

工作总结与计划（7 月 12 日到 7 月 25 日）

代寄

dg310012@mail.ustc.edu.cn

2021-07-25

目录

第一章 文档简介.....	3
第二章 工作总结.....	4
2.1 阅读论文.....	4
2.2 阅读代码.....	5
2.3 实验运行.....	7
第三章 工作计划.....	14

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 Specter 相关的论文。
- 2、阅读方师兄余下的 python 程序和 R 程序。
- 3、完成了常规 DIA 实验分析中单一样品分析这一部分的实验复现。

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

FIGS 是在 Specter 的基础上修改而来的，并且方师兄的很多实验也涉及到了 Specter 论文中的一些内容，所以这段时间我阅读了一下 Specter 的相关论文。

该论文主要是提出了一种利用最小二乘法求解二级图谱的方法，可以识别并得出所采集 MS2 图谱的各种图谱库组合。并针对 Specter 的性能作了一系列的实验验证。其中包括：

1、Specter 具有良好的定性及定量准确性。在具体实验中，将合成磷酸肽以一定的浓度加入到 HEK293T 混合物中，并设置五种浓度梯度进行三次重复实验，分析每个肽段的量化结果。

2、Specter 具有较低的错误发现率。在具体实验中，使用同一仪器上生成的大肠杆菌光谱库作为诱饵（样品中并不存在大肠杆菌蛋白质），计算其对应的错误发现率。

3、Specter 对于图谱的不完整性具有较好的鲁棒性。删去光谱库中的一些图谱后，再对 Specter 运行结果进行测试，发现其结果不会受很大的影响。

4、Specter 可以区分一些高度相似的肽段。在具体实验中，分析并选择了三种合成肽家族，将每个家族的一组随机成员加入到大肠杆菌裂解物消化物中。尽管每个家族中的库光谱十分相似，但是 Specter 仍然可以正确识别出几乎所有的肽段。

5、Specter 可以区分磷酸蛋白质组学数据中的位置异构体。在具体实验中，在 Thermo Q-Exactive Plus HF 上进行了 84 次的 DIA 数据实验分析，其中有 75 次鉴定出了位置异构体。

6、Specter 的实验结果具有高度可重复性。使用 Specter 分析 DIA 数据，并进行多次重复实验，通过 pearson 相关系数量化实验结果后发现 Specter 在 DIA 中的鉴定和定量上具有高度可重复性。

7、Specter 可以用于分析混合样本的定量结果。在具体实验中，是对人类、酵母、大肠杆菌这三个物种的蛋白质混合物样本进行分析，比较得出不同样品浓度下的蛋白质定量关系。

阅读代码

这段时间我先是阅读了方师兄余下的.py文件以及.R文件。具体如下：

1、Specter.py。该程序是 Specter 用于解谱的主要程序，也是 FIGS 的基础代码，Specter.py 与 FIGS 代码有许多相似之处。例如对于质谱数据和图谱库数据相同的读取方式，同样都选择 TopTen>5 作为肽谱匹配的条件，但是不同之处在于 Specter 是利用所有的已匹配离子峰计算质谱数据与所匹配图谱库之间的系数，而 FIGS 则是考虑了特征离子峰这一概念，仅仅考虑每个图谱对应的特征离子峰来计算该图谱在质谱中的权重，不仅缩短了运行时间，而且避免了不同图谱之间重叠离子峰的干扰。

2、SpecterFast_usePickle.py 和 SpecterFast.py。这两个程序和 Specter.py 基本一致，区别在于是否使用了并行方式执行程序以及是否计算反库的解谱结果。

3、UniqueQuan_Coeff+QuantByPredictRT.py。该程序是一个比较完整的蛋白质定量程序，它包括使用 FIGS 方法对二级图谱进行解谱，基于母离子预测保留时间对解谱结果进行定量，将正反库定量结果合并后执行 FDR 控制。

4、UniqueQuant_std_WindowWidth.py。该程序的作用与 FIGS 原始程序基本一致，都是对二级图谱进行解谱，区别在于该程序最终写入文件内容略有不同，该程序在写入 csv 文件时还将特征离子峰位置、强度，强度最低的特异峰排名写入文件。

5、UniqueQuant_std.py。该程序的作用也是二级图谱进行解谱，但是特别的地方在于，在计算质谱数据与已匹配图谱库系数时只选取那些稳定出现的特异峰。

6、wnnls.py。该程序的作用也是二级图谱进行解谱，但是在这个程序考虑到罚项的存在，会对原始的图谱库强度向量和质谱数据强度向量进行修改。

7、以 FIGS_Deconvolute 开头的一系列 python 程序。这些代码都是在 FIGS 程序的基础上修改而来的，只是其中有些细微的差别，比如在解谱时使用了 SN 选择的 3 个峰，在求解系数 coeff 时不使用最小二乘法，而是计算质谱数据特征峰强度之和与已匹配图谱特征峰强度之和的比值。

