NMR TOOLS
Pre-processing

Marie Tremblay-Franco
Cécile Canlet
Manon Martin

02/10/2018 v 1.0.0
NMR workflow

1. Read
2. TopSpin
3. Alignment
4. Bucketing
5. Normalize
6. Analyze
7. Annotate

Standard
- Bruker files
- Equal size
- Quantitative variable
- Univariate
- Total intensity
- PCA
- PLS-DA

Advanced
- Jcamp-dx
- Phase correction
- Baseline correction
- PQN
- ASICS

Legend
- available
- in progress

1 Preprocessing using TopSpin: Fourier transformation, phase correction, baseline correction and calibration (TSP: δ 0ppm)
2 Free Induction Decay
3 Fourier transformation
Galaxy NMR processing tools

Workflow4metabolomics

Current version: 3.0

Help and support: support@workflow4metabolomics.org

Latest news
- 10/05/2017 - LC-MS: A new tutorial video explain how to run xcmsSet in parallel on single files [link]
- 20/04/2017 - Workflow4Metabolomics v3.0 starts today - Check the changelog section below

Changelog
3.0.0 - 20/04/2017

LC-MS
- Preprocessing
  - UPGRADE - xcms.* (2.1.0): upgrade the xcms version from 1.44.0 to 1.46.0
  - NEW - xcms.* (2.1.0): The W4M tools will be able now to take as input a single file. It will allow to submit in parallel several files and merge them afterward using *xcms.xcmsSet Merger* before *xcms.group*.
  - BUGFIX - xcms.xcmsSet (2.1.0): the default value of "matchedFilter" -> "Step size to use for profile generation" which was 0.01 has been changed to fix with the XCMS default values to
MTBLS1 DATA DESCRIPTION

A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human

R. M. Salek,1* M. L. Maguire,1* E. Bentley,2 D. V. Rubtsov,2 T. Hough,3 M. Cheeseman,3 D. Nunez,4 B. C. Sweatman,4 J. N. Haselden,4 R. D. Cox,4 S. C. Connor,4 and J. L. Griffin4

1Department of Biochemistry, University of Cambridge, Cambridge; 2Mammalian Genetics Unit, Medical Research Council (MRC) Harwell, Oxfordshire; 3The Mary Lyon Centre, MRC Harwell, Harwell, Oxfordshire; and 4Safety Assessment, GlaxoSmithKline, Ware, Herts, United Kingdom

Submitted 5 September 2006; accepted in final form 15 December 2006
Experimental design

- 1H-NMR analysis of 131 urine samples from type II diabetic or control patients (http://www.ebi.ac.uk/metabolights/MTBLS1)
- W4M00004_MTBL51 reference workflow

42 volunteers:
- 12 healthy patients
- 30 T2DM patients

4-weeks washout period

| t0 | t1 | t2 | t3 | t4 | t5 | t6 | t7 |
Data description

• 134 midstream urine samples from human
  – 12 healthy volunteers (8 males/4 females): 7 time points of urine collection
  – 30 T2DM patients (17 males/13 females): 1-3 time points of urine collection

• NMR sample preparation\(^a\)

• NMR analysis:
  – Bruker DRX-700 spectrometer using a NEOSY pulse sequence\(^b\)
  – Fourier transformation was applied, then all spectra were phased and baseline corrected using TopSpin\(^f\)
HUMAN SERUM DATA DESCRIPTION
Experimental design and data description

- 1 **blood** sample was collected for 4 **different donors**
- For each blood sample, 8 sub-samples were measured across 8 days with one sub-sample of each donor per day and permutations according to a latin hypercube sampling method
- The total number of samples is then $4 \times 8 = 32$
- Data were acquired with a 500 MHz Bruker Avance spectrometer equipped with a TCI cryoprobe and using a CPMG relaxation-editing sequence with pre-saturation
- Spectra are labelled as: JxDx where Jx is the day of measurement and Dx is the donor label
- This dataset was designed to collect multiple measures from the same experimental unit and capture other technical sources of variation, allowing to compare the inter- and intra-unit variability.
PRE-PROCESSING TOOL
Reading and pre-processing tools in W4M

NMR Read

Description
Nuclear Magnetic Resonance Bruker files reading (from the PEPS-NMR R package (https://github.com/ManonMartin/PEPSNMR))

Workflow position
Upstream tools

<table>
<thead>
<tr>
<th>Name</th>
<th>output file</th>
<th>format</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

Downstream tools

<table>
<thead>
<tr>
<th>Name</th>
<th>Output file</th>
<th>Format</th>
</tr>
</thead>
</table>

