**基于DIA图谱分解的肽段定量分析软件V1.0**

**用户手册**

目录

[1引言 3](#_Toc88762818)

[1.1编写目的 3](#_Toc88762819)

[1.2项目背景 3](#_Toc88762820)

[1.3研究现状 3](#_Toc88762821)

[1.4项目简介 4](#_Toc88762822)

[2实验环境与环境配置 5](#_Toc88762823)

[2.1硬件环境 5](#_Toc88762824)

[2.2软件环境 5](#_Toc88762825)

[2.3环境配置 6](#_Toc88762826)

[3软件界面 6](#_Toc88762827)

[3.1概述 6](#_Toc88762828)

[3.2文件输入模块 7](#_Toc88762829)

[3.3参数模块 7](#_Toc88762830)

[3.4运行模块 8](#_Toc88762831)

[4软件使用 8](#_Toc88762832)

[4.1示例数据 8](#_Toc88762833)

[4.1.1质谱数据 8](#_Toc88762834)

[4.1.2库文件 8](#_Toc88762835)

[4.2软件运行 9](#_Toc88762836)

[4.2.1进入软件界面 9](#_Toc88762837)

[4.2.2添加质谱数据 9](#_Toc88762838)

[4.2.3添加图谱库数据 9](#_Toc88762839)

[4.2.4修改参数（可选） 10](#_Toc88762840)

[4.2.5执行 10](#_Toc88762841)

[4.2.6终止程序（可选） 11](#_Toc88762842)

[4.3结果保存 11](#_Toc88762843)

[4.3.1生成的文件 11](#_Toc88762844)

[4.3.2计算结束标志 13](#_Toc88762845)

[4.4多线程执行说明 13](#_Toc88762846)

# 1引言

## 1.1编写目的

本用户手册面向的用户主体是研究定量蛋白质组学相关方向的科研人员，主要目的是详细说明本软件所实现的功能及其运行方式，以便使用者了解本软件的适用场景和使用方法。

## 1.2项目背景

蛋白质是生命体的直接生物功能分子，为生命活动提供物质基础，因此对蛋白质的研究能够帮助人们探索各种生命活动机理，为生物学与医学研究作出重要贡献。蛋白质组的概念于 20 世纪末期被首次提出，指在生物组织或生物学环境中产生的一组蛋白质。继而衍生出对蛋白质组实验进行系统性大规模分析的蛋白质组学 (proteomics)，主要应用于癌症研究、药物和药物靶标发现、蛋白质间的相互作用以及发现生物标志物和阐明疾病机理，为直接在蛋白质水平上对基因和细胞功能进行大规模研究提供有力支撑。

蛋白质组学的进步是生物分子分离技术、质谱 (mass spectrometry, MS) 以及生物信息学分析等领域共同推动的结果，其中质谱技术在蛋白质组学中起着举足轻重的作用。早期质谱技术通常用于鉴定生物系统中的蛋白质及其水解产生的肽段。随着质谱仪器精度逐渐提高以及定性分析方法的日趋完善，对于生物样本的大规模蛋白质定量需求日益增加，对蛋白质和肽段的定量分析逐渐受到广泛关注，从而催生出定量蛋白质组学 (quantitative proteomics)。定量蛋白质组学不仅要鉴定存在于样本中的蛋白质，还要准确评估蛋白质的丰度。相比于传统的双向电泳，质谱技术具有更高的灵敏度和分辨率，因此逐渐成为定量蛋白质组学研究的主要手段。

自下而上的蛋白质组学 (bottom­up proteomics) 是当前主流的工作流程，通过分析蛋白质水解的肽段来测量蛋白质，主要包括以下步骤：（1）从生物样品中分离蛋白质混合物；（2）使用特定的蛋白酶将蛋白质水解为肽段；（3）通过质谱仪采集肽段样本的质谱数据；（4）对肽段进行定性、定量分析；（5）对蛋白质进行定性、定量分析。自下而上的蛋白质组学利用了肽段相对于蛋白质的优势，即肽段信息更容易被质谱仪采集，从而可以进行高通量分析，从复杂裂解物中鉴定和定量数千种蛋白质。

