工作总结与计划（9 月 5 日到 9 月 28 日）

代驾

dg310012@mail.ustc.edu.cn

2021-09-28

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，最近我主要是针对 FIGS 部分母离子特征峰匹配较差这一缺点尝试进行了改进，实验验证了之前想的一个解决方案，但是从实验结果来看这个方案并不能提升 FIGS 的定量准确性。

第二部分是对未来工作计划的一个安排。

工作总结

实验设计

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，即使用 Specter 的方法来计算匹配较差的肽段母离子的系数，这种方法可以看成是 FIGS 和 Specter 的一次结合。

近期，我编写了实验代码对上述的想法进行了验证。在 FIGS 的代码中，每次图谱分解都会计算特征峰与对应 lib 峰的余弦相似度、协防差、lib 的特征峰强度、图谱的特征峰强度等参数，首先考虑特征峰与对应 lib 峰的余弦相似度这一信息，任意两个向量的余弦相似度是一个位于 0 到 1 之间的实数，余弦相似度越接近 1 就说明这两个向量越相似，越接近于 0 就说明这两个向量的相对差异越大，因此可以设计一个阈值 G，当特征峰与对应 lib 峰的余弦相似度小于这个阈值时就说明这次匹配是一次较差的匹配。

```
'''计算特异峰与lib峰的相似度'''
if len(unique_mz_ids_of_thislib) > 1:
    FeaturedPeakCosSim = cosine_similarity(
        [FeaturedPeakIntensitiesInDIA, FeaturedPeakIntensitiesInLib])[0][1]

    pdMultiQuant = pd.DataFrame(
        [FeaturedPeakIntensitiesInDIA, FeaturedPeakIntensitiesInLib])
    pdMultiQuant = pdMultiQuant.T
    corrQuant = pdMultiQuant.corr()
    FeaturedPeakCorr = corrQuant[0][1]
    if FeaturedPeakCosSim < GoodOrBad1:  # 匹配程度较差
        FeaturedPrecursorsIndexBad.append(i)
        GoodOrBadFlag = True
```

在对该图谱进行一轮解谱后就可以区分出哪些肽段母离子的特征峰匹配情况较优，哪些肽段母离子的特征峰匹配情况较差。如果存在着一系列匹配程度较差的母离子，就从原始质

谱中减去那些匹配程度较优的肽段母离子的强度与 coeff 的乘积，之后利用最小二乘法来重新计算匹配程度较差的肽段母离子的系数，使得这些母离子的线性组合结果与新得到的质谱数据最接近。

```
if len(FeaturedPrecursorsIndexBad) > 0:
    DIASpectrumDeepCopy = copy.deepcopy(DIASpectrum)
    # 从原始质谱中减去匹配程度较好的特征母离子强度
    for i in range(len(FeaturedPrecursorsGood)):
        for j in range(len(RowIndicesMatrix[FeaturedPrecursorsGood[i][0]])):
            RemainDIASpectrum = DIASpectrumDeepCopy[RowIndicesMatrix[FeaturedPrecursorsGood[i][0]][j]][1] - \
                FeaturedPrecursorsGood[i][1] * \
                MatrixIntensitiesMatrix[FeaturedPrecursorsGood[i][0]][j]
            # if RemainDIASpectrum > 0:
            DIASpectrumDeepCopy[RowIndicesMatrix[FeaturedPrecursorsGood[i]
                                                 [0]][j]][1] = RemainDIASpectrum
```

```
# 仿照Specter计算匹配程度较差的特征母离子对应的系数
UniqueRowIndicesBad = [
    i for j in FeaturedPrecursorsIndexBad for i in RowIndicesMatrix[j]]
UniqueRowIndicesBad = list(set(UniqueRowIndicesBad))
UniqueRowIndicesBad.sort()
DIASpectrumIntensitiesBad = DIASpectrumDeepCopy[UniqueRowIndicesBad, 1]
DIASpectrumIntensitiesBad = np.append(DIASpectrumIntensitiesBad, [0])
SparseColumnIndicesForPeaksNotInDIABad = np.arange(
    len(FeaturedPrecursorsIndexBad))
NumRowsOfLibraryMatrixBad = max(UniqueRowIndicesBad)
SparseRowIndicesForPeaksNotInDIABad = [
    NumRowsOfLibraryMatrixBad+1]*len(SparseColumnIndicesForPeaksNotInDIABad)
SparseMatrixEntriesForPeaksNotInDIABad = np.array([np.sum([NormalizedRefPeptideCandidateList[j][k]
                                                           for k in range(len(NormalizedRefPeptideCandidateList[j]))
                                                           if IdentPrecursorsLocations[j][k] % 2 == 0])
                                                   for j in FeaturedPrecursorsIndexBad])
RefSpectraLibrarySparseRowIndicesBad = np.array([
    i for j in FeaturedPrecursorsIndexBad for i in RowIndicesMatrix[j]])
SparseRowIndicesBad = np.append(
    RefSpectraLibrarySparseRowIndicesBad, SparseRowIndicesForPeaksNotInDIABad)
RefSpectraLibrarySparseColumnIndicesBad = np.array([
    i for j in range(len(FeaturedPrecursorsIndexBad))
    for i in [j] * len([k for k in IdentPrecursorsLocations[FeaturedPrecursorsIndexBad[j]] if k % 2 == 1])
])
SparseColumnIndicesBad = np.append(
    RefSpectraLibrarySparseColumnIndicesBad, SparseColumnIndicesForPeaksNotInDIABad)
RefSpectraLibrarySparseMatrixEntriesBad = np.array([
    i for j in FeaturedPrecursorsIndexBad for i in MatrixIntensitiesMatrix[j]])
SparseMatrixEntriesBad = np.append(
    RefSpectraLibrarySparseMatrixEntriesBad, SparseMatrixEntriesForPeaksNotInDIABad)
SparseRowIndicesBad = stats.rankdata(SparseRowIndicesBad, method='dense').astype(
    int) - 1
LibrarySparseMatrixBad = sparse.coo_matrix(
    (SparseMatrixEntriesBad, (SparseRowIndicesBad, SparseColumnIndicesBad)))
LibraryCoeffsBad = sparse_nnls.lsqnonneg(
    LibrarySparseMatrixBad, DIASpectrumIntensitiesBad, {'show_progress': False})
LibraryCoeffsBad = LibraryCoeffsBad['x']
```