Authors Manon Martin (manon.martin@lucilouvain.be) and Marie Tremblay-Franco (marie.trembley-franco@inra.fr; Galaxy integration)
Developed from the R package PepsNMR

https://github.com/ManonMartin/PepsNMR


- Read Bruker files
- Group delay correction
- Solvent residuals suppression
- Apodization
- Fourier Transform
- Referencing to 0 ppm
- Zero order phase correction
- Baseline correction

Included in W4M
Data importation (1)

The **NMR_Read tool**: import raw Bruker data from a zip file

Choose the zip file **MTBLS1.zip**

Specify the line in the title file to recover the FID names (usually in *pdata*/1/title)

Default value is line 1

Presence of subdirectories? **FALSE**

Select ‘FALSE’ when there is no subdirectories, ‘TRUE’ if there are subdirectories

Use (sub)directories names as FID names? **FALSE**

Select ‘TRUE’ to use the subdirectories names as the FID names (instead of looking in the title file)

**Execute**

**Authors**: Manon Martin (manon.martin@uclouvain.be) and Marie Tremblay-Franco (marie.tremblay-franco@inra.fr; Galaxy integration)

Input for the **Preprocessing tool**: formatted datasets as outputted by the NMR_Read tool

**NMR_Preprocessing**: Preprocessing of 1D NMR spectra (Galaxy Version 1.2.0)

Data matrix of FIDs

Choose **MTBLS1_sampleMetadata.tsv**

Sample metadata matrix

Choose **MTBLS1_sampleMetadata.tsv**
Data importation (2)

**Format:** .tabular (as outputted by NMR_Read) or .csv

- Field separator: tab
- Decimal separator: .

**Sample metadata** with the acquisition parameters:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TD</td>
<td>BYTORDA</td>
<td>DIGMOD</td>
<td>DECIM</td>
<td>DSPFVS</td>
<td>SW_h</td>
<td>SW</td>
<td>O1</td>
<td>DT</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ADG10003u_007</td>
<td>65536</td>
<td>1</td>
<td>1</td>
<td>12</td>
<td>12</td>
<td>14005.60224</td>
<td>20.01162551</td>
<td>3290.5</td>
<td>3.57E-05</td>
</tr>
<tr>
<td>3</td>
<td>ADG10003u_009</td>
<td>65536</td>
<td>1</td>
<td>1</td>
<td>12</td>
<td>12</td>
<td>14005.60224</td>
<td>20.01162551</td>
<td>3290.5</td>
<td>3.57E-05</td>
</tr>
<tr>
<td>4</td>
<td>ADG10003u_015</td>
<td>65536</td>
<td>1</td>
<td>1</td>
<td>12</td>
<td>12</td>
<td>14005.60224</td>
<td>20.01162551</td>
<td>3290.5</td>
<td>3.57E-05</td>
</tr>
<tr>
<td>5</td>
<td>ADG10003u_017</td>
<td>65536</td>
<td>1</td>
<td>1</td>
<td>12</td>
<td>12</td>
<td>14005.60224</td>
<td>20.01162551</td>
<td>3290.5</td>
<td>3.57E-05</td>
</tr>
<tr>
<td>6</td>
<td>ADG10003u_021</td>
<td>65536</td>
<td>1</td>
<td>1</td>
<td>12</td>
<td>12</td>
<td>14005.60224</td>
<td>20.01162551</td>
<td>3290.5</td>
<td>3.57E-05</td>
</tr>
</tbody>
</table>

- **TD:** Time domain size
- **BYTORDA:** Determine the endianness of stored data. If 0 -> Little Endian; if 1 -> Big Endian
- **DIGMOD:** Digitization mode
- **DECIM:** Decimation rate of digital filter
- **DSPFVS:** DSP firmware version
- **SW_h:** DSP width in Hz
- **SW:** Sweep width in ppm
- **O1:** Spectrometer frequency offset
- **DT:** Dwell time in microseconds
Preprocessing tool: incentive

- An inadequate pretreatment will never be compensated by a powerful data analysis

  ⇒ Some crucial steps must be carefully performed before the statistical analysis

- Metabolomics studies can have several hundreds of samples

- Manual preprocessing is time-consuming

- In contrast, the **Preprocessing tool** in W4M:
  - Works semi-automatically ⇒ gain in time
  - Includes advanced methods (solvent suppression, baseline correction, etc.)
  - Has its code in open source (GitHub)
Preprocessing tool: overview

Steps

1. Group Delay suppression
2. Solvent suppression
3. Apodization (increase the resolution/sensitivity)
4. Fourier Transform
5. Zero order phase correction
6. Referencing to 0 ppm
7. Baseline correction
8. Negative values put to zero
Phase shifts

- Spectra have a mix of absorptive and dispersive signals
  ⇒ The phase must be corrected
- Crucial step: the integration of signals depends on an adequate phase correction
1. Group Delay suppression