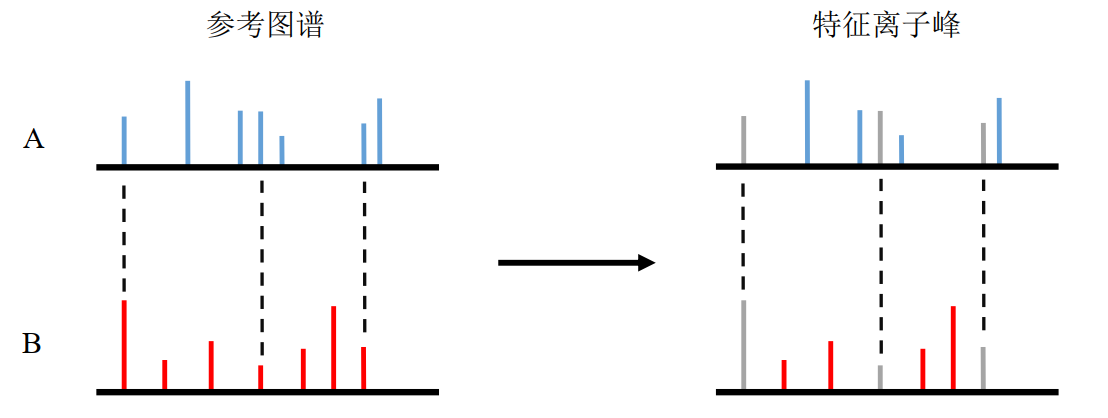
## 1.3研究现状

由于生物系统的复杂性，现阶段蛋白质组学实验在质谱采集之前通常要进行充分洗脱，并且对于复杂样品，早期的质谱模式鉴定同源蛋白质的能力不足，这些推动了液相色谱­串联质谱 (liquid chromatography­tandem mass spectrometry, LCMS/MS) 技术的迅速发展。在基于 LC­MS/MS 的蛋白质组学实验中，蛋白质样品通过水解转化为肽段，然后经过液相色谱洗脱和离子化，肽段母离子以不同的时间进入质量分析器并进行串联质谱分析。串联质谱是目前高通量蛋白质分析的主要技术，由一级质谱和二级质谱交替进行，其中一级质谱采集母离子信号形成一级质谱图 (MS1)，二级质谱根据一定策略从一级质谱中选择部分母离子进行碎裂，并采集相应的碎片离子信号形成二级质谱图 (MS2)。

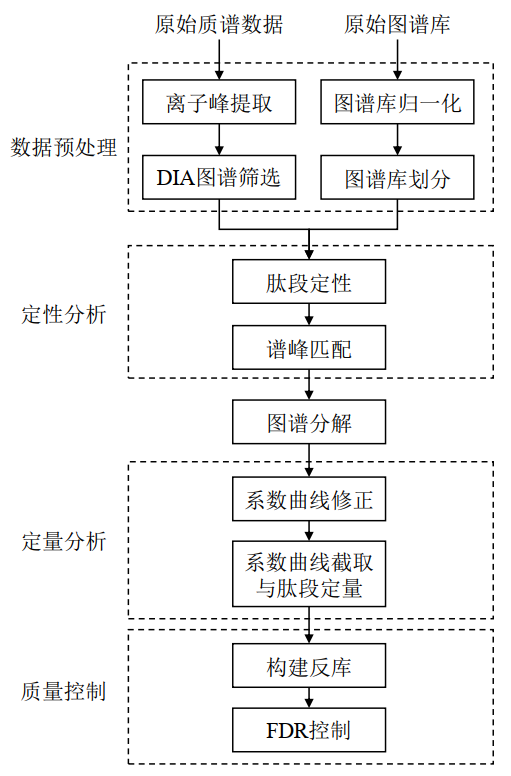
由于共洗脱的特点，LC-MS/MS 图谱极易生成混合图谱 (chimeric spectra），混合图谱是由至少两种肽段母离子共碎裂后形成的二级质谱图，通常被认为是对蛋白质组学研究的一个限制，因为它降低了图谱库 (spectral library) 搜索模型的鉴定性能，其中图谱库是存储单一肽段母离子碎裂产生的参考图谱的数据库。混合图谱分解的目标是从混合图谱中鉴定其肽段组成以及确定各肽段对混合图谱的贡献。由于近年来对面向 DIA质谱数据的定量方法研究逐渐增加，而 DIA 图谱通常为混合图谱，因此混合图谱分解问题的重要性逐渐得以凸显。针对图谱分解问题，目前虽然已经提出了多种策略与算法，但其准确度与效率仍存在较大的提升空间，这给肽段的定性及定量分析带来了巨大挑战。

## 1.4项目简介

为了更好地求解图谱分解问题，我们设计并实现了一种基于特征离子的DIA图谱分解算法Fisdec (featured ions-based spectral deconvolution)。对于一种肽段母离子，其特征离子是该母离子经串联质谱碎裂产生的一种碎片离子。在特定的肽段母离子集合中，特征离子仅由其中一种肽段母离子碎裂产生。由于特征离子的质荷比在这些肽段母离子产生的所有碎片离子中具有唯一性，因此特征离子能够唯一标识该集合中的一种肽段母离子。相应地，在这些肽段母离子的参考图谱中，反映特征离子相对强度的峰为特征离子峰。



与现有的基于全离子峰拟合的图谱分解方法相比，Fisdec在准确度和效率上均有了大幅提升。在Fisdec算法的基础上，我们设计研发了一个基于DIA图谱分解的肽段定量分析软件FIGS (featured-ions guided stoichiometry)。FIGS的主要流程如下图所示：



FIGS 以质谱数据和图谱库数据作为输入，在程序运行完毕后，就可以输出得到相应的肽段定量结果文件。该软件大大简化了定量蛋白质组学相关方向科研人员的工作量，为定量蛋白质组学领域提供了一个有效的计算工具，促进了该领域的发展。

