MetaboDirect

Release 0.2.1

Christian Ayala

Nov 03, 2022

CONTENTS

CHAPTER

ONE

INTRODUCTION

MetaboDirect is a Python and R based pipeline for the analysis of Direct Injection FT-ICR Mass Spectrometry data. The MetaboDirect pipeline takes a *report file*, generated using Formularity (Tolić et al., 2017), and a *sample information file* as inputs and automatically performs all the analysis described below including: sample filtering, m/z filtering, normalization of intensities, thermodynamic index calculation, annotation of molecular formulas using the KEGG database (Kanehisa & Goto, 2000), statistical analysis and construction of transformation networks using Cytoscape (Shannon et al., 2003).

1.1 Installation

MetaboDirect can be installed directly from PyPi using:

```
pip install metabodirect
```

Additionally it can be installed from source by cloning its GitHub repository:

```
git clone https://github.com/Coayala/MetaboDirect.git
cd MetaboDirect
python setup.py install
```

1.1.1 Required modules

MetaboDirect requires Python (3.5 and above), R (4 and above) and
Cytoscape (3.8 and above) with the following libraries/modules:

Python

- argparse
- numpy
- pandas
- seaborn
- more-itertools
- py4cytoscape

R

- tidyverse
- RColorBrewer
- vegan
- ggnewscale
- ggpubr
- ggvenn
- KEGGREST
- factoextra
- UpSetR
- pmartR (for normalization tests)
- SYNCSA

Cytoscape

- FileTransfer

1.2 Quick start

To quickly run the MetaboDirect pipeline use the following command.

```
metabodirect DATA_FILE METADATA_FILE -g GROUPING_VARIABLE
```

Check *metabodirect* in the User's Guide to learn more about the required *input data* and the analysis options that are offered.

Information about the arguments can be obtained using the -h/–help function.

```
metabodirect -h
```

1.2.1 Example using test data

Example data can be downloaded from the MetaboDirect repository `example` directory, or from the command line with:

```
# Report file
wget https://raw.githubusercontent.com/Coayala/MetaboDirect/main/example/Report.csv --no-
↪check-certificate

# Metadata file
wget https://raw.githubusercontent.com/Coayala/MetaboDirect/main/example/metadata.csv --
↪no-check-certificate
```

Try analizing example data using:

```
metabodirect Report.csv metadata.csv -o test -m 200 400 -g Habitat Depth -t -k
```

1.3 MetaboDirect pipeline

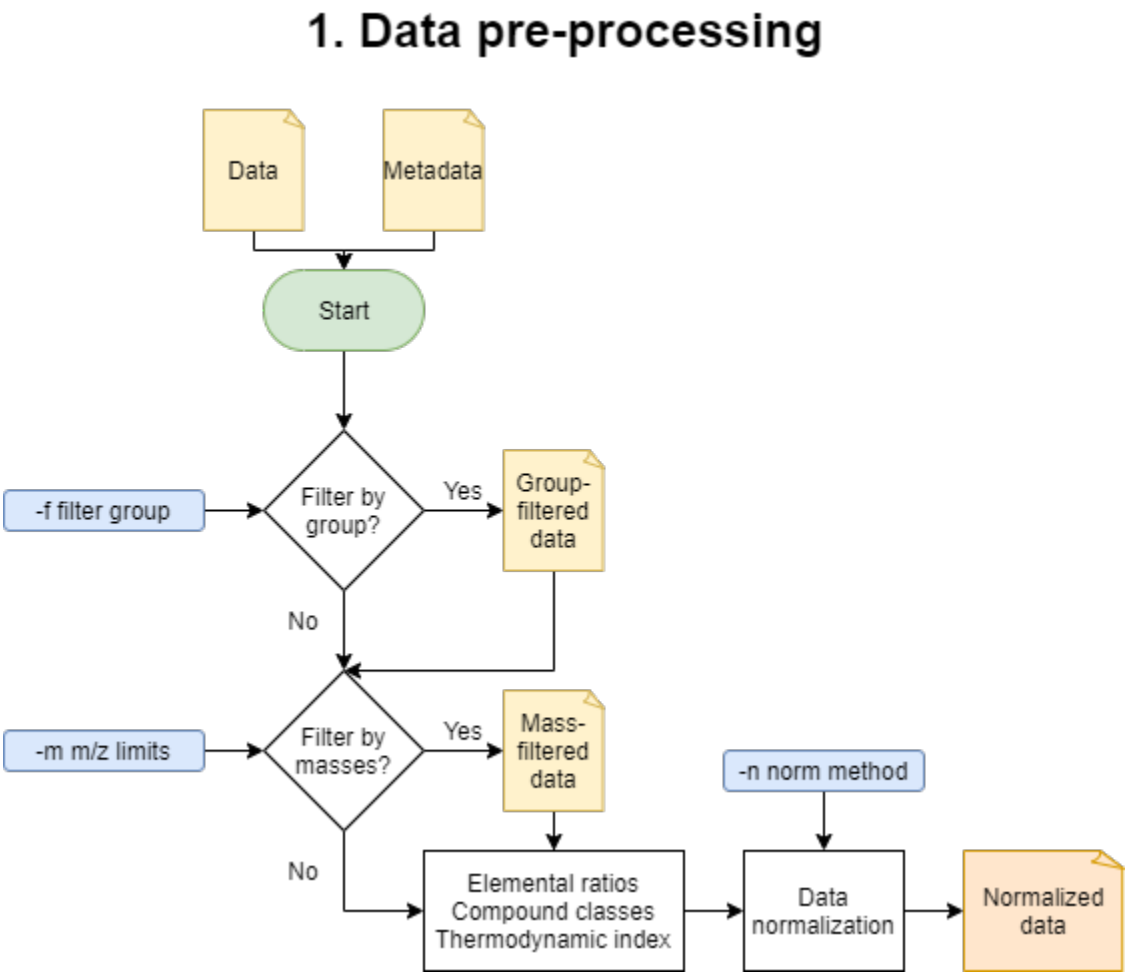
The **MetaboDirect** pipeline includes 5 major steps: data pre-processing, data diagnostics, data exploration, statistical analysis, and transformation network analysis, which can be run with the `metabodirect` command. For more information check *metabodirect* in the User's Guide.