8、wnnls_Quant.R。该程序的作用是对解谱后的结果做定量分析，在 FIGS 程序对二级图谱进行解谱后，由于同一种肽段母离子通常会出现同一窗口的连续多张二级质谱中，所以对于定性出来的同一肽段往往会有多个系数，每个系数对应一个采集时间，表明该肽段在各个时间的相对丰度，因此可以利用这些系数构建肽段母离子的系数曲线完成肽段定量。具体方法就是在系数曲线上截取一段较为连续的区间求积分来作为该肽段母离子的定量结果。

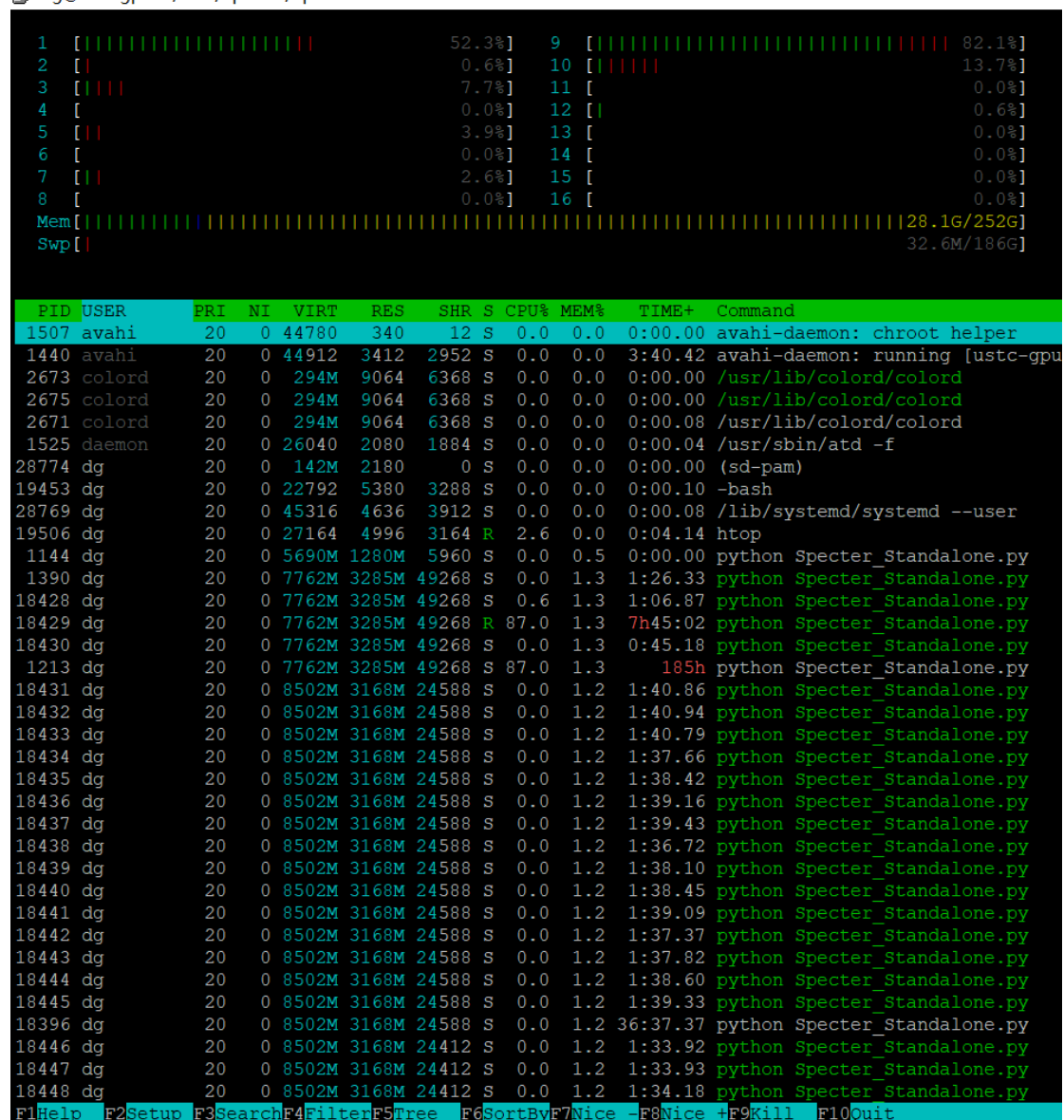
9、SpecterQuant_NoDecoy.R。该程序与 wnnls_Quant.R 程序大体一致，不一样的地方在于该程序只对正库解谱结果进行定量，也没有使用 FDR 控制的方法对定量结果进行筛选。

除了 python 程序和 R 程序以外，方师兄的代码还有一部分是用 jupyter notebook 写的，这部分代码本质上还是 python 程序，主要作用是根据生成好的蛋白质定量结果绘制相应的图表。方师兄论文的实验结果基本上都出自于“数据分析与绘图.ipynb”这个文件，我这段时间已经开始一边阅读这一块的代码一边复现其中的一部分实验。

实验运行

首先对 Specter 程序进行运行，我已经在服务器上配置好了实验环境，目前程序正在运行当中，但是 Specter 程序运行速度较慢，在服务器上跑了很多天以后也只得到了一部分的实验结果。因为后续的实验还要用到 Specter 实验的肽段定量结果，所以这里我先把方师兄跑好的 Specter 实验结果拿来用了。

```
dg@ustc-gpu: ~/FIGS/Specter/SpecterResults
```