# 2实验环境与环境配置

## 2.1硬件环境

开发环境：PC，Intel(R) Core(TM) i7-7700K CPU @4.20GHz，16G内存。

运行环境：最低1GHz CPU，推荐2.5GHz及以上；最低1G内存，推荐2G内存及以上。

## 2.2软件环境

开发环境：Windows 10, Microsoft Visual Studio 2015开发平台。

运行环境：Windows 7及以上，Apache 2.4，MongoDB 4.2。

## 2.3环境配置

首先使用git clone https://github.com/dg310012/FIGS.git命令将代码下载到本地。

代码包括以下几个文件：

requirements\_conda.txt --python3.6.5环境以及库

requirements\_R.txt --R语言环境所需库（3.6.x版本）

FIGS\_Deconvolute.py --FIGS解谱，python程序

myms.py

sparse\_nnls.py

FIGS\_Quant.R --FIGS进行定量，R语言程序

FIGS.py --FIGS界面程序，实现了解谱时多个文件同时计算，以及定量时多个文件同时计算

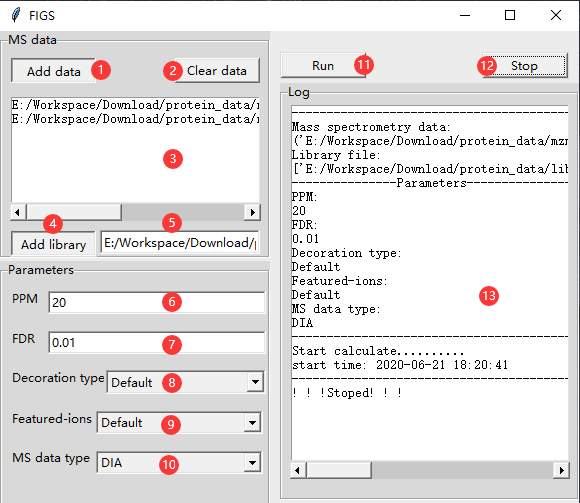
下面开始进行环境配置：

1. 先配置好python语言环境，利用pip或者anaconda配置python3.6.5环境以及需要的库（库的名称与版本号详见requirements\_conda.txt）。
2. 之后配置R语言环境，如果没有安装R软件，需要进入R的官网（<https://www.r-project.org/>）点击”download R”，之后根据操作系统选择相应的R软件进行安装。在安装完毕后，在R的控制台输入”install.packages("xxx")”命令对所需要的库进行安装，其中xxx表示所要安装的库的名称（需要用到的库的名称详见requirements\_R.txt）。
3. 为了利用python中的rpy2库来让python调用R语言程序，需要配置环境变量。需要配置R\_HOME为R语言安装路径，如 E:\Software\R。配置R\_USER为rpy2安装路径，如E:\Software\Anaconda3\envs\py36\Lib\site-packages\rpy2。

# 3软件界面

## 3.1概述

在配置完环境后，在Windows下使用CMD进入代码所在目录，执行”python FIGS.py”命令即可进入软件界面。软件界面如下图所示，主要分为三个部分：左上方的文件输入模块、左下方的参数模块、以及右边的运行和日志输出模块，下面对这三个模块进行逐一地介绍。



## 3.2文件输入模块

本模块用来输入质谱数据和库文件。对应上图序号1-5。

①点击“Add data”添加质谱文件，可以一次选择多个质谱文件作为输入。这里质谱数据文件必须为“.mzML”格式的文件。

②如果想删除已经添加的质谱文件，点击“Clear data”可以清除所有已选择的质谱数据文件。

③显示已经添加的质谱数据文件。

④点击“Add library”添加质谱数据对应的库文件，只能添加一个库文件作为输入。这里库文件必须为“.blib”格式的文件。

⑤显示已经添加的库文件。

## 3.3参数模块

本模块用来输入进行计算的参数。每个参数都设置了默认值，可以手动更改。对应上图序号6-10。

⑥“PPM”，(parts-per-million)，仪器的质量精度，以百万分之一为单位，默认为20，需要为整数。

⑦“FDR”，(False Discovery Rate)，默认为0.01，范围为大于0小于1的小数。

⑧“Decoration type”，修饰类型。Default（默认），暂时不能修改。

⑨“Featured-ions”，特征离子。Default（默认）：使用所有特征离子。或者Top10-first：优先使用属于相应光谱库中强度最高的十个离子作为特征离子来计算系数。当光谱库中没有满足此条件的特征离子时，将使用默认策略。

⑩“MS data type”，质谱数据类型。默认为DIA数据，RT-free暂时未集成。

## 3.4运行模块

本模块用来进行计算的控制和日志的输出。对应上图序号11-13。

⑪当质谱数据和库文件按要求选择完成，以及参数输入完成后，点击“Run”即可进行计算。

⑫若需要终止当前正在进行的计算，点击“Stop”即可终止。如果想修改参数或者修改计算的文件等，也需要先终止正在执行的计算才能进行新的计算。

⑬这里会显示一些提示信息，如未选择文件、文件格式不正确等。还会输出当前计算的日志，如已经完成的计算、计算进行到哪一步了、计算开始的时间、计算结束的时间等。同时会把计算过程中的关键信息保存到log.txt文件中。

