工作总结与计划（6 月 24 日到 7 月 11 日）

代驾

dg310012@mail.ustc.edu.cn

2021-07-11

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、阅读张师兄和方师兄的毕业论文，学习并理解论文的主要内容和思想。

2、阅读方师兄的部分实验代码，其中包括 FIGS_Deconvolute.py 这个核心程序，该程序的作用是利用图谱库对质谱数据进行解谱。

3、配置实验环境，在实验室服务器和个人电脑上安装所需要的各种依赖包。

4、运行了 FIGS_Deconvolute.py 和 FIGS_Quant.R 这两个程序，得到了共计 15 组质谱数据的蛋白质定性及定量结果。

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

首先阅读了张振航师兄的毕业论文《面向无色谱 DIA 质谱数据的计算系统设计及实现》。该论文主要包括以下内容：

1、基于 LC-MS/MS 流程设计出了无色谱 DIA 质谱数据的计算模型，并给出了无色谱 DIA 数据的计算系统。在蛋白质测定流程中，色谱洗脱环节占据了大部分的时间，该论文中所研究的无色谱 DIA 数据就是去除了色谱洗脱环节的质谱数据。目前的质谱数据分析软件还无法对无色谱 DIA 质谱数据进行处理，该论文开发了一个可以处理无色谱 DIA 质谱的计算系统。

2、基于 RT-free 数据处理系统进行降噪处理和肽谱匹配的改进。RT-free 数据中存在着噪声信号，该论文中通过叠加同一窗口下所有图谱的方法，将满足给定误差范围的信号峰叠加起来，而噪声信号峰则会因为其随机分布的特点被筛选出来。在对窗口中的质谱数据进行解谱时，当不同图谱库中的离子峰出现重叠时，会对定量的准确性带来影响。由于较低质荷比的子离子更容易出现重叠峰，该论文提出了一个仅考虑 300m/z 以上子离子的混合图谱求解方法，并给出了一个适用于该模式下的新的肽段过滤条件，将原来的 TopTenIntensity>5 这一过滤条件修改为 TopTenIntensity>30%。

3、面向特异性定量方法的研究和实现。在一张二级图谱中，某子离子信号峰仅由一个肽段碎裂生成，该信号峰称为特异峰。该论文中提出了如何从一个混合图谱中找到每个理论图谱对应的特异峰，并得出相应理论图谱在混合图谱中的比例系数 Coeff，之后代入定量模块完成定量。

之后阅读了方言师兄的毕业论文《面向 DIA 质谱数据的定量蛋白质组学方法研究》的前 3 章内容。方师兄的论文是在张师兄的基础上具体提出了一种基于特征离子的图谱分解方法（FIGS），并给出了具体的算法流程和实验代码。FIGS 的主要算法流程如下：

1、对数据进行预处理。包括对质谱数据进行筛选和对图谱库数据进行归一化。

2、定性分析。一个肽段母离子被认为是当前质谱数据的离子源需要满足两个条件。一是母离子的质荷比在采集窗口内，二是母离子参考图谱强度前 10 的离子峰中匹配成功的离子峰数超过 5 个。

3、定量分析。根据特征离子确定对应图谱库的系数 coeff，对系数曲线进行修正后确

定定量结果。

4、质量控制。利用正反库的思想，构造一个反库，同时求解正反库的定量结果，用线性判别分析方法对正反库定量结果进行打分，选取高于打分阈值的正库肽段母离子的定量结果输出。

最后阅读了 200311-FIGS-投稿 NBT 论文。该论文的主要内容在方师兄的毕业论文中都有所体现，其各项实验的实验内容和结果在方师兄的毕业论文中也有更加详细的说明。

阅读代码

方师兄的代码主要包括三种类型。一是.py 文件，编程语言是 python，方师兄的实验代码主要也都是 python 程序。二是.ipynb 文件，使用的工具是 jupyter notebook，该应用程序可以用于运行代码，展示结果，文档编写等等，主要用来绘制各种实验图表。三是.R 文件，编程语言是 R 语言。这一部分程序主要用来对解谱后的数据做定量结果分析。我目前只是详细阅读了其中的一部分代码，具体如下：