- Zero ($\varphi_0$) and first order ($\varphi_1(\nu)$) phase shifts: $S = S_{phased}e^{i(\varphi_0 + \varphi_1(\nu))}$

- Presence of a **digital filter**
  - The first tens of points in the FID are not part of the recorded signal and are called the **group delay**
  - Since the phase shift differs across signals, it introduces a first order phase shift linearly related to $\nu$

- The Bruker pre-acquisition $\tau$ is known from the acquisition parameters

- Procedure
  - Apply Fourier transform on the FID
  - Remove the delay $\tau$
  - Apply inverse Fourier transform on the spectrum to recover the FID
2. Solvent suppression

- Solvent residuals signal is problematic:
  - Highly variable
  - Not of interest & can mask useful information
  - Affects other pre-processing steps (baseline correction, phase corrections, etc.)

- **Hypothesis**: solvent is the main component in the samples

- **Principle**:
  - Non-parametric estimation of the solvent signal ($W$) with a penalized smoother (Eilers, 2003):
    
    $$\text{Min} \ \text{Least Squares criterion} + \lambda \times \text{roughness penalty}$$
    
    - Remove the estimated solvent signal from the FIDs

- **Meta-parameter**: smoothing parameter lambda
  - Determine how smooth is the solvent signal: $\lambda \uparrow \Rightarrow$ smoother signal
Illustration of solvent residuals suppression

FID before solvent suppression (zoom)

FID after solvent suppression (zoom)

Spectrum with solvent suppression

Spectrum without solvent suppression
3. Apodization

- **Purpose**: improve the sensitivity
- **Procedure**: multiply the FID by a factor called a **weighting function**
- Several classes of factors in W4M:
  - Negative exponential
  - Gaussian
  - Hanning
  - Hamming
  - Cos2

- **Example**: the negative exponential function:
  \[ W(t) = \exp(-LB \times t) \text{ with } LB : \text{line broadening factor} \]
Example of apodization with a negative exponential

No apodization

Exponential
LB=0.3

Exponential
LB=1

Exponential
LB=5
4. Fourier Transform and ppm scale

- Mathematical function to convert the signal in the time domain into a spectrum in the frequency domain.
- Frequencies are expressed in Hertz.
- Hertz units are dependant of the external magnetic field.
- Frequencies in Hertz are re-expressed as chemical shift $d$ in parts-per-million (ppm), a dimensionless scale.
5. Zero order phase correction

- The axis along which the signal is recorded cannot be predicted
  => arbitrary phase $\varphi_0$

- **Hypothesis:**
  - the real part of the spectrum should be in pure absorptive mode (with strictly (+) intensities)

- **Correction principle:**
  - Rotate the spectrum with a series of angles
  - Measure a positiveness criterion on the real part of the spectrum
  - Select the angle that maximises this criterion

- **Positiveness criteria:**
  - rms: $\frac{\sum([\text{positive intensities}]^2)}{\sum([\text{intensities}]^2)}$
  - max: maximum positive intensity
6. Referencing to 0 ppm

- **Incentive**: a known standard called an internal reference compound (TMS or TSP) is usually added to the samples to refine the scale calibration
  
  $\Rightarrow$ The chemical shift is now defined relative to this reference

- A ppm value is attributed to the reference peak (usually 0 ppm) and spectra are aligned to this peak
7. Baseline correction

- For a good integration, baseline should be flat with no distortion ⇒ Baseline artefacts need to be removed

- **Principle:**
  1. Non-parametric estimation of the baseline (Z) for the real spectrum with AsLS
  2. Removal of the estimated baseline from the real part of the spectrum

- **Asymmetric Least Squares smoothing (AsLS) algorithm** (Eilers & Boelens, 2005):

  \[
  \text{Min } \frac{1}{2} \sum \left[ (Z - y)^2 p + (y - Z)^2 (1-p) \right] + \lambda \text{ roughness}
  \]

  - Uneven weights \( p \) for (+) and (-) deviations of \( Z \) from the spectrum will favour positive corrected intensities
  - Roughness penalty

- **Meta-parameters:**
  - **Asymmetry parameter** (\( p \)): = weight for (-) deviations, if \( p \uparrow \) => the estimated baseline will more frequently be above the spectrum. Default value = 0.05
  - **Smoothing parameter** (\( \lambda \)): \( \lambda \uparrow \) => smoother signal
8. Negative values are put to zero

- After the baseline correction step, the remaining negative values are replaced by 0 since they cannot be interpreted.
- Final spectra are ready for further alignment, bucketing, normalization and data analysis steps.
Preprocessing tool in W4M (1)

Data importation:
- Data matrix: spectral intensities
- Sample metadata: acquisition parameters

First order phase correction

Solvent suppression
- Smoothing parameter: lambda

Apodization
- Method: window function
Preprocessing tool in W4M (2)