Additionally, the `test_normalization` command can be run before the main **MetaboDirect** pipeline, to help deciding which normalization method to use for the analysis.

1.3.1 (Optional Step) Test normalization methods

The command `test_normalization` uses the Statistical Procedure for the Analysis of Normalization Strategies (SPANS) (Webb and Robertson et al., 2011), which has been previously demonstrated to work well with FT-ICR MS data (Thompson et al., 2021). This method systematically evaluates the effect of several normalization methods on the between-group variance structure to identify which method improves the structure of the data while introducing the less amount of bias. The SPANS method has been previously demonstrated to work well with FT-ICR MS data (Thompson et al., 2021). For more information check *test_normalization* in the User's Guide.

1.3.2 1. Data pre-processing

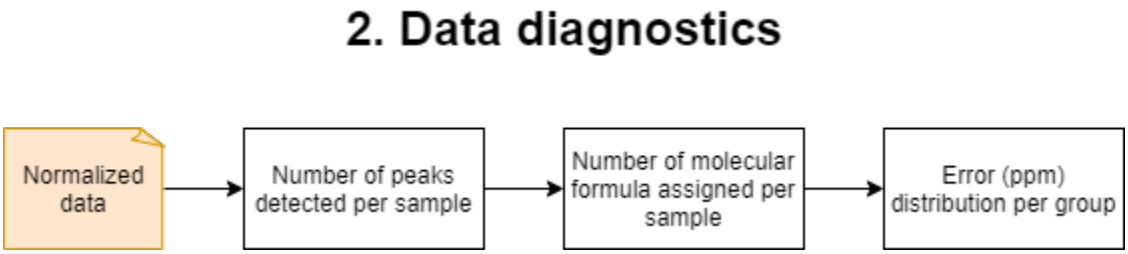


These is the beginning of the main **MetaboDirect** pipeline. During this step this step samples will samples will be filtered out based on the `-f` option. Detected peaks will also be filtered out inf the `-m` option was specificied. After filtering elemental ratios will be calculated. Compound classes will be defined based on elemental ratios, the boundaries to determine compound classes can be found in the file: compound_class_table.csv. Thermodynamic indices are calculated based on the assigned molecular formula as follows:

Index	Formula
Nominal Oxidation State of Carbon (NOSC)	$NOSC = \frac{4C+H-3N-2O+5P-2S}{C} + 4$
Gibbs Free Energy (GFE)	$GFE = 60.3 - 28.5 * NOSC$
Double Bond Equivalent (DBE)	$DBE = 1 + 0.5(2C - H + N + P)$
Aromatic Index (modified) (AI_mod)	$AI = 1 + C - 0.5O - S - \frac{0.5(H+P+N)}{C-0.5O-S-N-P}$

