工作总结与计划（2 月 7 日到 2 月 24 日）

代驾

dg310012@mail.ustc.edu.cn

2022-2-24

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，最近我仔细阅读了唐师兄 RI-FIGS 的所有代码，掌握了代码的逻辑和涉及到的方法，之后执行 RI-FIGS 程序运行得到了各项实验的实验结果。

第二部分是对未来工作计划的一个安排。

工作总结

代码阅读

这段时间，我仔细阅读了唐师兄的 RI-FIGS 代码，RI-FIGS 代码共包括 4 个 python 文件，它们分别是 RI-FIGS-ID.py、RI-FIGS_Quant.py、myms.py 和 sparse_nnls.py。其中 RI-FIGS-ID.py 用于肽段的定性，RI-FIGS_Quant.py 用于肽段的定量，myms.py 和 sparse_nnls.py 中主要是包括了一些函数用于调用，其中大部分函数在 FIGS 代码中已经出现过，有些函数则是新添加的。

sparse_nnls.py 中主要是包括了一些矩阵处理的函数，在 FIGS 代码中也都全部出现过。

myms.py 中主要包括了以下几个函数：

1、LoadMS2(path)

该函数的作用是读取质谱数据。

2、GenerateDecoyLibrary(SpectraLibrary, distance)

该函数的作用是使用母离子交换-离子峰偏移的反库构建策略去构建反库。

3、cal_nnls(LibIntensity, MS2Intensity, penalty)

该函数的作用是根据图谱库离子峰强度和质谱数据离子峰强度来计算对应的量化系数。

4、Peaks_match(lib_spc, exp_spc, tol=2e-5)

该函数是之前 FIGS 中没有的函数，该函数的作用是根据输入的图谱库离子峰强度和质谱数据离子峰强度，输出匹配的 mz，lib 对应的匹配离子峰强度和质谱对应的匹配离子峰强度。在具体实现时，由于质谱数据和 lib 数据的 mz 都是从小到大排序的，所以可以利用归并排序的思想，输出匹配的 mz（两者的 mz 误差绝对值在 tol 之内）。

5、deconvolute(SpectrumInfo, SpectraLibrary, tol, Top10First, level=1)

该函数就是原先 FIGS 中的解谱函数，对当前的某个单元质谱数据进行解谱得到肽段组成及其系数。

RI-FIGS-ID.py 的作用是进行肽段的定性分析。该文件包含以下几个部分：

1、获取命令行参数

在终端执行时，需要给出以下几个参数：

mzML_file：质谱数据文件。

SSM_file：csodiaq 生成的肽谱匹配结果（没有 FDR 过滤之前的）。

lib_file：图谱数据。

start_cycle：设置的开始 cycle。

end_cycle：这是的结束 cycle。

good_shared_limit：判断是否为较优 target 的标准之一。

good_cos_sim_limit：判断是否为较优 target 的标准之一。

good_sqrt_cos_sim_limit：判断是否为较优 target 的标准之一。

good_count_within_cycle_limit：判断是否为较优 target 的标准之一。

tol：离子峰匹配误差

scans_per_cycle：每个 cycle 下有多少个 scan

seed：随机数种子。

其中 SSM_file 是 csodiaq 软件生成的文件，该文件是 csodiaq 软件 FDR 过滤前的 SSM 结果文件，文件中存有定性出的肽段的名称、图谱库匹配离子峰的 m/z 与电荷量 z、查询峰和图谱峰的强度、匹配余弦相似度、MACC 分数等信息。

2、读取并处理图谱库数据

读取图谱库数据后，只取图谱库中每个图谱强度最高的 10 个峰，并做归一化处理。

3、读取 SSM 文件

读取 SSM 文件后，设置每条数据的 cycle，并选择位于所设置的 cycle 范围内的数据，并添加母离子项（肽段名称+“-”+图谱库离子峰电荷数）。

4、读取质谱数据

5、特征提取

对于 SSM 文件中的每条数据，先根据对应图谱库离子峰和质谱数据离子峰得到离子峰匹配结果（匹配 mz，lib 和质谱数据对应匹配离子峰的强度）。之后计算 lib 和质谱的匹配余弦相似度、开根号后的余弦相似度、均方误差 MSE、平均绝对误差 MAE、lib 中匹配的峰强度除以总峰强度。这些特征可以在后续说明定性结果的可靠性，将其写入 csv 文件中备用。

6、根据肽段字段前缀分辨出肽段来源（正/反库）

7、计算每个 cycle 下每个肽段出现次数以及最好的 MaCC_Score

MaCC_Score 是匹配计数与余弦分数，是 csodiaq 论文中提出的一种新的肽谱匹配分数，

该分数提高了目标诱饵分析中的目标识别能力，csodiaq 软件输出的结果文件就包含这一字段。而在这一段代码中，对于每个 cycle，统计每个肽段出现的次数，并按照肽段分组，选出最好的 MaCC_Score。将肽段出现次数和 MaCC_Score 最优值都暂存下来。