Data importation:
- Data matrix: spectral intensities
- Sample metadata: acquisition parameters

First order phase correction

Solvent suppression
- Smoothing parameter: lambda

Apodization
- Method: window function

Line broadening
Preprocessing tool in W4M (3)

Fourier Transform

- Display the Fourier transformed spectra?
  - NO
  - Select 'YES' to display the spectra or 'NO' to not display them

Zero order phase correction

- Method: positiveness criterion
- Exclusion area(s): one or several exclusion areas can be specified for the computation of the positiveness criterion
Preprocessing tool in W4M (4)

Shift referencing
- Search zone: define the area where to look for the reference compound peak
- shiftHandling: how to deal with spectral borders during realignment: 0 filling, cut, etc.
Preprocessing tool in W4M (4)

Shift referencing
- Search zone: define the area where to look for the reference compound peak
- shiftHandling: how to deal with spectral borders during realignment: 0 filling, cut, etc.
Shift referencing
- Search zone: define the area where to look for the reference compound peak
- shiftHandling: how to deal with spectral borders during realignment: 0 filling, cut, etc.
Preprocessing tool in W4M (4)

Shift referencing
- Search zone: define the area where to look for the reference compound peak
- shiftHandling: how to deal with spectral borders during realignment: 0 filling, cut, etc.

**Shift Referencing:**
- **Definition of the search zone**
  - Near value:
  - Percentage of the ppm axis around the ppm value to look for the reference compound peak: 0.02
    - Default value is 0.02

- **Shift Referencing: shiftHandling**
  - Zero filling
    - How to deal with shifts between spectra on their left and right sides

- **Shift Referencing: the value of the reference peak in ppm**
  - 0.0
    - Default value is 0

- **Display the spectra after Shift Referencing?**
  - NO
    - Select 'YES' to display the spectra or 'NO' to not display them
Shift referencing
- Search zone: define the area where to look for the reference compound peak
- shiftHandling: how to deal with spectral borders during realignment: 0 filling, cut, etc.
Preprocessing tool in W4M (4)

Shift referencing
- Search zone: define the area where to look for the reference compound peak
- shiftHandling: how to deal with spectral borders during realignment: 0 filling, cut, etc.

![Shift Referencing Interface](image-url)
Preprocessing tool in W4M (5)

Baseline correction
- Smoothing parameter: lambda
- Asymmetry parameter: p

Negative values put to zero

Final spectra
Results : graphs (1)

Spectrum after Fourier transform

- First order phase correction
- Solvent suppression with lambda = $10^6$
- Apodization : exp with lb=0.3
- Fourier transform
Results: graphs (2)

Spectrum after zero order phase correction and chemical shift referencing

- zero order phase correction: method rms
- Shift referencing: nearvalue; zerofilling; 0 ppm
Results: graphs (3)

Spectrum after baseline correction

- Baseline correction: \( \text{lambda} = 100000; p=0.05; \) numerical precision = \( 10^{-8} \)
Results : graphs (4)

Final preprocessed spectrum

Negative values are put to 0
ALIGNMENT TOOL
Alignment: why?

- **Variation of chemical shifts between samples**
  - Chemical shift depends on the pH
  - For urine samples, the pH can vary greatly
  - Addition of phosphate buffer is not sufficient to ensure the same pH in all samples (500 µl of urine and 200 µl of phosphate buffer pH 7)

- **Problems with bucketing**
  - A bucket can correspond to different metabolites in the different spectra if the chemical shift is different between spectra
  - The statistical results may be affected by this problem
Example: superposition of urine NMR spectra

Problematic for the bucketing: the bucket at 2.53 ppm corresponds to citrate in the red spectrum and noise in the other spectra.
• Vu et Laukens. (2013): classification of identified algorithms
Vu et Laukens. (2013): some of identified algorithms in their review have been compared in different papers.