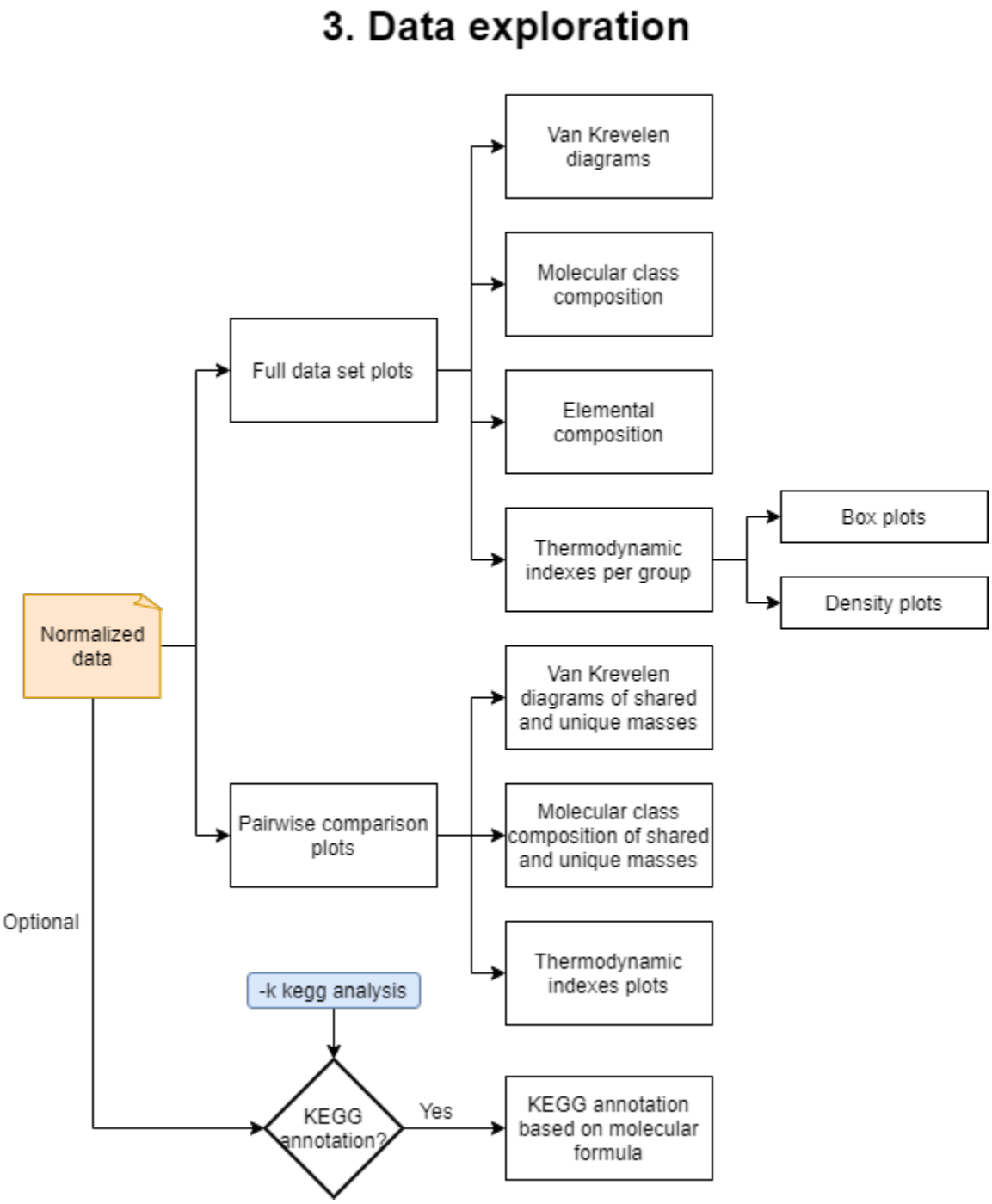
Data will be normalized during this step to be used in all of the subsequent analysis.

1.3.3 1. Data diagnostics



In this step plots with the number of detected peaks on each sample and the number of molecular formulas that were assigned per each sample will be generated. In addition the error (in ppm) during the formula assigment will also be plotted based on the grouping variables.