PID	USER	PRI	NI	VIRT	RES	SHR	S	CPU%	MEM%	TIME+	Command
1507	avahi	20	0	44780	340	12	S	0.0	0.0	0:00.00	avahi-daemon: chroot helper
1440	avahi	20	0	44912	3412	2952	S	0.0	0.0	3:40.42	avahi-daemon: running [ustc-gpu]
2673	colord	20	0	294M	9064	6368	S	0.0	0.0	0:00.00	/usr/lib/colord/colord
2675	colord	20	0	294M	9064	6368	S	0.0	0.0	0:00.00	/usr/lib/colord/colord
2671	colord	20	0	294M	9064	6368	S	0.0	0.0	0:00.08	/usr/lib/colord/colord
1525	daemon	20	0	26040	2080	1884	S	0.0	0.0	0:00.04	/usr/sbin/atd -f
28774	dg	20	0	142M	2180	0	S	0.0	0.0	0:00.00	(sd-pam)
19453	dg	20	0	22792	5380	3288	S	0.0	0.0	0:00.10	-bash
28769	dg	20	0	45316	4636	3912	S	0.0	0.0	0:00.08	/lib/systemd/systemd --user
19506	dg	20	0	27164	4996	3164	R	2.6	0.0	0:04.14	htop
1144	dg	20	0	5690M	1280M	5960	S	0.0	0.5	0:00.00	python Specter_Standalone.py
1390	dg	20	0	7762M	3285M	49268	S	0.0	1.3	1:26.33	python Specter_Standalone.py
18428	dg	20	0	7762M	3285M	49268	S	0.6	1.3	1:06.87	python Specter_Standalone.py
18429	dg	20	0	7762M	3285M	49268	R	87.0	1.3	7h45:02	python Specter_Standalone.py
18430	dg	20	0	7762M	3285M	49268	S	0.0	1.3	0:45.18	python Specter_Standalone.py
1213	dg	20	0	7762M	3285M	49268	S	87.0	1.3	185h	python Specter_Standalone.py
18431	dg	20	0	8502M	3168M	24588	S	0.0	1.2	1:40.86	python Specter_Standalone.py
18432	dg	20	0	8502M	3168M	24588	S	0.0	1.2	1:40.94	python Specter_Standalone.py
18433	dg	20	0	8502M	3168M	24588	S	0.0	1.2	1:40.79	python Specter_Standalone.py
18434	dg	20	0	8502M	3168M	24588	S	0.0	1.2	1:37.66	python Specter_Standalone.py
18435	dg	20	0	8502M	3168M	24588	S	0.0	1.2	1:38.42	python Specter_Standalone.py
18436	dg	20	0	8502M	3168M	24588	S	0.0	1.2	1:39.16	python Specter_Standalone.py
18437	dg	20	0	8502M	3168M	24588	S	0.0	1.2	1:39.43	python Specter_Standalone.py
18438	dg	20	0	8502M	3168M	24588	S	0.0	1.2	1:36.72	python Specter_Standalone.py
18439	dg	20	0	8502M	3168M	24588	S	0.0	1.2	1:38.10	python Specter_Standalone.py
18440	dg	20	0	8502M	3168M	24588	S	0.0	1.2	1:38.45	python Specter_Standalone.py
18441	dg	20	0	8502M	3168M	24588	S	0.0	1.2	1:39.09	python Specter_Standalone.py
18442	dg	20	0	8502M	3168M	24588	S	0.0	1.2	1:37.37	python Specter_Standalone.py
18443	dg	20	0	8502M	3168M	24588	S	0.0	1.2	1:37.82	python Specter_Standalone.py
18444	dg	20	0	8502M	3168M	24588	S	0.0	1.2	1:38.60	python Specter_Standalone.py
18445	dg	20	0	8502M	3168M	24588	S	0.0	1.2	1:39.33	python Specter_Standalone.py
18396	dg	20	0	8502M	3168M	24588	S	0.0	1.2	36:37.37	python Specter_Standalone.py
18446	dg	20	0	8502M	3168M	24412	S	0.0	1.2	1:33.92	python Specter_Standalone.py
18447	dg	20	0	8502M	3168M	24412	S	0.0	1.2	1:33.93	python Specter_Standalone.py
18448	dg	20	0	8502M	3168M	24412	S	0.0	1.2	1:34.18	python Specter_Standalone.py

F1Help F2Setup F3Search F4Filter F5Tree F6SortBy F7Nice F8Nice F9Kill F10Quit

图 1: 服务器上运行的 Specter 程序

下面对常规 DIA 实验中的单一样品实验进行了复现:

为了评估 FIGS 在单一样品的常规 DIA 质谱数据上的性能, 实验选取 HEK293T 质谱数据

集进行验证。数据集包括五种浓度梯度，每个梯度下三次重复试验，在分别运行了 FIGS 程序和 Specter 程序后即可得到每次实验的肽段定量结果，根据肽段定量结果统计相应的数据并绘制对应的图表来说明各自的性能。