<table>
<thead>
<tr>
<th>Short Name</th>
<th>Full Name</th>
<th>Reference</th>
<th>Technique</th>
<th>Target Function</th>
<th>Peak Picking?</th>
<th>Number of Parameters</th>
<th>Original Applied Data</th>
<th>Segment-Wise?</th>
<th>Pair-Wise?</th>
<th>Correction Method</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAFFT</td>
<td>Recursive alignment by FFT</td>
<td>[19]</td>
<td>FFT + Recursive segmentation model from global to local</td>
<td>FFT cross-correlation</td>
<td>No</td>
<td>1 (max. allowable shift)</td>
<td>Chromatographic data</td>
<td>Yes</td>
<td>Yes</td>
<td>Shift</td>
<td>(3)</td>
</tr>
<tr>
<td>icoshift</td>
<td>interval correlation shifting</td>
<td>[10]</td>
<td>Segmentation model by equal size segments or manually selecting segments</td>
<td>FFT cross-correlation</td>
<td>No</td>
<td>2 (the number of intervals or the length of interval $l$, max. allowable shift)</td>
<td>1D NMR</td>
<td>Yes</td>
<td>Yes</td>
<td>Shift &amp; Insert and deletion</td>
<td>(6)</td>
</tr>
<tr>
<td>ChuPA</td>
<td>hierarchical Cluster-based Peak Alignment</td>
<td>[8]</td>
<td>Segmentation model by hierarchical clustering</td>
<td>FFT cross-correlation</td>
<td>Yes</td>
<td>1 (max. allowable shift)</td>
<td>1D NMR</td>
<td>Yes</td>
<td>Yes</td>
<td>Shift</td>
<td>(7)</td>
</tr>
<tr>
<td>COW</td>
<td>Correlation Optimized Warping</td>
<td>[12]</td>
<td>Dynamic programming</td>
<td>Pearson correlation coefficient</td>
<td>No</td>
<td>2 ($m$: length of segments and $r$: slack or the max. allowable shift)</td>
<td>Chromatographic data</td>
<td>Yes</td>
<td>Yes</td>
<td>Insert and deletion</td>
<td>(1)</td>
</tr>
<tr>
<td>RUNAS</td>
<td>Recursive Unreferenced Alignment of Spectra</td>
<td></td>
<td>Intensity Weight slope transform + recursive alignment</td>
<td>FFT cross-correlation</td>
<td>No</td>
<td>1D NMR</td>
<td>Yes</td>
<td>No</td>
<td>Shift</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CluPA algorithm (1)

- Vu et al. (2011): Cluster-based Peak Alignment
- Three-steps algorithm
  - **Peak peaking**: spectra denoising, baseline correction and peak selection: R package based on wavelet transform, developed for mass spectra
Peak picking algorithm
CluPA algorithm (2)

- Vu et al. (2011): Cluster-based Peak Alignment
- Three-steps algorithm
  - Peak peaking: spectra denoising, baseline correction and peak selection: R package based on wavelet transform, developed for mass spectra
  - **Reference finding**: reference pattern against which all the spectra will be aligned. Automatic method proposed
Reference determination
Reference determination
Reference determination
CluPA algorithm (3)

• Vu et al. (2011): Cluster-based Peak Alignment

• Three-steps algorithm
  – Peak peaking: spectra denoising, baseline correction and peak selection: R package based on wavelet transform, developed for mass spectra
  – Reference finding: reference pattern against which all the spectra will be aligned. Automatic method proposed
  – **Peak alignment**: iterative method based on hierarchical clustering

![Diagram showing the steps of the CluPA algorithm](image-url)
Alignment algorithm (1)

• Peak lists of reference spectrum and target spectrum are merged

• Steps
  • 1 – Definition of the spectral zone to align
  • 2 – Determination of a shift step, based on the FFT cross correlation
  • 3 – Alignment
  • 4 – Hierarchical Clustering Algorithm of the shifted merged peak list
Alignment algorithm (2)
Alignment algorithm (1)

- Peak lists of reference spectrum and target spectrum are merged
- Steps
  - 1 – Definition of the spectral zone to align
  - 2 – Determination of a shift step, based on the FFT cross correlation
  - 3 – Alignment
  - 4 – Hierarchical Clustering Algorithm of the shifted merged peak list
  - 5 – Tree is cut at level 2
- Steps 1 – 5 are repeated for each sub-tree
Alignment algorithm (2)
Alignment algorithm (2)
Alignment algorithm (1)

- Peak lists of reference spectrum and target spectrum are merged
- Steps
  - 1 – Definition of the spectral zone to align
  - 2 – Determination of a shift step, based on the FFT cross correlation
  - 3 – Alignment
  - 4 – Hierarchical Clustering Algorithm of the shifted merged peak list
  - 5 – Tree is cut at level 2
- Steps 1 – 5 are repeated for each sub-tree, as long as
  - More than 1 peak in the peak list
  - Peaks from reference and target spectra in the peak list
Alignment algorithm (2)
CluPA evaluation (1)

- NMR spectra were simulated using MetAssimulo (Muncey et al. 2010)
  - MatLab toolbox to simulate 1H NMR spectra of complex mixtures
  - Based on a database including 49 reference compounds and on HMDB
  - Spectra including Dimethylamine, Citric acid, L-Histidine, Taurine, Trimethylamine, Hippuric acid, Glycine and TMAO were simulated under different pH conditions (from 4.0 to 10.0)
CluPA evaluation (2)
CluPA evaluation (3)
CluPA evaluation (4)
Form