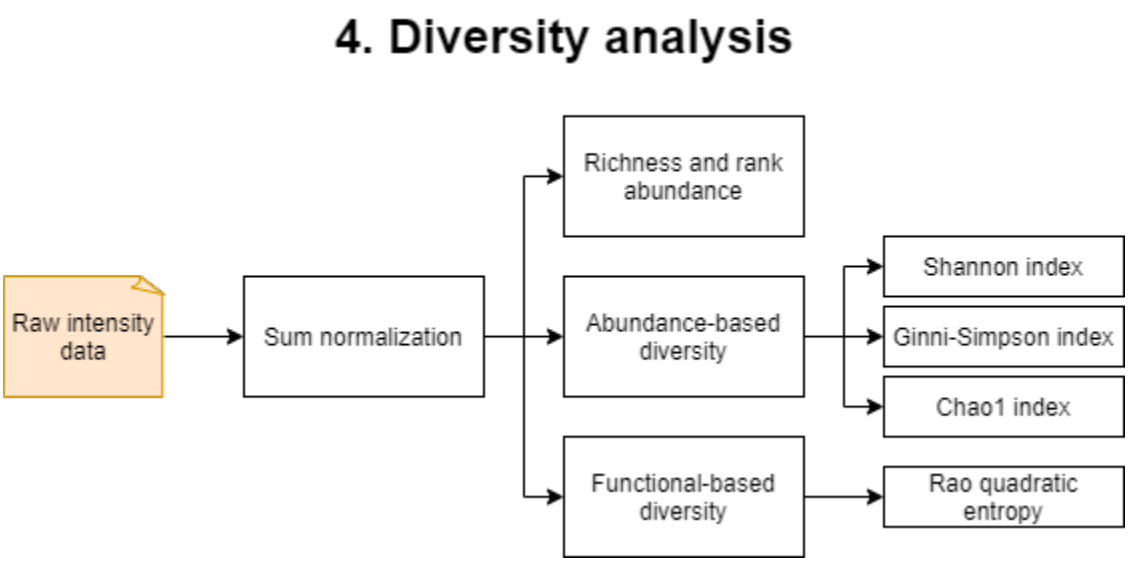
1.3.4 3. Data exploration



This step will generate and run an R script named **data_exploration.R**. This part of the analysis will generate plots for the elemental and molecular compound class composition, Van Krevelen diagrams of the detected masses, and violin and density plots of the previoulsy mentioned thermodynamic indices. In addition, it will generate directories that contain plots of the pairwise comparisons among the different values in the specified grouping variables.

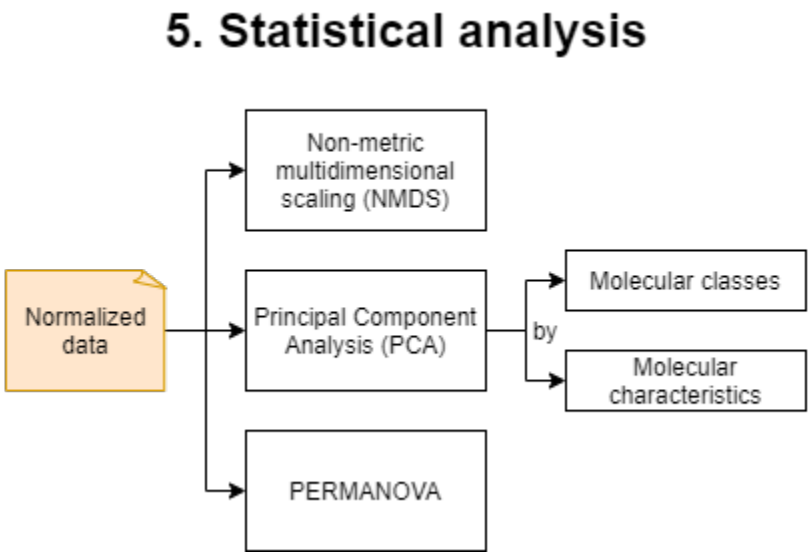
If the option `-k` was selected another R script named **KEGG_annotation.R** will be created. It will produce an additional .csv file with KEGG annotations of the detected masses based on the molecular formula.