首先对 FIGS 和 Specter 在 HEK293T 质谱数据的定量母离子数量进行了评估：

对于 FIGS 来说，在 15 次实验中，单次实验可以定量 13021 至 14229 个母离子，平均可以定量 13519 个母离子。每对 DIA 质谱数据之间定量的共同母离子数量为 11002 到 12169 个，平均每对实验的共同母离子数量为 11418 个。

```
单个实验定量母离子数量(不含磷酸肽): 13021 14229
平均单个实验定量母离子数量(不含磷酸肽): 13519.466666666667
共同母离子数量: 11002 12169
平均每对实验的共同母离子数量: 11418.095238095239
```

图 2: FIGS 定量母离子数量

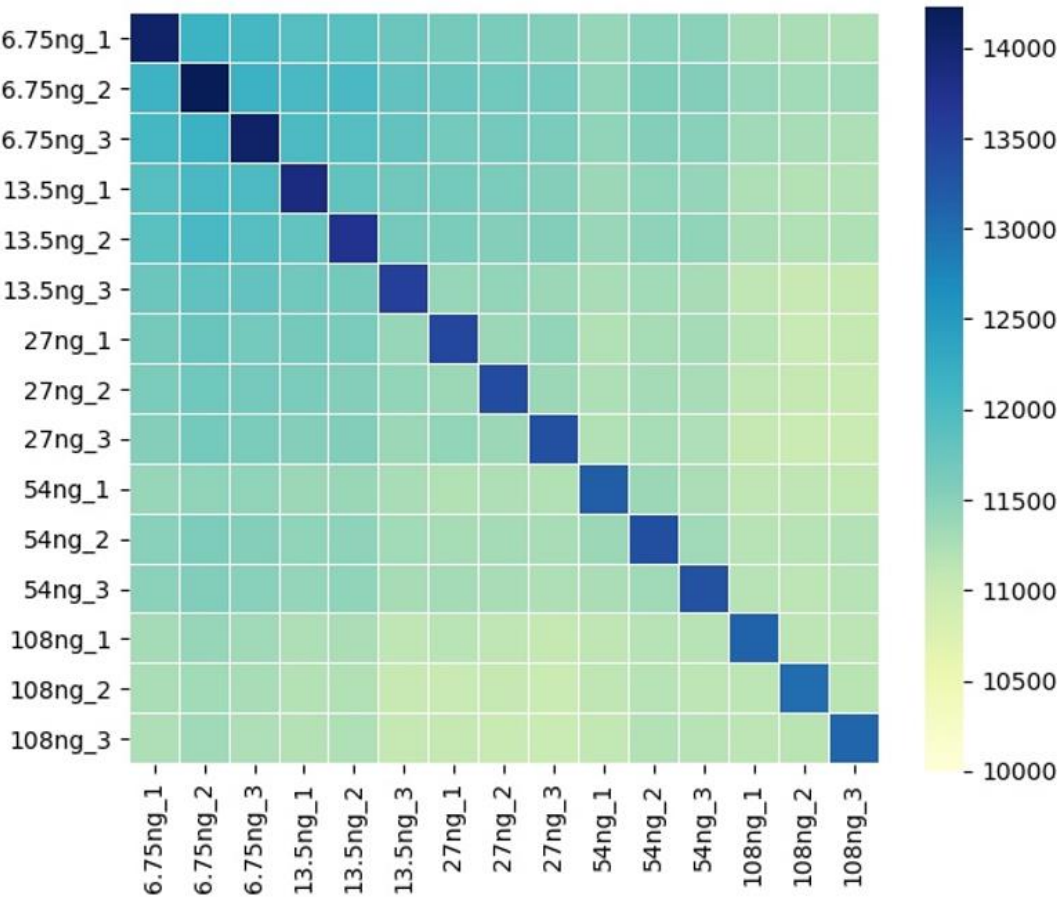


图 3: FIGS 每对数据共同母离子数量

对于 Specter 来说，在 15 次实验中，单次实验可以定量 4542 至 5009 个母离子，平均

可以定量 4762 个母离子。每对 DIA 质谱数据之间定量的共同母离子数量为 3564 到 4099 个，平均每对实验的共同母离子数量为 3760 个。

单个实验定量母离子数量(不含磷酸肽): 4542 5009
平均单个实验定量母离子数量(不含磷酸肽): 4762.2
共同母离子数量: 3564 4099
平均每对实验的共同母离子数量: 3760.057142857143