Type of input files
Type of input files
Bruker file

- Preprocessing using TopSpin performed: Fourier transformation, spectra phased and baseline corrected, calibrated (TSP: $\delta 0$ ppm)
- For each individual sample, one directory as follows:
Form

Type of input files

Spectral width

Region(s) to exclude
Form

**Type of input files**

**Spectral width**

**Region(s) to exclude**
Form

**Type of input files**

**Spectral width**

**Region(s) to exclude**

**Left and right boundaries**
Form

- Type of input files
- Spectral width
- Region(s) to exclude
- Reference spectrum
- Segment size
- Minimal intensity
Form

Type of input files
Spectral width
Region(s) to exclude
Reference spectrum
Segment size
Minimal intensity
Graphical zone(s)
Form

<table>
<thead>
<tr>
<th>Graphical zone(s)</th>
<th>1: graphical zones</th>
<th>2: graphical zones</th>
<th>3: graphical zones</th>
<th>4: graphical zones</th>
<th>5: graphical zones</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Left graphical zone border</td>
<td>Right graphical zone border</td>
<td>Left graphical zone border</td>
<td>Right graphical zone border</td>
<td>Left graphical zone border</td>
</tr>
<tr>
<td></td>
<td>8.3</td>
<td>5.6</td>
<td>3.19</td>
<td>3.19</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>4.34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lactic acid

Citric acid

Lactic acid
Form

Type of input files

Spectral width

Region(s) to exclude

Reference spectrum

Segment size

Minimal intensity

HELP
Results: aligned matrix

Variables’ ID = chemical shift

Samples’ ID
Results: graph
Results: graph
BUCKETING TOOL
Bucketing or binning

- Data Reduction

16384 variables

600 variables
Bucketing or binning

- **Fixed size bucketing:**
  - Spectra segmentation in 600 (depends on the bucket width) windows of equal size, for example 0.01 ppm;
  - Exclusion of contaminants;
  - Bucket integration

- **Variable size bucketing:**
  - Each window is determined graphically;
  - The entire signal (singlet, doublet, triplet....) is in the same bucket;
  - Contaminants and noise are not chosen;
  - In case of shifts, the signal of a same proton is in the same bucket for all spectra;
  - The number of buckets is small (100-200)
  - Bucket integration
Bucketing or binning
Fixed-size Bucketing

- Binning: spectra segmentation in fixed-size windows (e.g. 0.01ppm)

- Spectra regions corresponding to water, solvents, ... resonances excluded
Integration

• Sum of intensities inside each bucket = area under the curve computation

• Trapezoidal method: cumtrapz function (pracma package)
  – For each bucket
    • Discretization into $N$ equally spaced “small” rectangles
Integration

- Sum of intensities inside each bucket = area under the curve computation
- Trapezoidal method: `cumtrapz` function (pracma package)
  - For each bucket
    - Discretization into $N$ equally spaced “small” rectangles
Integration

- Sum of intensities inside each bucket = area under the curve computation
- Trapezoidal method: cumtrapz function (pracma package)
  - For each bucket
    - Discretization into \( N \) equally spaced “small” rectangles
Integration

- Sum of intensities inside each bucket = area under the curve computation
- Trapezoidal method: cumtrapz function (pracma package)
  - For each bucket
    - Discretization into $N$ equally spaced “small” rectangles
Integration

• Sum of intensities inside each bucket = area under the curve computation

• Trapezoidal method: cumtrapz function (pracma package)
  – For each bucket
    • Discretization into $N$ equally spaced “small” rectangles
**Integration**

- Sum of intensities inside each bucket = area under the curve computation
- Trapezoidal method: `cumtrapz` function (pracma package)
  - For each bucket
    - Discretization into \( N \) equally spaced "small" rectangles
    - In each "small" rectangles, linear approximation of the curve = trapezium area is used to compute the area under the curve
  - **Absolute values of rectangles** are summed
**Form**

### Galaxy / Metabolomics

**Tools**
- Upload File from your computer
- LC-MS
- Preprocessing
- Normalisation
- Quality Control
- Statistical Analysis
- Annotation

**NMR Preprocessing**
- Read NMR raw files
- Preprocessing of 1D NMR spectra
- NMR spectra alignment based on the Cluster-based Peak Alignment (CluPA) algorithm

---

**NMR Bucketing**

**Type of input files**

**Choose your inputs method**
- Zip file
  - E.2: Mmuscule_sPA.zip

**Bucket width**
- 0.04
- Default value is 0.04 ppm

**Left Border**
- 10.0
- Default value is 10 ppm

**Right Border**
- 0.5
- Default value is 0.5 ppm

**Exclusion zone(s)**
- No
  - Choose if you want to exclude particular zone(s)

**Spectra representation**
- None
  - Select 'None' for no representation, 'Overlaid' to overlay all spectra on a unique chart and 'One per individual' to generate an individual chart for each observation