1.3.5 4. Diversity analysis



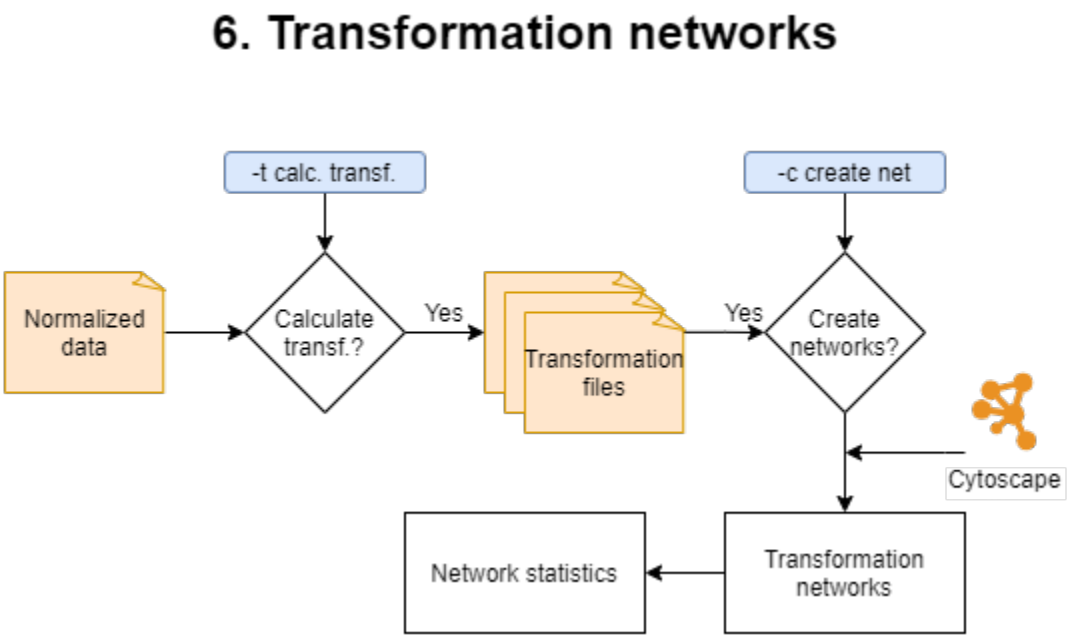
This step will generate and run an R script named **data_chemodiversity.R**. For this analysis, raw intensity data will be sum-normalized and used to obtain several diversity metrics. Diversity metrics generated include species (metabolite) richness and rank abundance. Abundance-based diversity is measured with the *Shannon* diversity index, the *Gini-Simpson* index and the *Chao1* index. Functional based diversity, based on the compounds elemental composition, reactivity and insaturation/aromaticity is measured with the Rao's quadratic entropy index.

1.3.6 5. Statistical analysis



This step will generate and run an R script named **data_statistics.R**. During this step a Permutational multivariate analysis of variance (PERMANOVA) test will be applied to the dataset. Multiple plots with sample ordination will be generated. Non-metric Multidimensional Scaling (NMDS) plots are based on the normalized peak intensity, while Principal Component Analysis (PCA) plots are based on the molecular class composition and the molecular characteristics of the sample.

1.3.7 6. Transformation networks



This step is optional since it is time consuming and it requires that Cytoscape is installed on the machine. It comprises two phases:

1) Calculation of the transformations. During this phase mass differences between all peaks from a given sample are calculated. Mass differences matching trasformations specified in the **biochemical_transformation_key** (transf_key.csv) are retained to become the edges for the transformation networks. A different biochemical transformation key can be provided when running MetaboDirect using the `-b` option. Once transformation are calculated, transformation networks will be generated using Cytoscape. Edge/transformation files will be saved in the `./5_transformations/transf_by_sample` directory inside the **results** directory. MetaboDirect can be stopped in this phase if you want to build the transformation networks at a later time.

2) Creation of transformation networks. During this phase, transformation networks will be built in Cytoscape using the mass differences previously calculated. If edge files were previously calculated, the command `create_networks` can be used to build the networks based on those files. For more information check *create_networks* in the User's Guide. These networks will have the nodes colored based on the compound classes determined during the **pre-processing** step. Finally an R script named **network_stats.R** will be generated and run to plot network statistics.

References

- Hill, M. O. (1973). Diversity and evenness: a unifying notation and its consequences. Ecology, 54(2), 427-432.
- Thompson, A. M., Stratton, K. G., Bramer, L. M., Zavoshy, N. S., & McCue, L. A. (2021). Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR-MS) peak intensity normalization for complex mixture analyses [https://doi.org/10.1002/rcm.9068]. Rapid Communications in Mass Spectrometry, 35(9), e9068. https://doi.org/https://doi.org/10.1002/rcm.9068
- Webb and Robertson, B. J. M., Matzke, M. M., Jacobs, J. M., Pounds, J. G., & Waters, K. M. (2011). A statistical selection strategy for normalization procedures in LC-MS proteomics experiments through dataset-dependent ranking of normalization scaling factors. Proteomics, 11(24), 4736-4741.