8、计算所有循环下每个肽段最好的 MaCC_Score

根据刚才的结果统计所有循环下每个肽段最好的 MaCC_Score。

9、筛选出较优的 target

首先根据 protein 字段将 target 和 decoy 筛选出来，之后对于 target 数据，根据之前计算的特征信息，如共享峰数量，余弦相似度，开根号后的余弦相似度，每个 cycle 肽段出现次数，筛选出较优的 target，只有这些特征值都大于各自阈值时，才会被认为是较优的 target。

10、训练 LDA 分类器

在得到 good_target 集合后，从 decoy 集合中随机选取元素，保证选出的元素个数等于 good_target 集合大小，并将这些元素与 good_target 合并成 DataSet_df。这样就得到了一个集合，一半元素由 decoy 数据组成，另一半由较优的 target 数据组成。利用样本特征和标签进行 LDA 线性判别分析，计算出相应的权重值，再将权重值乘以各个样本特征就得到了 LDA_Score。

11、计算 FDR

将处理过的 SSM 数据按照 LDA_Score 降序排序，这样 target 数据将更多的排在前面，而 decoy 数据则会更多的排在后面，根据数据标签计算 FDR 值，FDR 的计算公式为 FDR=decoy_number/target_number。因此 FDR 值最开始时为 0，后面逐渐增加，直至一个定值，该值等于 SSM 结果中 decoy 集合大小/target 集合大小。

12、计算 p-value

同样根据 LDA_Score 降序排序的 SSM 数据，计算每条数据的 p-value，根据 p-value 的计算公式，p-value=(decoy_number+1)/(total_decoy+1)。可知，p-value 值开始时为一个接近于 0 的正数，后面逐渐增加至 1。

13、筛选出 FDR<=0.01 的肽段

根据求得 FDR 值，筛选出 FDR<=0.01 的所有肽段。

RI-FIGS_Quant.py 的作用是进行肽段的定量分析。该文件主要包括以下几个部分：

1、获取命令行参数

在终端执行时，需要给出以下几个参数：

mzML_file：质谱数据文件。

SSM_file： RI-FIGS-ID.py 生成的定性结果。

lib_file：图谱数据。

topN：图谱库选择强度最高的峰个数

tol：离子峰匹配误差

2、读取并处理图谱库数据

读取图谱库数据后，只取图谱库中每个图谱强度最高的 topN 个峰，并做归一化处理。

3、读取 SSM 文件

读取 SSM 文件后，添加母离子项（肽段名称+“-”+图谱库离子峰电荷数）。

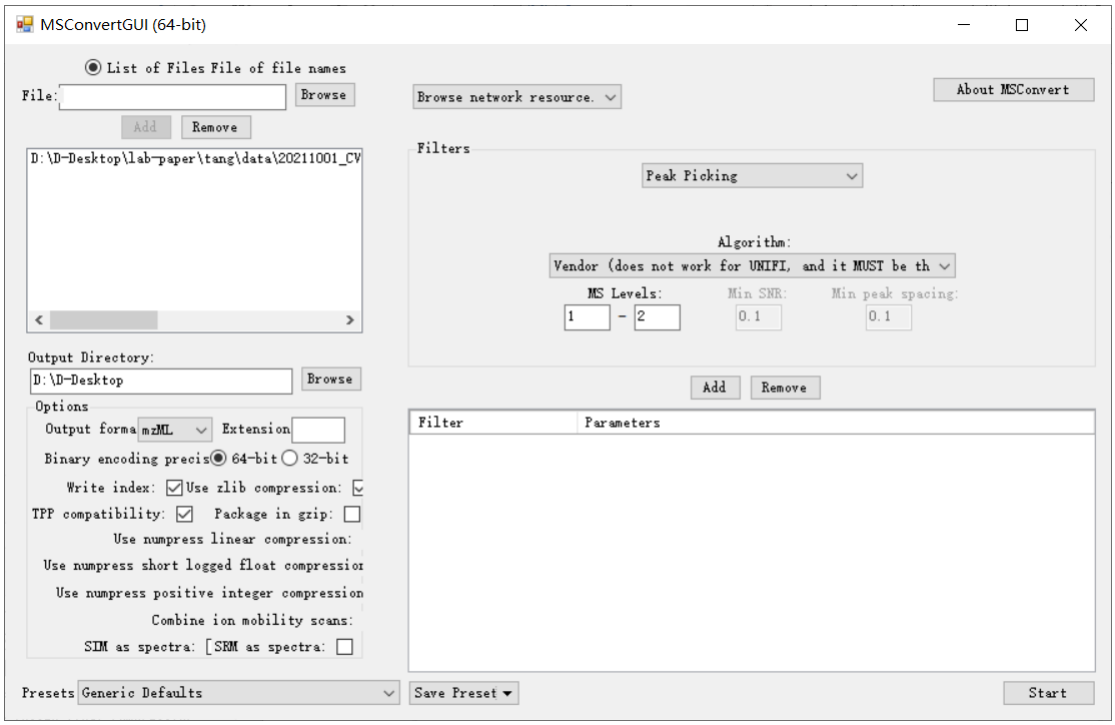
4、定量

对于每一个 scan，获取该 scan 下的定性信息，将定性结果中定性出的所有肽段作为当前 scan 的图谱库，调用 FIGS 程序的 deconvolute 函数进行解谱，即可得到每个图谱对应的 coeff 值。