**Execute**

---

**Authors**
- Marie Tremblay-Franco (marie.tremblay-franco@toulouse.inra.fr), Marion Landi (marion.landi@clermont.inra.fr) and Franck Giacomoni (fgiacomon@clermont.inra.fr)

**NMR Bucketing**

**Description**

Bucketing / Binning (spectra segmentation in fixed-size windows) and integration (sum of absolute intensities inside each bucket) to preprocess NMR data

**Workflow position**

Upstream tools
Form

**NMR Bucketing**

**Description**

Bucketing / Binning (spectra segmentation in fixed-size windows) and integration (sum of absolute intensities inside each bucket) to preprocess NMR data.

**Workflow position**

Upstream tools

**Bin size**

- Left Border
  - Default value is 0.04 ppm
- Right Border
  - Default value is 10 ppm

**Spectral width**

- Default value is 0.5 ppm

**Region(s) to exclude**

- No selection

**Type of input files**

- Zip file from your history containing your Bruker directories

---

**Authors**

- Marie Tremblay-Franco (marie.tremblay-franco@toulouse.inra.fr)
- Marion Landi (marion.land@clermont.inra.fr)
- Franck Giacomoni (giacomoni@clermont.inra.fr)
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of input files</td>
<td>NMR_Bucketing bucketing and integration of NMR Bruker raw data (Galaxy Version 1.0.3)</td>
</tr>
<tr>
<td>Bin size</td>
<td>Bucket width: 0.04, default value is 0.04 ppm</td>
</tr>
<tr>
<td></td>
<td>Left Border: 10.0, default value is 10 ppm</td>
</tr>
<tr>
<td></td>
<td>Right Border: 0.5, default value is 0.5 ppm</td>
</tr>
<tr>
<td>Spectral width</td>
<td>Exclusion zone(s): yes (optional)</td>
</tr>
<tr>
<td>Left and right boundaries</td>
<td>Left exclusion zone border: 10.0</td>
</tr>
<tr>
<td></td>
<td>Right exclusion zone border: 10.0</td>
</tr>
<tr>
<td>Spectra representation</td>
<td>Select 'None' for no representation, 'Overlay' to overlay all spectra on a unique chart and 'One per individual' to generate an individual chart for each observation</td>
</tr>
</tbody>
</table>
Form

- Bin size
- Spectral width
- Region(s) to exclude
- Spectra visualisation
- Type of input files
Results: bucketed and integrated data matrix

Variables’ ID = Bucket center

Samples’ ID

File download
Results: variableMetadata matrix

Variables’ ID=Bucket center: row names of the variable Metadata file have to be identical to the row names of the dataMatrix file.
Results: spectra representation
NORMALIZATION TOOL
Normalization

- Operation applied to each sample to make the data from all samples directly comparable with each other (to take into account for variations of the overall concentrations of samples due to biological and technical reasons)

⇒ To ensure that a measured concentration observed for a metabolite at the lower end of the dynamic range is as reliable as it is for a metabolite at the upper end
Total intensity

• Intensity of each feature is divided by the intensity of total spectrum

• Each spectrum has a total intensity of 1
Quantitative variable

- Intensity of each feature is divided by the value of a quantitative variable: weight for tissue, osmolality, ...
Probabilistic Quotient Normalization (1)

- Hyp.: a majority of molecule concentrations remain unchanged across the samples = biologically interesting concentration changes influence only parts of the NMR spectrum, whereas dilution effects will affect all metabolite signals

- Comparison of spectra to a reference sample (the best approach is the computation of the median spectrum of control samples, Dieterle et al. 2006)
Probabilistic Quotient Normalization (2)

- Hyp.: a majority of molecule concentrations remain unchanged across the samples = biologically interesting concentration changes influence only parts of the NMR spectrum, while dilution effects will affect all metabolite signals
- Comparison of spectra to a reference sample (the best approach is the computation of the median spectrum of control samples, Dieterle et al. 2006)
- Steps
  - Total intensity normalization
Probabilistic Quotient Normalization (2)
Probabilistic Quotient Normalization (3)

- Hyp.: a majority of molecule concentrations remain unchanged across the samples = biologically interesting concentration changes influence only parts of the NMR spectrum, while dilution effects will affect all metabolite signals

- Comparison of spectra to a reference sample (the best approach is the computation of the median spectrum of control samples, Dieterle et al. 2006)

- Steps
  - Total intensity normalization
  - Reference spectrum computation
Probabilistic Quotient Normalization (3)
Probabilistic Quotient Normalization (4)

• Hyp.: a majority of molecule concentrations remain unchanged across the samples = biologically interesting concentration changes influence only parts of the NMR spectrum, while dilution effects will affect all metabolite signals