1、FIGS_deconvolute_std.py。该程序的主要作用是对质谱数据进行解谱，得到当前质谱数据对应的图谱系数。

2、Loadmsms.py。该程序的作用是生成图谱库数据，首先修改了原始文件中每个母离子的 seq 信息，对于 seq 和 charge 相同的母离子，选取其中 score 最高的母离子加入到 newmsms 中，之后基于 newmsms 取出匹配离子峰数量 TopTen>5 的母离子作为输出写入到文件中。

3、LoadmsmsWithProteinGroups.py。该程序与 Loadmsms.py 程序的功能大致相同，区别在于 LoadmsmsWithProteinGroups.py 处理的是人类、酵母、大肠杆菌蛋白质胰蛋白酶消化物混合物数据，需要每个母离子额外添加所属类别信息。

4、MergeSpectra.py。该程序的作用是对同一窗口下的二级图谱进行合并，在合并 m/z 相近的离子峰时可以选择求和或者求平均值。

5、myms.py。该程序提供了一系列函数供其他 python 程序调用，如读取质谱数据、图谱库数据，生成反库，计算图谱库特征离子强度和质谱特征离子强度的相似系数等等。

6、mzmlPreprocess.py。该程序的作用在于将每个二级图谱与同一窗口下的前一张二级图谱进行对比，只保留相同的离子峰。

7、NormByWindow.py。该程序的作用是求解每一个窗口下所有二级图谱的总强度大小。

8、PlotRatios.py。该程序的作用是对人类、酵母、大肠杆菌蛋白质混合物数据得到的定量结果进行分析。绘制了样品 A 和样品 B 的相对丰度比率 log2(A/B)分布图，并按照 B 样品相对丰度的三分位数分组，计算每组肽段的绝对中位差。

9、QuantBySpectrumScore.py。该程序的作用是对肽段进行定量。根据 FIGS_deconvolute.py 程序生成的 csv 文件可以得知每张二级图谱对应的图谱的系数 coeff，选取质谱数据和图谱库数据特征离子相似度最高的三张二级图谱对肽段进行定量，使用积分的方式得到该肽段的定量值。

10、RTfree_Quant.py。该程序的作用是进行 FDR 控制。首先将正库和反库的定量结果

进行合并，之后利用线性判别方法确定分数阈值，将高于阈值的正库肽段母离子的定量结果输出。

11、sparse_nnls.py。该程序提供了一些矩阵处理的基本函数用于其他的程序进行调用。

配置实验环境

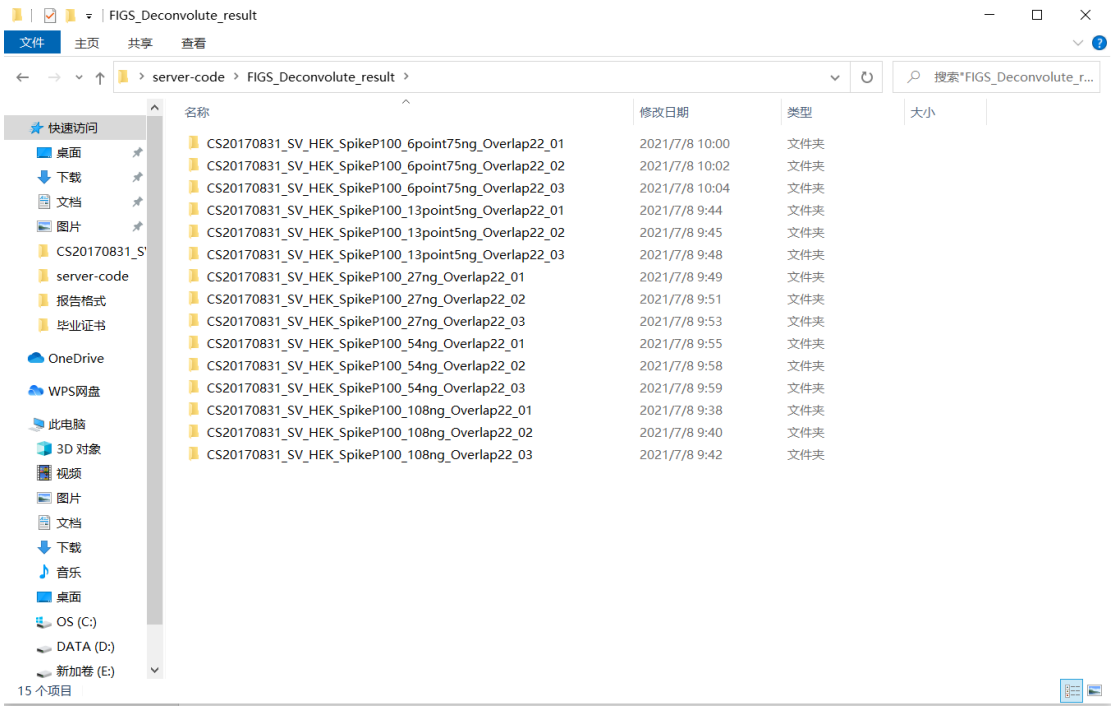
我在自己的电脑和服务器上都对实验环境进行配置，主要是安装了一些 python 的依赖包以及 R 语言环境的安装，依赖包包括 scipy、bidict、collections、pandas、math、numpy、time、tqdm 等等，在具体安装时主要是使用 anaconda 和 pip 进行安装。