# 4软件使用

## 4.1示例数据

### 4.1.1质谱数据

可以使用下面的链接下载测试的质谱数据：

CS20170831\_SV\_HEK\_SpikeP100\_27ng\_Overlap22\_01.mzML：

链接：<https://pan.baidu.com/s/14H-0xRMcxqlJBTtx5B_Qpg> 提取码：1234

CS20170831\_SV\_HEK\_SpikeP100\_27ng\_Overlap22\_02.mzML：

链接：<https://pan.baidu.com/s/1ojXmawv99V540a09jSP5kg> 提取码：1234

CS20170831\_SV\_HEK\_SpikeP100\_27ng\_Overlap22\_03.mzML：

链接：<https://pan.baidu.com/s/1X-cpoTGCf_UzfVfsk5ssyQ> 提取码：1234

上面三个链接下载下来为”.mzML”格式的文件，本软件只允许输入“.mzML”格式的质谱数据文件，如果质谱数据为其他格式，例如”.raw”格式，请将其转化为“.mzML”格式。

### 4.1.2库文件

可以使用下面的链接下载测试的库文件：

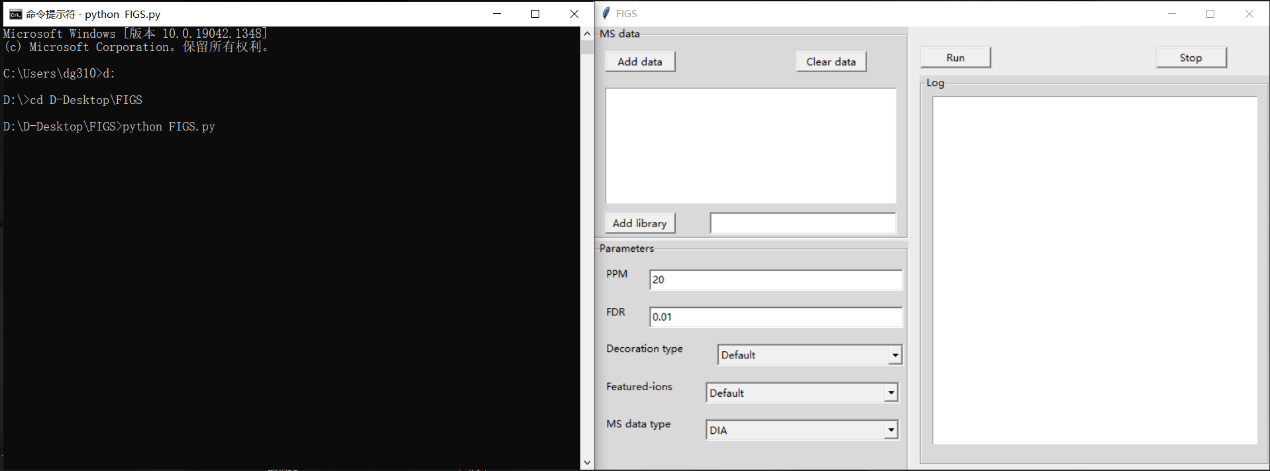
链接：<https://pan.baidu.com/s/13ZRw8rn8QIGILb3WZpgJKA> 提取码：1234