1.4 User's Guide

1.4.1 metabodirect

MetaboDirect requires two files, a report from Formularity (data file) and a sample information file (metadata file), and at least one grouping variable (defined in the metadata file).

Information about the arguments can be obtained using the `-h/--help` function. A detailed description of each argument is provided below.

```
metabodirect -h
```

```
```

usage: metabodirect [-h] -g STR [STR ...] [-o OUTDIR] [-m FLOAT FLOAT] [-p INT] [-e FLOAT] [-f STR STR] [-b STR] [-k]
[-v] [-n STR] [–norm_subset STR] [–subset_parameter FLOAT] [–log_transform] [-t] [-c] DATA METADATA

Program for running all the MetaboDirect analysis pipeline

positional arguments:

DATA Name of the file with the Direct Injection MS data in .csv format METADATA Name of the file with the sample information (metadata) in .csv format

optional arguments:

-h, --help show this help message and exit

-g STR [STR ...], –group STR [STR ...]
Grouping variables for coloring and faceting figures (Max 2) (default: None)

-o OUTDIR, --outdir OUTDIR Output directory (default: MetaboDirect_output)

-m FLOAT FLOAT, –mass_filter FLOAT FLOAT
Range to filter m/z data (min_mz, max_mz). The pipeline will not filter m/z values by default (default: None)

-p INT, --peak_filter INT Minimum number of samples a peak must be present to be conserved for the analysis (default: 2)

-e FLOAT, --error_filter FLOAT Max error (e) allowed in formula assignment. Peaks with **|error|** > e will be removed from the analysis (default: 0.5)

-f STR STR, –filter_by STR STR
Filter samples based on metadata. First enter the name of the feature, followed by the values associated with the samples you want to keep in the analysis.(Example -f Habitat Bog,Palsa) (default: None)

-b STR, --biochem_key STR File with the biochemical key to use for the transformation network (default: Default key)

-k, --kegg_annotation Set this option to perform annotation of the molecular formulas usingthe KEGG database (default: False)

-v, --version show program's version number and exit

Normalization methods:

Options to define how data normalization will be carried out

-n STR, --norm_method STR Available methods to normalize data are: 'mean', 'median', 'zscore', 'sum', 'max', 'minmax', 'binary', 'none' (default: max)

--norm_subset STR Subset of the data to use for normalization purpouses. Available subset methods: ALL, LOS, PPP. LOS uses peaks in the top L order statistics, PPP uses peaks having a minimum percentage of observed values. (default: ALL)

--subset_parameter FLOAT If using a sample subset for nomalization, this parameter defines the subsample of peaks that will be used for normalization. If not defined, the default values will be 0.3 for LOS and 0.5 for PPP (default: None)

--log_transform Set this option to log transform the data. (Program will fail if there are peaks with intensities of 0. Consider tranforming this values into 1 if log transformation is desired (default: False)

Transformation network options:

Options to control wheter transformations will be calculated and if networks will be constructed

-t, --calculate_transformations Set this option to calculate transformations based on biochemical key (default: False)

-c, --create_networks Set this option to build transformation networks based on transfomations calculatedwith the biochemical key (this options turns -t automatically) (default: False)

Input data file (`DATA`)

The input data for **MetaboDirect** is the .csv report file generated by default by Formularity. If Formularity was not used, any data file can be arranged to have columns with the **exact same names** that are shown in the table below.

Mass	C	H	O	N	C13	S	P	Na	El_comp	Class	Neutral_Mass	Error_ppm	Candidates	*Sample1*	*Sample2*	*Sample3*	…
Mass1	4	4	4	0	0	0	0	0		NA	111.634709		NA	6.237	0	0	
Mass2	5	2	2	0	0	0	0	0		NA	111.712005		NA	0	6.343	6.166	
Mass3	3	6	2	2	0	1	0	0		NA	112.125106		NA	7.549	7.363	6.75	
Mass4	5	6	3	0	0	0	1	0		NA	112.395704	0.45	NA	0	0	6.145	
Mass5	6	2	3	0	0	0	1	0		NA	112.457004	3	NA	0	6.133	0	
…																	

Mass1, Mass2, … refer to the m/z values detected by the software (i.e. each peak) while *Sample1, Sample2, ..* refer to the name of each sample. An example dataset is included in the MetaboDirect repository `example` directory.

