谱和质谱参数空间方面对目标肽的特性进行密切建模，但不应存在于样品中。诱饵 PQP 是从目标肽列表中通过反转或改组氨基酸序列在计算机上生成的。诱饵肽对于控制 SWATH-MS 数据中肽检测和蛋白质推断的错误率至关重要，因为它们可用于估计基础数据中碎片离子色谱图随机共洗脱的影响。

以肽为中心的数据分析策略的下一步是使用定义的 PQP 来提取感兴趣的肽前体和片段离子色谱图。当肽从色谱柱洗脱时，其查询片段离子的信号强度将随着通常的高斯样肽洗脱轮廓同步变化，形成色谱“峰组”，将用于评估肽检测和估计其数量。用同样的方法提取目的肽和诱骗肽的片段离子色谱图。为了大幅度提高以肽为中心的 SWATH-MS 数据分析的选择性，色谱图提取通常不贯穿整个色谱梯度长度，而只在以查询的肽的预期洗脱时间为中心的保留时间窗口内进行。

之后算法定义一个或几个潜在色谱峰组的左右边界，并分别对每个候选肽信号打分。在评分过程中，会计算出各种分数。这些分数会被应用到 SWATH-MS 软件工具 OpenSWATH、Skyline、PeakView 和 Spectronaut 中，这些工具会自动将它们合并成一个最终的鉴别分数。

通过将之前描述的单个分数应用于从 SWATH-MS 数据中提取的目标肽和诱饵肽的离子色谱图，可以计算单个分数分布。理想情况下，在这些分布中，“真阳性”目标和“假阳性”诱饵肽明显可分离。然而，由于实验存在误差，某些分数可能比其他分数表现更好，而单个单独的分数通常没有足够的辨别力来确定检测和正确量化给定的肽。因此，可以用肽查询召回的方式将单个分数组合成单个判别分数。这个目标可以通过半监督学习算法来实现，该算法迭代地收敛于最优化的加权个人分数组合集，以分离目标和诱饵。

SWATH-MS 提供肽水平的定量数据，而肽数量通常是通过对几个片段离子的积分峰面积求和或求平均值来计算的。而为了从 SWATH-MS 测量中推断蛋白质的丰度，需要将一个或多个多肽的测量强度聚集到每个样品的最终蛋白质强度值中。

实验运行

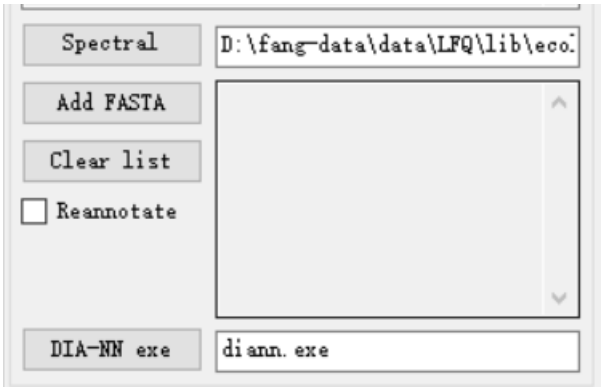
学习并使用了 DiaNN 这个软件。

首先从 https://github.com/vdemichev/DiaNN/releases/tag/1.8 处下载.exe 安装程序并运行它。

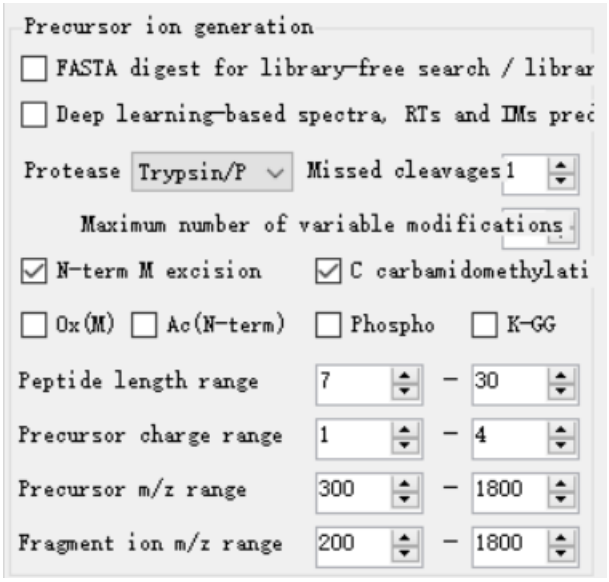
之后使用 DiaNN 执行 LFQbench 实验（样品由人类、大肠杆菌、酵母菌蛋白质混合物组成）。

第一步选择添加原始质谱数据文件：

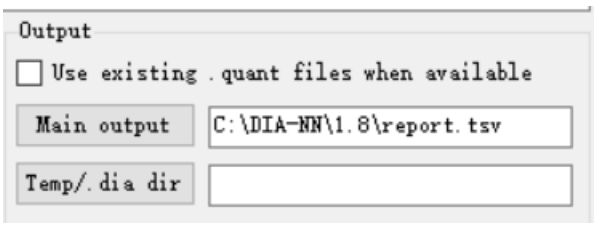
第二步选择添加光谱库：



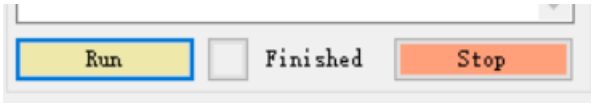
在第二步中也可以选择不添加光谱库进行无库分析，此时需要点击 Precursor ion generation 模块中的 FASTA digest for library-free search/library generation 按钮：