上面的链接下载下来为”.blib”格式的文件，本软件只允许输入“.blib”格式的图谱库文件，如果图谱库数据为其他格式，请将其转化为“.blib”格式。

## 4.2软件运行

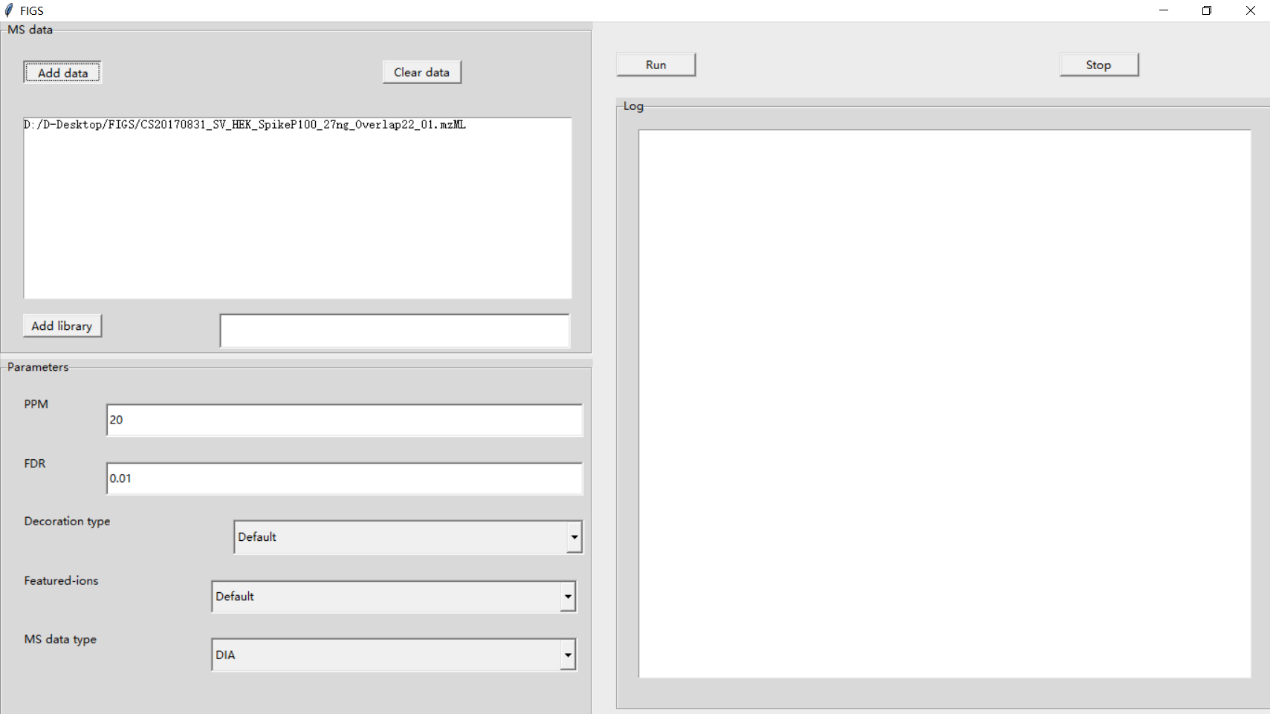
### 4.2.1进入软件界面

使用CMD进入代码所在目录，执行命令”python FIGS.py”，进入软件界面。



### 4.2.2添加质谱数据

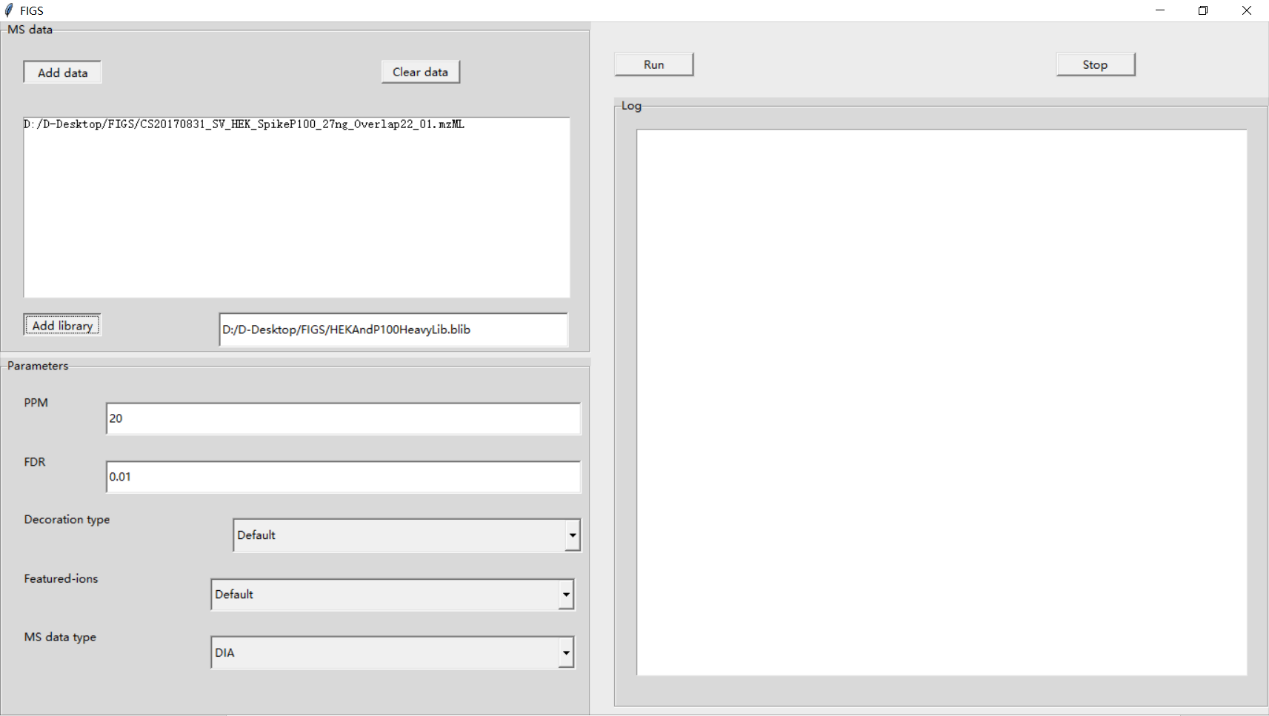
点击”Add data”按钮，添加质谱文件。



如果想删除已经添加的质谱文件，点击“Clear data”按钮可以清除所有已选择的质谱数据文件。