Sample information file (`METADATA`)

The sample information file (or metadata file) is a .csv file that has one column called *SampleID* with the names of all of the samples that are present in the report file. Please make sure that the sample names in the *input data* and the *sample information file* are **exactly the same**. At least one other column must be present in the sample information file and must contain information used to group the data for plotting and for the statistical analysis. Multiple grouping variables can be present in this file but only two can be used simultaneously in **MetaboDirect**. When running the pipelines the grouping variables can be defined with the *-g* option using the **exact name** that it is on this file. Additionally, please use only letters (Aa-Zz), numbers (0-9) and underscores (_) for both the **sample names** and the **grouping variables**. An example file is included in the `example` folder with the name *metadata.csv*.

SampleID	Grouping_var1	Grouping_var2	Grouping_var3
Sample1	A	M	X
Sample2	A	N	Y
Sample3	B	M	Y
Sample4	B	N	X
Sample5	A	N	Z

Output directory (`-o` | `--outdir`)

The name of directory where all the generated plots, tables and scripts will be saved. If it is not defined the directory will be named MetaboDirect_output by default.

Grouping variable (`-g` | `--group`)

This option accepts up to two grouping variables (e.g. `-g Grouping_var1` or `-g Grouping_var1 Grouping_var2`) whose names are **exactly the same** as they appear in the columns of the metadafile. The first grouping variable will be used for giving colors to the plots generated. Both variables will be used for the statistical analysis and the pairwise comparisons.

Filter samples (`-f` | `--filter_by`)

This option takes two arguments: **1)** a variable from the metadata file and **2)** values from that variable column that we want to keep in the analysis. For example `-f Grouping_var3 X`, will keep just the samples for whom the Groupin_var3 is equal to "X". Multiple values for the same variable can be defined separated by commas (without spaces) (i.e. `-g Grouping_var3 X,Z`).

Mass filter (`-m` | `--mass_filter`)

This option takes two arguments: lower and an upper m/z limits. Peaks with m/z (masses) outside of its limits will be filtered out and not considered in the analysis.

Peak filter (`-e` | `--error_filter`)

This option is to determine the maximum error that is allowed from formula assignment.

Error filter (`-p` | `--peak_filter`)

This option is for specified the minimum number of samples a peak must be present to be conserved for the analysis.

Normalization method (-n | --norm_method)

This option defines which normalization method will be used to normalize the intensities (I). It can take one of the following options for i samples and j peaks. Normalization methods are based on the ones used by Kitson, et al. (2021) and Thompson, et al. (2021):

Normalization method	Formula
max	$NormIntensity_{i,j} = \frac{I_{i,j}}{max(I)_i}$
minmax	$NormIntensity_{i,j} = \frac{I_{i,j} - min(I)_i}{max(I)_i - min(I)_i}$
mean	$NormIntensity_{i,j} = \frac{I_{i,j} - mean(I)_i}{max(I)_i - min(I)_i}$
median	$NormIntensity_{i,j} = \frac{I_{i,j} - median(I)_i}{max(I)_i - min(I)_i}$
sum	$NormIntensity_{i,j} = \frac{I_{i,j}}{\sum I_i}$
zscore	$NormIntensity_{i,j} = \frac{I_{i,j} - mean(I)_i}{std.dev(I)_i}$
none	$NormIntensity_{i,j} = InputData_{i,j}$

Normalization subset method (--norm_subset)

If a normalization method other than binary or none is selected it is possible to use only a fraction of the peaks to calculate the normalization factors (normalization will still be applied to all the dataset). Possible subset methods are :

Subset method	Description
ALL	Use all present peaks to calculate normalization factors
LOS	Use a percentage of peaks in the top L order statistics
PPP	Uses peaks that are present in more than minimum percentage of samples

The option --subset_parameter defines the percentage of peaks that will be used in LOS or the minimum percentage of samples that a peak must be present for PPP.

Subset parameter (--norm_subset)

This option is only needed when LOS of PPP are selected as normalization methods. It defines either the minimum percentage of samples a peaks need to bre present to be considered (PPP) or the percentage of top peaks that will be used (LOS).

KEGG annotation (-k | --kegg_annotation)

This is an optional step as it may take a long time (~ couple of hours) depending on the number of peaks present in the data. If this option is present, peaks will be annotated with the KEGG database (Pathway, Module, Brite, etc.) based on their molecular formula.