图 4: Specter 定量母离子数量

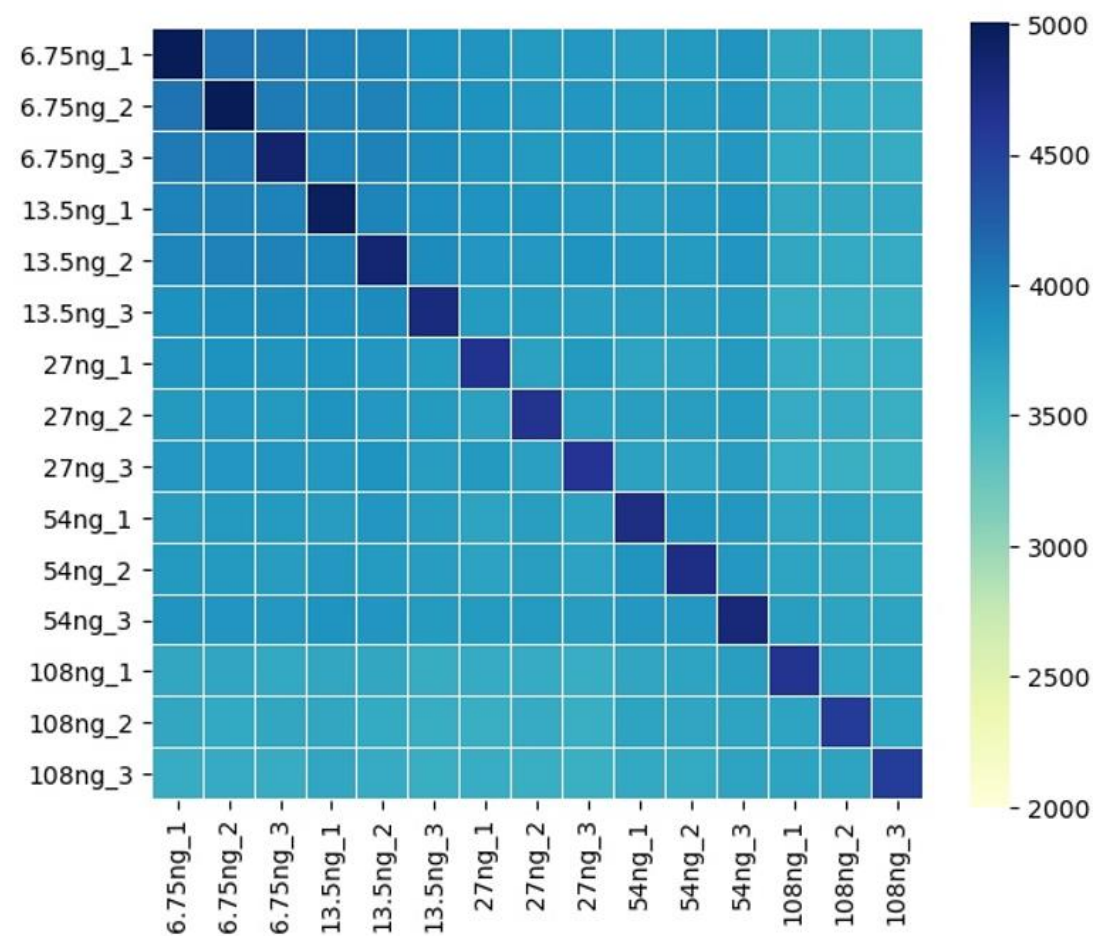


图 5: Specter 每对数据共同母离子数量

可以看出相比于 Specter, FIGS 显著提高了单次实验定量及多次试验重复定量的母离子数量。

之后计算了每对 DIA 质谱数据之间定量的共同母离子的相关系数:

