**README file for NMR data:**

**Design\_file (metadata) – ‘Design\_file\_NMR\_CDCl3\_14Aug2020.txt’ – tab delimited**

The design file contains information as to the study design and analysis

It is generated from the analyst whom prepared the data and analyzed the samples

(Contact Amanda Shaver for further details: [amanda.shaver@uga.edu](mailto:amanda.shaver@uga.edu)).

Each column corresponds to the following:

|  |  |
| --- | --- |
| sampleID | Unique sample name matching the data containing file |
| set | Grouping of samples for NMR analysis and sample prep |
| batch | Grouping of samples for NMR analysis and sample prep |
| rack\_position | Position of samples in the NMR autosampler |
| run\_order | Order in which each sample was run |
| tube\_sample | NMR tube identifier |
| genotype | C. elegans strain of that sample |
| wormgrowth\_sample\_name | sample name for that particular aliquot/sample |
| Solvent | Solvent used for reconstitution after metabolite extraction |
| parameters | Type of data collected |

Each row corresponds to a sample as defined by the different column identifiers.

**Datamatrix – ‘ph\_NMR\_CDCl3 \_14Aug2020.txt’ – tab delimited**

The .txt file named “**ph\_NMR\_CDCl3 \_14Aug2020.txt**” specifies the type of data ph = Peak Height, the analytical platform, the solvent it was collected “CDCl3” = non-polar extraction, and the date when the .txt file was saved.

This file contains data from all the samples in the study, including quality controls.

Solvent Blanks and Process Blanks are **not** included.

The data matrix is composed of the following:

Columns:

Columns 1 to 3:

These columns are comprised of the following:

blank\_peak: A value > 0 corresponds to a peak being detected in that particular bucket in any of the blanks (solvent + process). Values >1 indicate the number of blank peaks detected in that particular bucket.

blank\_intensity\_flag: Binary values only (1 or 0). 1 - indicates that the maximum height of the blank peak multiplied by 3, is bigger than the average height of the samples. Standard threshold for blank feature cutoff. 0 – indicates that the peak is absent or smaller than the threshold.

align\_score: This alignment metric is based on the Euclidean distance of the peaks (from all the samples) within a bucket, to the median of all those peaks. A scale is created from these alignment scores. <5 – **Excellent**; 5-10 - **V.** **good**; 10-15 – **Good**; 15-20 – **OK**; 20-25 – **Caution**; 25-30 – **Inspect**; 30-40 - **Not** **great**; 40> - **Unaligned**

A picture containing screenshot

Description automatically generated

Columns 4 to end:

The remaining columns represent each sample peak height and correspond to the order of the “Design\_file”.

Each column name is defined as:

sample\_name\_aliquot = unique name for that particular sample

Instrument = Analytical platform it was run on (i.e. NMR, MS)

Solvent = Reconstitution solvent (i.e. CDCL3 = non-polar metabolites)

runorder = represented by a number corresponding to the order of which that sample was run.

Example:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| aos127\_ga\_ms2\_NMR\_CDCL3\_5 | | | | solera\_NMR\_CDCL3\_3 | | | | pooled\_pd1074\_NMR\_CDCL3\_143 | | | |
| sample\_name\_aliquot | Instrument | Solvent | Run  order | sample\_name\_aliquot | Instrument | Solvent | Run  order | sample\_name\_aliquot | Instrument | Solvent | Run  order |

Rows:

Each ppm (chemical shift) value -row names- corresponds to the center of one bucket. A

bucket is defined as the boundaries of a peak.

**Data processing:**

The data matrix enclosed has been obtained after the following transformations.

**Software used:**

**nmrPipe** (only main steps described)**:**

Fourier transform:

NMR process transform data collected in the time domain to frequency domain.

Phasing:

NMR process to adjust the phase angles needed to put the real spectrum

entirely in absorption mode and the imaginary spectrum entirely in dispersion

mode. Adjusts peak shape and baseline to certain extent.

Baseline correction:

polynomial filter to adjust baseline distortions (automatic selection and

parametrization)

**nmrPipe scripts are available on Metabolomics Workbench Study ID: ST002096**

**Matlab:**

Referencing:

All chemical shifts are matched to a known reference feature (CDCL3 @ 7.24

ppm)

Alignment:

Each feature in each spectrum is aligned to each other, small drifts are habitual

for NMR data, several methods are used to align these features.

Baseline correction:

Baseline Correction for NMR Spectroscopic Metabolomics Data Analysis, small

adjustments are made to the baseline, so the spectra noise is closer to “0.00“

intensity, and remove small baseline distortions per each spectrum.

Bucketing and Peak picking:

Buckets are defined by an algorithm and further refined manually to encapsulate

a peak.

Peak height is then defined as the maximum value of each peak picked within

those boundaries. The process is repeated for each spectrum using the same

boundaries for each iteration.

**Matlab data input:**

paths.processed folder should contain “2\_processed\_nmr\_cdcl3” found on

MWB ID: ST002096

paths.metadata folder should contain “master\_nmr\_run\_designfile.csv” found on

Github (https://github.com/artedison/metaanalysis.git).

**Matlab data output:**

ph\_NMR\_CDCL3\_14Aug202.txt

Design\_file\_NMR\_CDCL3\_14Aug202.txt

**The Matlab workflow is available on GitHub:**

**All functions used are described in detail in:** https://github.com/artedison/Edison\_Lab\_Shared\_Metabolomics\_UGA/wiki

(Contact Goncalo Gouveia for further details: [goncalog@uga.edu](mailto:goncalog@uga.edu))