实验结果

首先针对特征峰匹配相似度这一信息，设置了多个阈值（0.1 到 0.9）重现了常规 DIA 实验分析中的单一样品分析实验。

考察平均每对DIA质谱数据之间定量的共同HEK293T母离子数量这一指标来说明定量结果好坏，得下表：

相似度阈值	共同 HEK293T 母离子数量
0.1	11396
0.2	11279
0.3	11041
0.4	10674
0.5	10193
0.6	9596
0.7	8862
0.8	9930
0.9	6305

相比之下，**FIGS 的共同母离子数量为 11418 个，Specter 的共同母离子数量为 3760 个，**从上表中可以看出，当阈值等于 0.1 时，定量的共同 HEK293T 母离子数量略小于 FIGS，随着阈值的增加，最终定量得出的共同 HEK293T 母离子数量在减少，但是均大于 Specter。考虑到特征峰与对应 lib 峰的余弦相似度未必能完全反映出两者之间的相似关系，实验又引入了特征峰与对应 lib 峰的协方差这一信息，协方差大于 0，说明两者正向相关，协方差小于 0 说明两者负向相关，最理想的情况是特征峰与对应 lib 峰的协方差等于 1，说明两者线性正相关，没有任何误差。

后面我又做了只考虑协方差信息和同时考虑相似度和协方差这两组实验，并且细化了阈值最小值附近的阈值区间，实验结果如下表：

协方差阈值	共同 HEK293T 母离子数量
-0.99	11350
-0.98	11348
-0.97	11347

-0.96	11346
-0.95	11345
-0.93	11342
-0.9	11339
-0.8	11325
-0.7	11308

可以看出即使不断降低阈值，新程序的定量结果也无法超过 FIGS 程序，当同时考虑相似度和协方差时，最终得到的也是类似的结果。

从实验结果可以看出，由于 FIGS 本身定量准确性明显优于 Specter，即使是舍弃 FIGS 中特征峰匹配非常差的结果，转而用 Specter 的方法重新计算相应的值，仍然无法提升算法的定量准确性。

工作计划

有关部分特征峰匹配的线性相关度较低这一问题，如果想要进行优化的话，还可以尝试去排除误差较大的特征离子峰，设计合适的算法选择相关度最高的一组特征离子峰进行图谱分解，有关这个想法我后面会进行实验设计并验证。另外，在和唐师兄进行交流后，FIGS程序并没有利用色谱信息进行定量，因此针对有色谱数据的定量还存在很大的提升空间，这个方向我后续也会进行学习和思考。