对于 FIGS 来说, 在 15 次实验中, 每对实验定量的共同母离子的相关系数为 0.804 到 0.9975, 平均相关系数为 0.963。

共同母离子的相关系数： 0.9307 0.9886
平均相关系数： 0.9706171428571427

图 6: FIGS 母离子相关系数

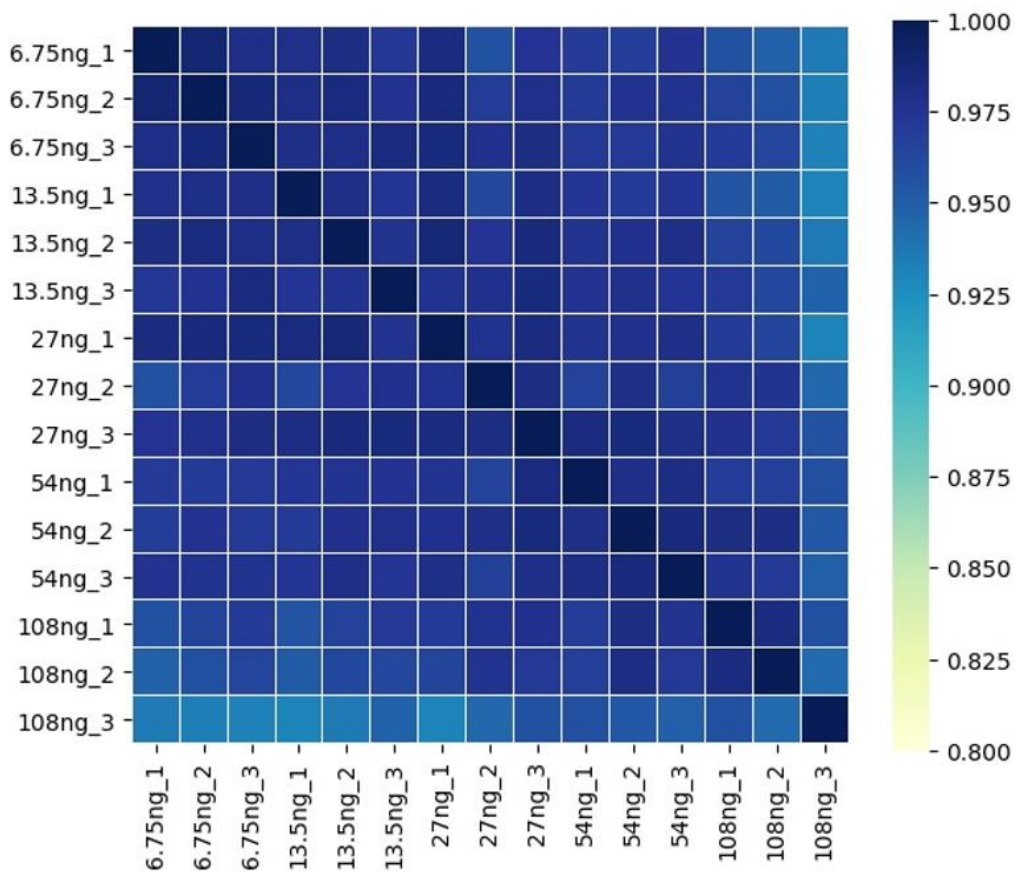


图 7: FIGS 每对共同母离子相关系数

对于 Specter 来说，在 15 次实验中，每对实验定量的共同母离子的相关系数为 0.804 到 0.9975，平均相关系数为 0.963。

共同母离子的相关系数： 0.804 0.9975
平均相关系数： 0.9630285714285713

图 8: Specter 母离子相关系数

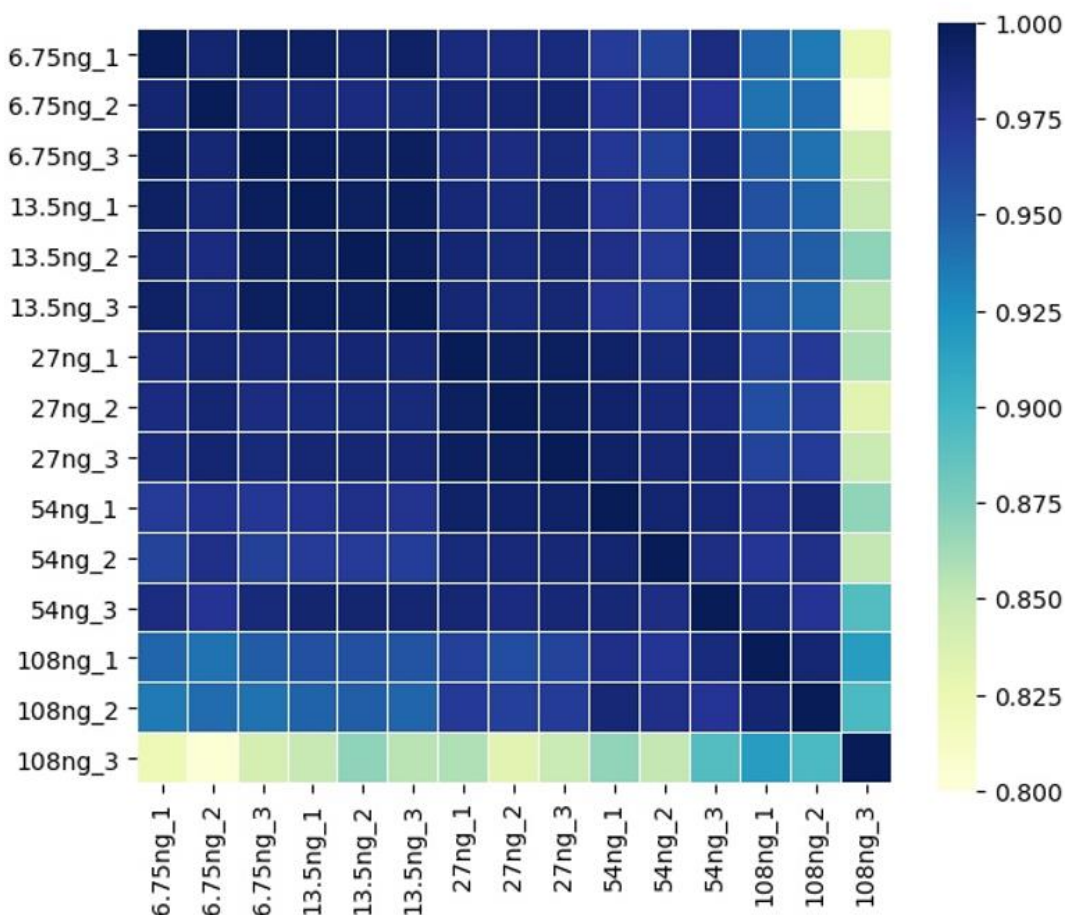


图 9: Specter 每对共同母离子相关系数

可以看出 FIGS 相对于 Specter 具有较高的定量的相关性。即使是不同浓度下的样品实验结果, FIGS 仍然有着较高的相关性 (大于 0.93), 而 Specter 为 0.8 以上。