程序运行

之前已经得到了 csodiaq 的 SSM 结果文件，在服务器上也已经配置好了环境，为了执行 RI-FIGS 代码，则需要制备 mzml 质谱数据。

使用 MSConvert 将 raw 格式文件转化为 mzml 格式。



需要注意的是，Filters 模块需要选择 Peak Picking，并设置 MS Levels 为 1-2。之前没有设置好这个，导致后续执行程序时出现错误。

之后开始进行实验，首先是 Hela DISPA 数据，使用 csodiaq 执行定性实验可以得到定性结果和 SSM 结果（FDR 过滤前）。之后根据 SSM 结果，执行 RI-FIGS 中的 RI-FIGS-ID.py 进行定性，即可得到 RI-FIGS 的定性结果文件如下图所示：

	peptide	scan	zLIB	cosine	Peaks(Libr	shared	MaCC_Sc	cycle	cos_sim	sqrt_cos_s	norm_msc	norm_ma	match_it_r	match_nu	protein	label	count_wit	cycle_cou	LDA_Score	FDR	p-value
2	VIQC+57.(	12678	2	0.949344	10	10	1.504608	9	0.807171	0.949344	0.038566	0.1576	1	10	TARGET	1	11	12	136.4594	0	1.49E-05
3	ALTVPELT	3466	2	0.997804	10	10	1.581413	3	0.997704	0.997804	0.000459	0.01898	1	10	TARGET	1	10	12	134.2105	0	1.49E-05
4	LVQSPNS	4933	2	0.997273	10	10	1.580571	4	0.992662	0.997273	0.001468	0.034167	1	10	TARGET	1	10	12	134.1912	0	1.49E-05
5	HFSVEGQ	17388	2	0.995796	10	10	1.578231	12	0.990082	0.995796	0.001984	0.030245	1	10	TARGET	1	10	12	134.1637	0	1.49E-05
6	ADLINNLC	2362	2	0.985944	10	10	1.562616	2	0.96896	0.985944	0.006208	0.049232	1	10	TARGET	1	10	12	134.0765	0	1.49E-05
7	IHFPLATY(	6240	2	0.988864	10	9	1.534564	5	0.98222	0.988864	0.003951	0.054166	0.968147	9	TARGET	1	12	12	133.0313	0	1.49E-05
8	VQNNLYH	17574	3	0.957593	10	8	1.45144	12	0.888472	0.957593	0.027882	0.15189	0.889429	8	TARGET	1	14	12	130.7184	0	1.49E-05
9	IQVLQQQ	14212	2	0.990063	10	10	1.569144	10	0.983605	0.990063	0.003279	0.038312	1	10	TARGET	1	9	12	130.2368	0	1.49E-05
10	TIAEC+57.	3459	2	0.991044	10	10	1.570699	3	0.980774	0.991044	0.003845	0.048917	1	10	TARGET	1	9	12	130.2077	0	1.49E-05
11	LDALVAEE	4964	2	0.987263	10	10	1.564707	4	0.965391	0.987263	0.006922	0.072405	1	10	TARGET	1	9	12	130.1106	0	1.49E-05
12	HLEINPDH	12497	2	0.983192	10	10	1.558255	9	0.94799	0.983192	0.010402	0.083294	1	10	TARGET	1	9	12	129.9407	0	1.49E-05

RI-FIGS 的定性结果除了包含肽段名称外，还包含了该肽段的各项特征信息，从结果中可以看出，定性出的肽段从上到下 LDA_Score 越来越小，FDR 值从 0 开始逐渐增加直至到达 0.01 左右，p-value 值也从一个接近于 0 的数字开始逐渐增加。

之后利用 RI-FIGS 的定性结果执行 RI-FIGS_Quant.py，通过调用 FIGS 中的解谱函数即

可得到定量结果，由于 RI-FIGS 的定性结果只取 MACC_Score 分数最高时的肽段信息，所以解谱时每个肽段只会对应一个系数，使用该系数即可代表该肽段的定量值。

之后使用同样的方法运行100次重复实验DISPA数据和大肠杆菌酵母菌人类蛋白质混合样品 DISPA 数据，得到相应的定性结果和定量结果。

工作计划

目前已经运行得到了各项实验的实验结果，后续则需要对重复实验结果进行分析，绘制相应的图表。为了提升自己的代码能力，为后续科研工作打下基础，之后我计划编写相应的代码对唐师兄论文中出现的图表进行复现，这一工作预计需要一到两周的时间完成。