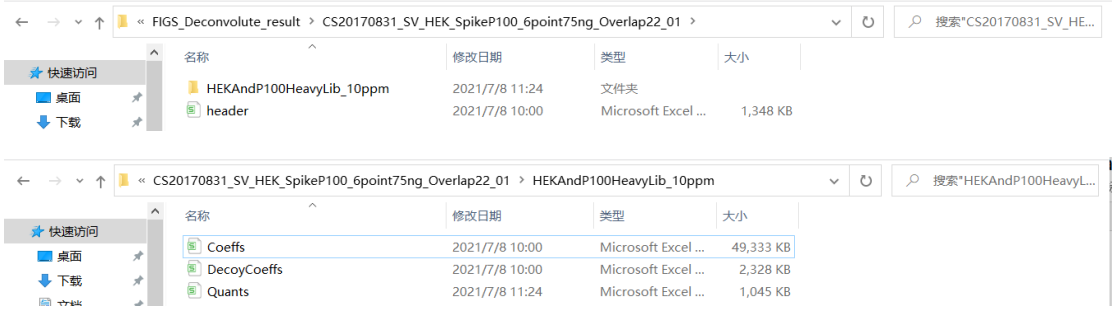
实验运行

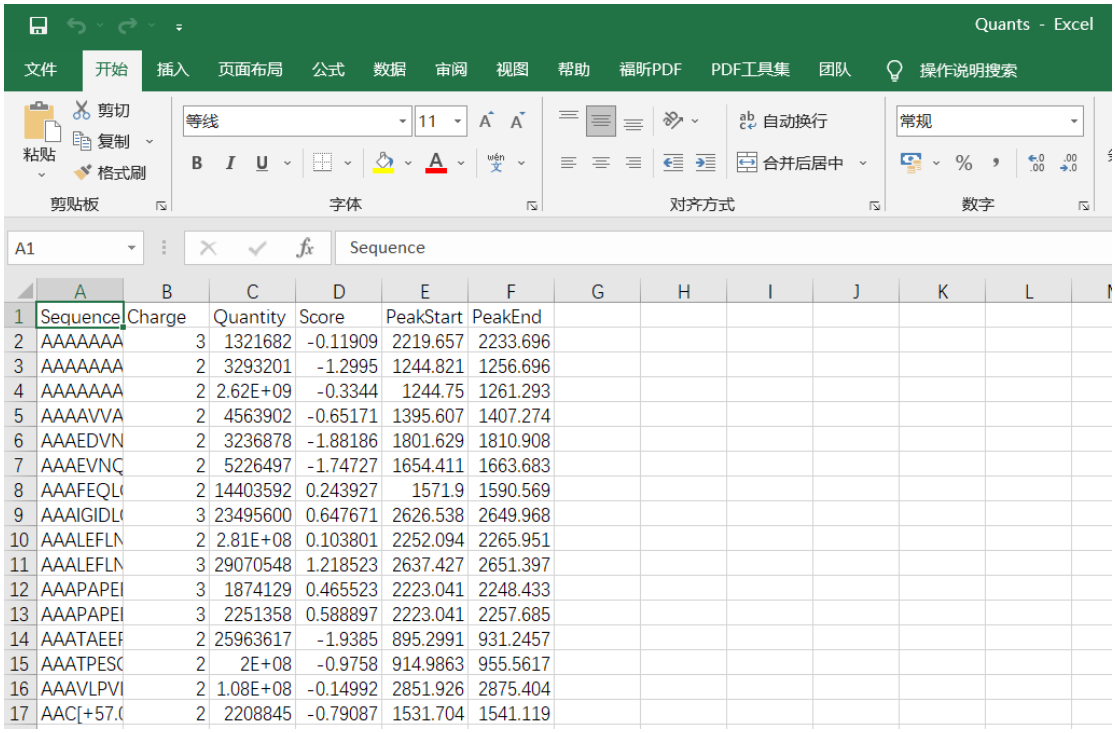
在配置好实验环境后运行了 FIGS_Deconvolute.py 和 FIGS_Quant.R 这两个程序，这两个程序的作用是进行解谱和肽段的定量，并且这里选取的是方师兄比较原始的一个版本，实际上后续的很多程序基本上都是在这两个程序的基础上添加了一些参数而来的。