Calculate transformations (-t | --calculate_transformations)

This option define whether or not a molecular transformations between the peaks will be calculated based on their mass differences. If this option is selected, **MetaboDirect** will end after generating the transformation files. Transformation files will be located in `./$outdir/6_transformations/transf_by_sample`.

Create networks (-c | --create_networks)

If this option is selected, it will automatically turn on the option `-t`. After the transformation files are generated, transformation networks will be built. This step requires Cytoscape (version 3.8 and above) to be installed in the machine. **MetaboDirect** will ask the user to open Cytoscape when required in order to construct the networks. When prompted in the screen, please open Cytoscape and then hit enter to continue with the analysis.

1.4.2 test_normalization

This a companion script that can be used to help choosing the best normalization method for the data using the SPANS method.

Information about the arguments can be obtained using the `-h/--help` function. A detailed description of each argument is provided below.

```
test_normalization -h
```

```
usage: test_normalization [-h] [-f STR STR] [--log_transform] DATA METADATA GROUP

Program for running all the MetaboDirect analysis pipeline

positional arguments:
DATA                  Name of the file with the DI-MS data in .csv format
METADATA              Name of the file with the sample information (metadata) in tabular␣
↪format
GROUP                 Grouping variables to test for normalization significance

optional arguments:
-h, --help            show this help message and exit
-f STR STR, --filter_by STR STR
                                                  Filter samples based on metadata. First␣
↪enter the name of the feature,followed by the values
                                                  associated with the samples you want to␣
↪keep in the analysis.(Example -f Habitat Bog,Palsa)
                                                  (default: None)
--log_transform       Set this if you plan to log transform your data before␣
↪normalization (default: False)
```

Input data file (DATA)

The same input data that will be used for **MetaboDirect**. A .csv report file generated by default by Formularity. If Formularity was not used, any data file can be arranged to have columns with the **exact same names** that are shown above for *metabodirect*.

Sample information file (METADATA)

The same input data that will be used for **MetaboDirect**. A .csv file that has one column called *SampleID* with the names of all of the samples that are present in the report file. Please make sure that the sample names in the *input data* and the *sample information file* are **exactly the same**.

Grouping variable (GROUP)

The grouping variable that will be tested for significance . It names should be **exactly the same** as they appear in the columns of the metadafile.

Filter samples (-f | --filter_by)

This option takes two arguments: **1)** a variable from the metadata file and **2)** values from that variable column that we want to keep in the analysis. For example `-f Grouping_var3 X`, will keep just the samples for whom the Groupin_var3 is equal to "X". Multiple values for the same variable can be defined separated by commas (without spaces) (i.e. `-g Grouping_var3 X,Z`).

Log transform data (--log_transform)

If this option is used, data will be log transformed before testing for significance.

1.4.3 create_networks

This a companion script that can be used to build the transformation networks using the transformation files created with the option `-t`.

Information about the arguments can be obtained using the `-h/--help` function. A detailed description of each argument is provided below.

```
create_networks -h
```

```
usage: create_networks [-h] OUTDIR METADATA STR [STR ...]

Program for creating molecular transformation networks, based on previously calculated␣
↪transformations

positional arguments:
OUTDIR      Output directory used to create networks with metabodirect and the -t option
METADATA    Metadata file used in the analysis, if a filtered metadata was generated␣
↪please enter that one
GROUP         Grouping variables for coloring and faceting figures (Max 2)
```

(continues on next page)

(continued from previous page)

```
optional arguments:
-h, --help  show this help message and exit
```

Output directory (OUTDIR)

It needs to be the same output directory that was used during the original run of the **MetaboDirect** pipeline with the `-t` option.

Sample information file (METADATA)

The same metadata file that was used during the original run of the **MetaboDirect** pipeline with the `-t` option.

Grouping variable (GROUP)

The grouping variable that will be used to compare the network statistics. It names should be **exactly the same** as they appear in the columns of the metadafile.

References

- Kitson, E., Kew, W., Ding, W., & Bell, N. G. A. (2021). PyKrev: A Python Library for the Analysis of Complex Mixture FT-MS Data. Journal of the American Society for Mass Spectrometry, 32(5), 1263-1267. https://doi.org/10.1021/jasms.1c00064
- Thompson, A. M., Stratton, K. G., Bramer, L. M., Zavoshy, N. S., & McCue, L. A. (2021). Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR-MS) peak intensity normalization for complex mixture analyses [https://doi.org/10.1002/rcm.9068]. Rapid Communications in Mass Spectrometry, 35(9), e9068. https://doi.org/https://doi.org/10.1002/rcm.9068