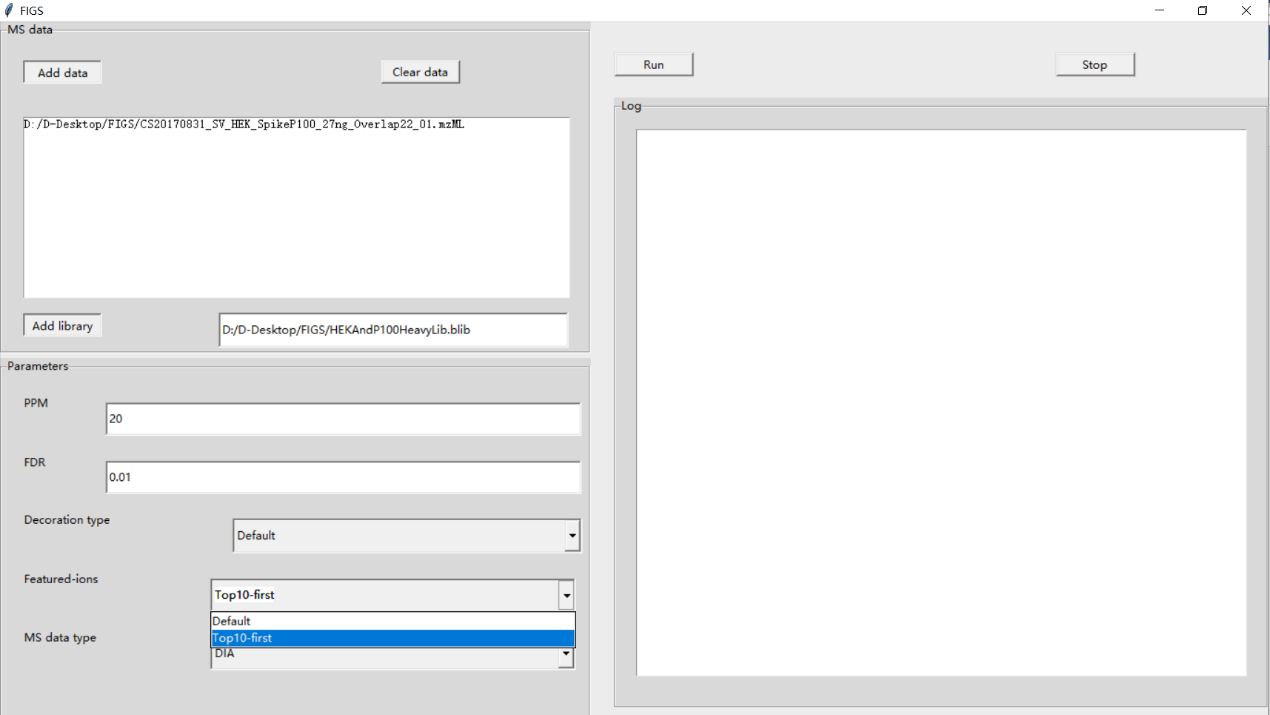
### 4.2.3添加图谱库数据

点击”Add library”按钮，添加图谱库文件。



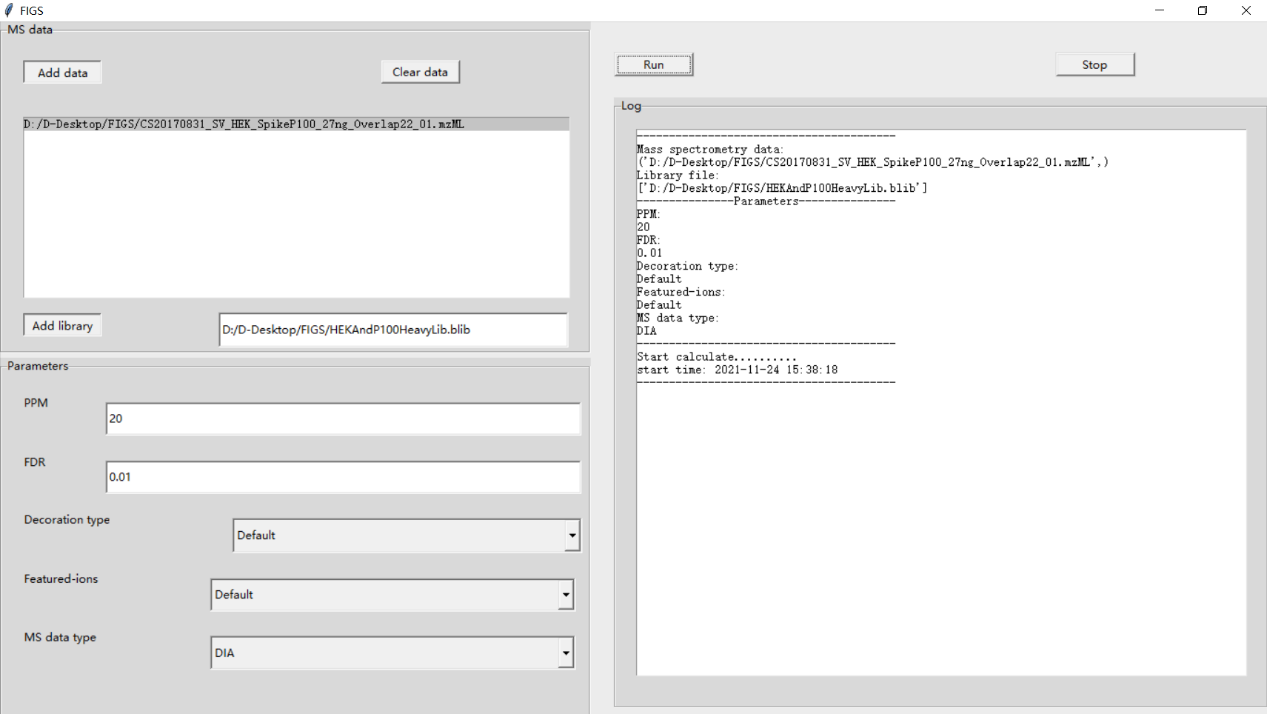
### 4.2.4修改参数（可选）

在Parameters模块下可以修改各个参数的值。



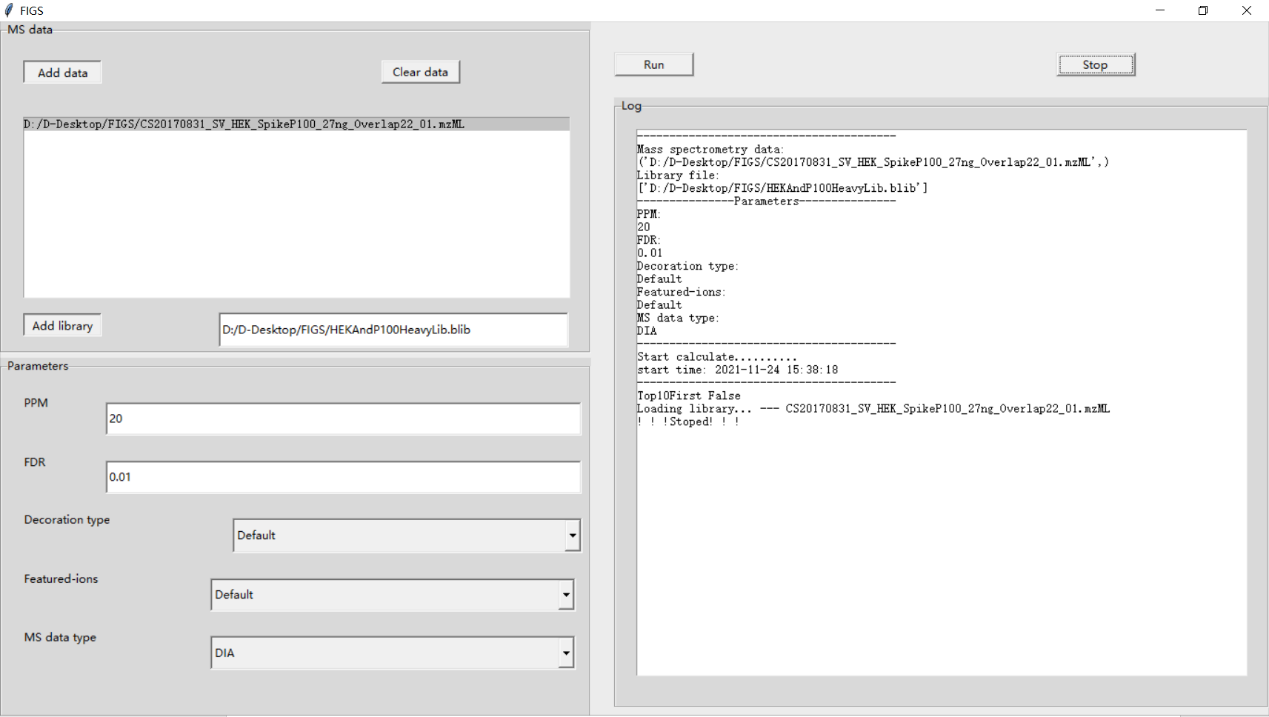
### 4.2.5执行

点击”Run”按钮即可开始执行。在Log模块下会显示一些提示信息，如未选择文件、文件格式不正确等。还会输出当前计算的日志，如已经完成的计算、计算进行到哪一步了、计算开始的时间、计算结束的时间等。



### 4.2.6终止程序（可选）

点击”Stop”按钮即可终止软件运行。



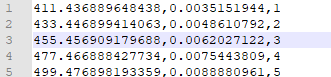
## 4.3结果保存

### 4.3.1生成的文件

在计算过程中会生成如下文件：

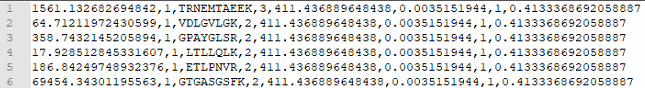
①在质谱文件 (.mzML格式文件)的同级目录下，会生成一个和质谱文件同名（不含后缀）的文件夹，所有的计算结果都会保存到该文件夹下。例如，对于前面的示例质谱数据1：“CS20170831\_SV\_HEK\_SpikeP100\_27ng\_Overlap22\_01.mzML”，会在质谱文件的同级目录下生成一个文件夹“CS20170831\_SV\_HEK\_SpikeP100\_27ng\_Overlap22\_01”。