实验数据选取的是 HEK293T 数据集，五种梯度浓度，每个梯度下三次重复试验，共计 15 组实验数据，实验截图和结果如下：

```
(FIGS) dg@ustc-gpu:~/FIGS/Specter$ cat FIGS_Deconvolute_record.txt
Loading library... Finished
Generating decoy library... Finished
Loading MS2... Finished
Header has written.
Deconvoluting 43900... Finished
Coeffs have written.
Re-deconvoluting 43900... Finished
Decoy coeffs have written.
End.
Loading library... Finished
Generating decoy library... Finished
Loading MS2... Finished
Header has written.
Deconvoluting 44000... Finished
Coeffs have written.
Re-deconvoluting 44000... Finished
Decoy coeffs have written.
End.
Loading library... Finished
Generating decoy library... Finished
Loading MS2... Finished
Header has written.
Deconvoluting 44000... Finished
Coeffs have written.
Re-deconvoluting 44000... Finished
Decoy coeffs have written.
End.
Loading library... Finished
Generating decoy library... Finished
Loading MS2... Finished
Header has written.
Deconvoluting 44000... Finished
Coeffs have written.
Re-deconvoluting 44000... Finished
Decoy coeffs have written.
End.
Loading library... Finished
Generating decoy library... Finished
Loading MS2... Finished
Header has written.
Deconvoluting 44000... Finished
Coeffs have written.
```







	A	B	C	D	E	F
1	Sequence	Charge	Quantity	Score	PeakStart	PeakEnd
2	AAAAAAA	3	1321682	-0.11909	2219.657	2233.696
3	AAAAAAA	2	3293201	-1.2995	1244.821	1256.696
4	AAAAAAA	2	2.62E+09	-0.3344	1244.75	1261.293
5	AAAAVVA	2	4563902	-0.65171	1395.607	1407.274
6	AAAEDVN	2	3236878	-1.88186	1801.629	1810.908
7	AAAEVNC	2	5226497	-1.74727	1654.411	1663.683
8	AAAFEQL	2	14403592	0.243927	1571.9	1590.569
9	AAAIGIDL	3	23495600	0.647671	2626.538	2649.968
10	AAALEFLN	2	2.81E+08	0.103801	2252.094	2265.951
11	AAALEFLN	3	29070548	1.218523	2637.427	2651.397
12	AAAPAPEI	3	1874129	0.465523	2223.041	2248.433
13	AAAPAPEI	3	2251358	0.588897	2223.041	2257.685
14	AAATAEEF	2	25963617	-1.9385	895.2991	931.2457
15	AAATPESC	2	2E+08	-0.9758	914.9863	955.5617
16	AAAVLPVI	2	1.08E+08	-0.14992	2851.926	2875.404
17	AAC[+57.(	2	2208845	-0.79087	1531.704	1541.119

工作计划

我目前已经大致了解了方师兄实验的一些理论基础和主要思想，但是方师兄的代码目前我还没有看完，后面我打算继续阅读方师兄剩下的代码，其中有一个比较关键的程序是用 R 语言编写的，由于我之前没有学过 R 语言，所以可能还要花一点时间把 R 语言的基本语法学习一下。

由于 FIGS 是在 Specter 的基础上修改而来的，方师兄的一些实验其实也涉及到了 Specter 的一些内容，所以我后面准备去看一下 Specter 的相关论文和代码。

因为我已经运行了 FIGS_deconvolute.py 和 FIGS_Quant.R 这两个最基本的程序，也得到了每个质谱数据对应的肽段定量结果，所以方师兄的一些具体实验我也可以开始进行复现了，后面我打算对照着论文的实验内容以及方师兄的具体代码来进行各个实验结果的复现。