在各个 HEK293T 样品中还掺入了已知量的合成磷酸化肽段, 下面是 FIGS 和 Specter 对图谱库中 85 种合成磷酸化肽段的定性结果:

CS20170831_SV_HEK_SpikeP100_6point75ng_Overlap22_01 :	FIGS 定量合成磷酸化肽段数量为 65	Specter 定量的合成磷酸化肽段数量为 38
CS20170831_SV_HEK_SpikeP100_6point75ng_Overlap22_02 :	FIGS 定量合成磷酸化肽段数量为 65	Specter 定量的合成磷酸化肽段数量为 49
CS20170831_SV_HEK_SpikeP100_6point75ng_Overlap22_03 :	FIGS 定量合成磷酸化肽段数量为 61	Specter 定量的合成磷酸化肽段数量为 45
CS20170831_SV_HEK_SpikeP100_13point5ng_Overlap22_01 :	FIGS 定量合成磷酸化肽段数量为 68	Specter 定量的合成磷酸化肽段数量为 60
CS20170831_SV_HEK_SpikeP100_13point5ng_Overlap22_02 :	FIGS 定量合成磷酸化肽段数量为 70	Specter 定量的合成磷酸化肽段数量为 62
CS20170831_SV_HEK_SpikeP100_13point5ng_Overlap22_03 :	FIGS 定量合成磷酸化肽段数量为 68	Specter 定量的合成磷酸化肽段数量为 58
CS20170831_SV_HEK_SpikeP100_27ng_Overlap22_01 :	FIGS 定量合成磷酸化肽段数量为 73	Specter 定量的合成磷酸化肽段数量为 69
CS20170831_SV_HEK_SpikeP100_27ng_Overlap22_02 :	FIGS 定量合成磷酸化肽段数量为 73	Specter 定量的合成磷酸化肽段数量为 69
CS20170831_SV_HEK_SpikeP100_27ng_Overlap22_03 :	FIGS 定量合成磷酸化肽段数量为 72	Specter 定量的合成磷酸化肽段数量为 66
CS20170831_SV_HEK_SpikeP100_54ng_Overlap22_01 :	FIGS 定量合成磷酸化肽段数量为 75	Specter 定量的合成磷酸化肽段数量为 74
CS20170831_SV_HEK_SpikeP100_54ng_Overlap22_02 :	FIGS 定量合成磷酸化肽段数量为 74	Specter 定量的合成磷酸化肽段数量为 73
CS20170831_SV_HEK_SpikeP100_54ng_Overlap22_03 :	FIGS 定量合成磷酸化肽段数量为 75	Specter 定量的合成磷酸化肽段数量为 74
CS20170831_SV_HEK_SpikeP100_108ng_Overlap22_01 :	FIGS 定量合成磷酸化肽段数量为 77	Specter 定量的合成磷酸化肽段数量为 75
CS20170831_SV_HEK_SpikeP100_108ng_Overlap22_02 :	FIGS 定量合成磷酸化肽段数量为 76	Specter 定量的合成磷酸化肽段数量为 77
CS20170831_SV_HEK_SpikeP100_108ng_Overlap22_03 :	FIGS 定量合成磷酸化肽段数量为 76	Specter 定量的合成磷酸化肽段数量为 75

图 10: FIGS 和 Specter 15 次实验定量的合成磷酸化肽段数量

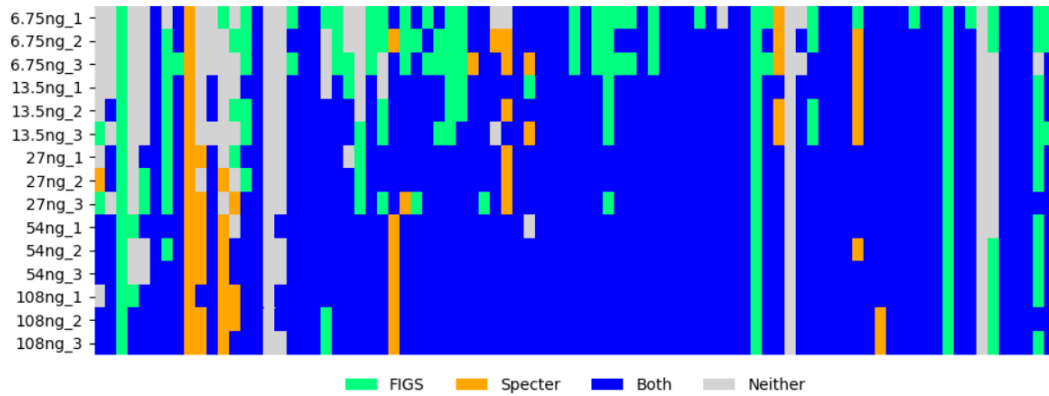


图 11: FIGS 和 Specter 定量的合成磷酸化肽段分布

FIGS 在 15 次实验中可以定量出 61 到 77 个合成磷酸化肽段，而 Specter 在 15 次实验中可以定量出 38 到 77 个合成磷酸化肽段。对于最低掺入浓度的 6.75ng 样品，FIGS 平均可以定量出 64 个合成磷酸化肽段，而 Specter 平均只能定量出 44 个合成磷酸化肽段。这说明 FIGS 具有较高的灵敏度，可以有效鉴定更多掺入水平较低的合成磷酸化肽段。