②在前面生成的文件夹下，会先生成一个“header.csv”文件，格式如下，对应了三个信息：前体m/z窗口中心、保留时间和扫描索引。

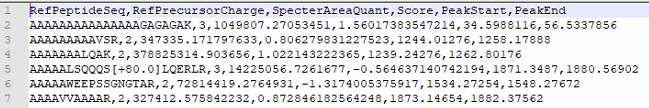


③在前面生成的文件夹下，同时创建了一个新的文件夹，文件夹的名字为：库文件的名字 + “\_”+ PPM的值。例如，对于前面的示例库文件：“HEKAndP100HeavyLib.blib”，以及PPM默认值20，会生成一个文件夹“HEKAndP100HeavyLib\_20ppm”。

④之后会在“HEKAndP100HeavyLib\_20ppm”这个文件夹下生成两个文件为：“Coeffs.csv”和“DecoyCoeffs.csv”。这两个文件格式一样，一个是使用原始图谱库计算得到的结果文件，另一个是使用诱饵库计算得到的结果文件。格式如下，对应信息为计算的非负系数、扫描索引、候选肽段、电荷值、前体m/z、前体保留时间、递归求解次数和相关系数。



⑤最后在“HEKAndP100HeavyLib\_20ppm”这个文件夹下生成最终结果文件“Quants.csv”。格式如下，对应信息为肽段序列、电荷、定量结果、分数、峰值起始和结束。



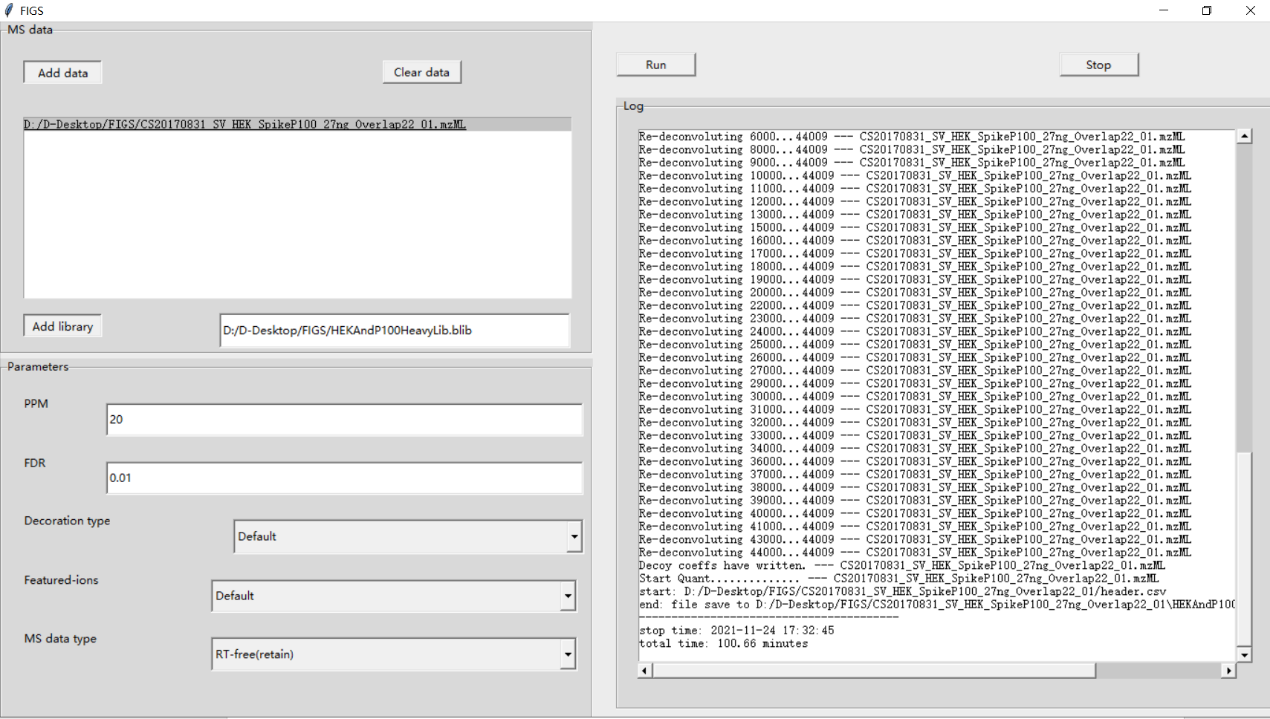
⑥除此之外，还会在源程序所在的文件夹下生成一个log.txt文件。其中记录了计算过程中的关键信息。



### 4.3.2计算结束标志

当一个质谱数据计算结束时，log日志框会输出“end: file save to······”，这里的省略号表示的是该质谱数据文件定量结果保存的路径。

当所有质谱数据都计算结束时，会输出 “stop time:······”和“total time:······”。分别表示结束时间和计算总共用时。



## 4.4多线程执行说明

软件默认使用多进程进行计算，对每个质谱数据文件创建一个新的进程进行计算。例如，当同时选择3个质谱数据进行计算时，本软件会创建3个进程，同时对这3个质谱数据进行定量计算。

所以，理论上只要电脑内存够大，cpu够快，可以同时选择很多个质谱数据文件，同时进行计算，并且这些文件基本上会同时计算结束，互不影响。这样计算一个质谱数据文件和计算很多个文件时间一样，可以使平均计算时间变少，但并不能改变一个文件的时间。