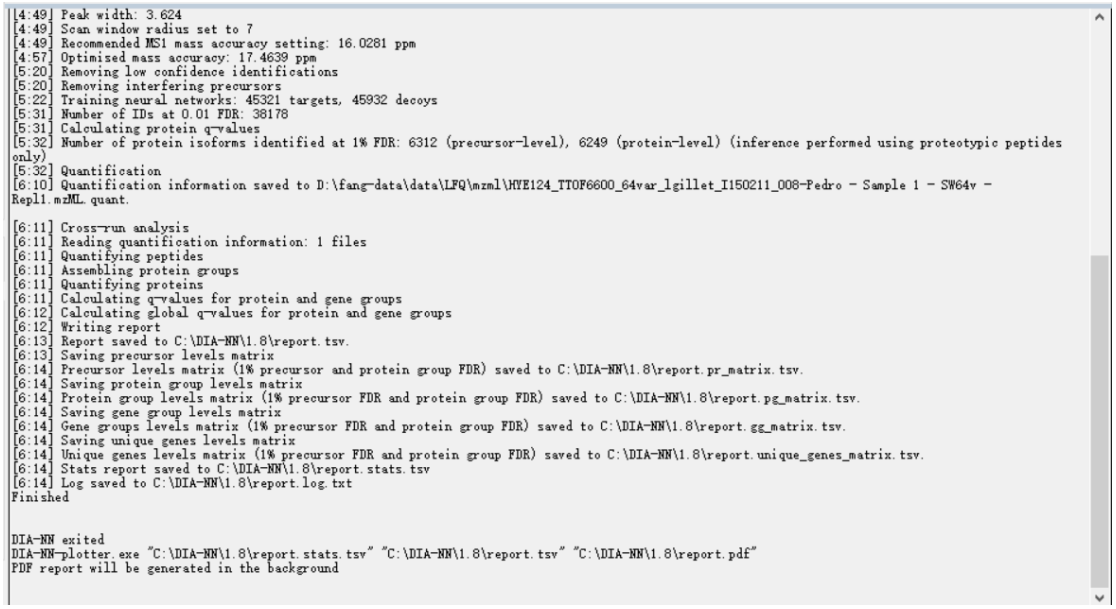
第三步在输出窗格中指定主输出文件名：



第四步点击 Run 按钮：



之后会在输出框中看到输出日志：



```
[4:49] Peak width: 3.624
[4:49] Scan window radius set to 7
[4:49] Recommended MS1 mass accuracy setting: 16.0281 ppm
[4:57] Optimised mass accuracy: 17.4639 ppm
[5:20] Removing low confidence identifications
[5:20] Removing interfering precursors
[5:22] Training neural networks: 45321 targets, 45932 decoys
[5:31] Number of IDs at 0.01 FDR: 38178
[5:31] Calculating protein q-values
[5:32] Number of protein isoforms identified at 1% FDR: 6312 (precursor-level), 6249 (protein-level) (inference performed using proteotypic peptides
only)
[5:32] Quantification
[6:10] Quantification information saved to D:\fang-data\data\LFQ\mzml\HYE124_TTOF6600_64var_lgillet_I150211_008-Pedro - Sample 1 - SW64v -
Repl1.mzML.quant.

[6:11] Cross-run analysis
[6:11] Reading quantification information: 1 files
[6:11] Quantifying peptides
[6:11] Assembling protein groups
[6:11] Quantifying proteins
[6:11] Calculating q-values for protein and gene groups
[6:12] Calculating global q-values for protein and gene groups
[6:12] Writing report
[6:13] Report saved to C:\DIA-NN\1.8\report.tsv.
[6:13] Saving precursor levels matrix
[6:14] Precursor levels matrix (1% precursor and protein group FDR) saved to C:\DIA-NN\1.8\report.pr_matrix.tsv.
[6:14] Saving protein group levels matrix
[6:14] Protein group levels matrix (1% precursor FDR and protein group FDR) saved to C:\DIA-NN\1.8\report.pg_matrix.tsv.
[6:14] Saving gene group levels matrix
[6:14] Gene groups levels matrix (1% precursor FDR and protein group FDR) saved to C:\DIA-NN\1.8\report.gg_matrix.tsv.
[6:14] Saving unique genes levels matrix
[6:14] Unique genes levels matrix (1% precursor FDR and protein group FDR) saved to C:\DIA-NN\1.8\report.unique_genes_matrix.tsv.
[6:14] Stats report saved to C:\DIA-NN\1.8\report.stats.tsv
[6:14] Log saved to C:\DIA-NN\1.8\report.log.txt
Finished


DIA-NN exited
DIA-NN-plotter.exe "C:\DIA-NN\1.8\report.stats.tsv" "C:\DIA-NN\1.8\report.tsv" "C:\DIA-NN\1.8\report.pdf"
PDF report will be generated in the background
```

当程序运行完毕后，会得到相应的输出文件。DIA-NN 生成四种类型的输出文件：主报告、矩阵、“统计”报告（用于质量控制）和 PDF 报告。

主报告中包括 report.tsv（包含所有已识别母离子的列表，以及不同种类的数量、质量指标和注释）、report.pg_matrix.tsv （包含蛋白质组数量）、report.gg_matrix.tsv（基因组数量）、report.pr_matrix.tsv（前体离子数量）。

矩阵表示这些包含蛋白质组、基因组、独特基因（即仅使用蛋白质型，即基因特异性肽）和前体鉴定和量化的基因的标准化数量。

统计报告包括许多可用于数据过滤的 QC 指标，例如排除失败的运行，或作为方法优化的读数。

PDF 报告是基于主报告和统计报告的许多 QC 指标的可视化。

工作计划

和唐师兄交流后，他让再看一下 di-spa 这篇论文，论文的创新点就是优化了采集方式，取消了液相色谱，而采用了补偿电压，然后结合之前学习的所有内容开始思考自己的研究方向。