由于在 HEK293T 样品中掺入的合成磷酸化肽段具有 5 种已知的浓度梯度，因此下面验证 FIGS 相应 DIA 质谱数据中这些合成磷酸化肽段的相对丰度与掺入的浓度梯度之间的变化关系：

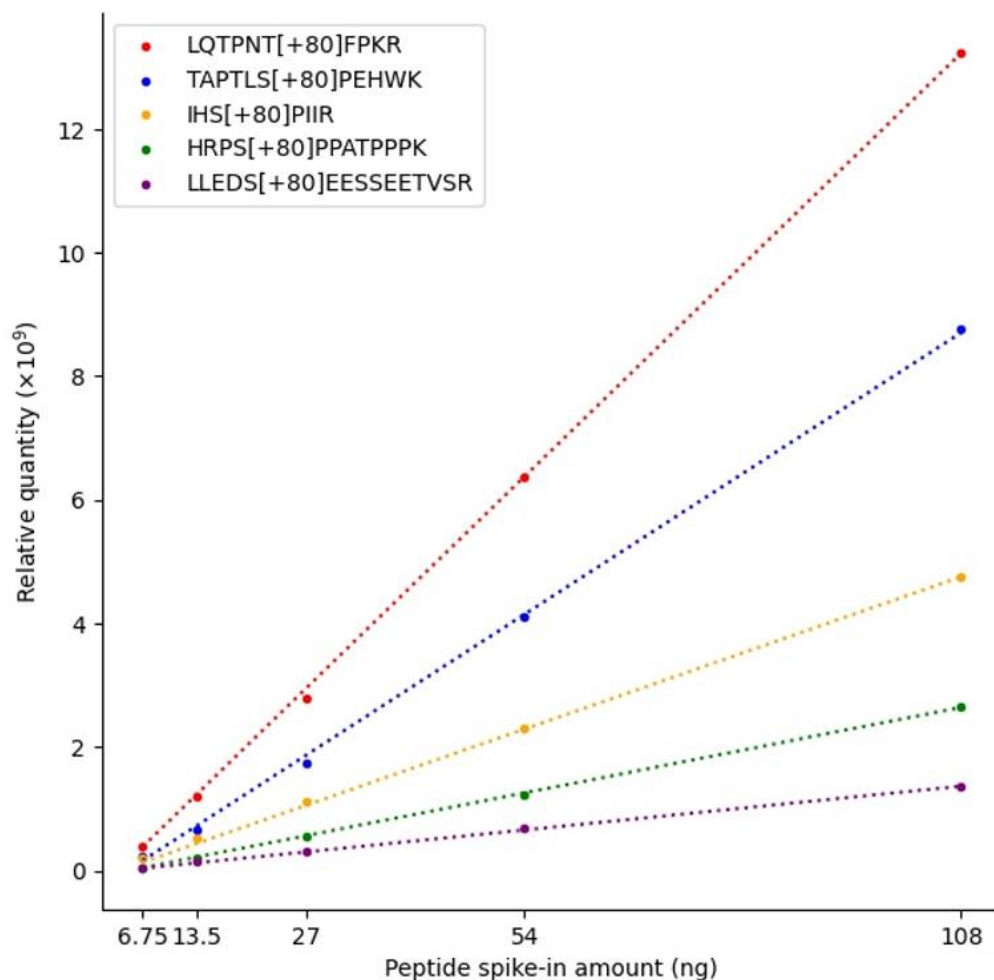


图 12: 合成磷酸化肽段的相对丰度

图中展示了 5 种浓度梯度下部分肽段的平均相对丰度, 其中虚线表示对应肽段相对丰度的理论校准线, 可以看出 FIGS 估计的肽段相对丰度与实际含量呈现出高度相关。

经过以上的验证实验, 可以得出以下结论:

- 1、相较于 Specter, FIGS 单次 DIA 实验可以显著增加定量的母离子数量。
- 2、相较于 Specter, FIGS 多次 DIA 实验 (同一种样品) 的母离子相对丰度的相关系数更高, 即使样品浓度差异较大, FIGS 估计的相对丰度仍然服从一致分布。
- 3、相较于 Specter, FIGS 具有更高的灵敏度, FIGS 可以有效鉴定更多掺入水平较低的合成磷酸化肽段。
- 4、在所有掺入浓度下, FIGS 估计的相关丰度与合成磷酸化肽段的实际含量高度线性相关。

工作计划

后面我会继续一边进行实验的复现一边阅读对应的代码。接下来的实验就是对不同物种的混合样品进行分析，所使用的数据集为 HYE124 质谱数据集，它们是由人类、酵母、大肠杆菌三种物种的蛋白质混合而成的。这一部分实验和 LFQbench 有很大的关系，所以我还会先阅读一下有关的论文。