• Comparison of spectra to a reference sample (the best approach is the computation of the median spectrum of control samples, Dieterle et al. 2006)

• Steps
  – Total intensity normalization
  – Reference spectrum computation
  – Computation of quotients of all variables of interest of the test spectrum with those of the reference spectrum
Probabilistic Quotient Normalization (4)
Probabilistic Quotient Normalization (5)

- Hyp.: a majority of molecule concentrations remain unchanged across the samples = biologically interesting concentration changes influence only parts of the NMR spectrum, while dilution effects will affect all metabolite signals

- Comparison of spectra to a reference sample (the best approach is the computation of the median spectrum of control samples, Dieterle et al. 2006)

- Steps
  - Total intensity normalization
  - Reference spectrum computation
  - Computation of quotients of all variables of interest of the test spectrum with those of the reference spectrum
  - Median of these quotients computation
Probabilistic Quotient Normalization (5)
Probabilistic Quotient Normalization (6)

- Hyp.: a majority of molecule concentrations remain unchanged across the samples = biologically interesting concentration changes influence only parts of the NMR spectrum, while dilution effects will affect all metabolite signals
- Comparison of spectra to a reference sample (the best approach is the computation of the median spectrum of control samples, Dieterle et al. 2006)
- Steps
  - Total intensity normalization
  - Reference spectrum computation
  - Computation of quotients of all variables of interest of the test spectrum with those of the reference spectrum
  - Median of these quotients computation
  - Division of all variables of the test spectrum by this median
Normalization (2)
Preprocessed datamatrix
Normalization method

Authors Marie Tremblay-Franco (marie.tremblay-franco@toulouse.inra.fr) and Marion Landi (marion.landi@clermont.inra.fr)


Please cite
Preprocessed datamatrix
Normalization method

Normalization
Description
Normalization (operation applied on each (preprocessed) individual spectrum) of preprocessed data
Workflow position
Upstream tools
<table>
<thead>
<tr>
<th>Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preprocessed datamatrix</td>
</tr>
<tr>
<td>Normalization method</td>
</tr>
<tr>
<td>Sample metadata</td>
</tr>
<tr>
<td>« Reference » group</td>
</tr>
<tr>
<td>Biological factor of interest</td>
</tr>
</tbody>
</table>

1 mandatory for QuantitativeVariable and PQN normalization
Preprocessed datamatrix

Normalization method

Sample metadata\(^1\)

Biological factor of interest

« Reference » group

Spectra visualization

\(^1\) mandatory for QuantitativeVariable and PQN normalization
### Results: normalized data matrix

<table>
<thead>
<tr>
<th>Sample</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPA_c21_ag_126_BPA25ng</td>
<td>8.65583995148145e-06</td>
</tr>
<tr>
<td>BPA_c21_ag_133_BPA25ng</td>
<td>2.25962506196409e-06</td>
</tr>
<tr>
<td>BPA_c21_ag_135_BPA25ng</td>
<td>6.0889630355511e-06</td>
</tr>
<tr>
<td>BPA_c21_ag_136_BPA25ng</td>
<td>1.03275816559584e-05</td>
</tr>
<tr>
<td>BPA_c21_ag_137_BPA25ng</td>
<td>8.65363855074581e-06</td>
</tr>
<tr>
<td>Samples</td>
<td>Features</td>
</tr>
<tr>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>B9.495</td>
<td>7.162904050517e-06</td>
</tr>
<tr>
<td>B9.486</td>
<td>3.9203983334388e-06</td>
</tr>
<tr>
<td>B9.475</td>
<td>2.2460534766752e-06</td>
</tr>
<tr>
<td>B9.465</td>
<td>4.1060545127584e-06</td>
</tr>
<tr>
<td>B9.456</td>
<td>9.890011130262e-06</td>
</tr>
<tr>
<td>B9.445</td>
<td>1.0510435211246e-06</td>
</tr>
<tr>
<td>B9.436</td>
<td>8.9212833148396e-06</td>
</tr>
<tr>
<td>B9.426</td>
<td>8.2938857710687e-06</td>
</tr>
<tr>
<td>B9.415</td>
<td>7.867116842899e-06</td>
</tr>
<tr>
<td>B9.406</td>
<td>6.997021864198e-07</td>
</tr>
<tr>
<td>B9.396</td>
<td>9.076715263609e-07</td>
</tr>
<tr>
<td>B9.385</td>
<td>7.260895726553e-07</td>
</tr>
<tr>
<td>B9.376</td>
<td>8.99184168873114e-07</td>
</tr>
<tr>
<td>B9.366</td>
<td>2.83998857710687e-07</td>
</tr>
<tr>
<td>B9.355</td>
<td>7.688070906391e-07</td>
</tr>
<tr>
<td>B9.345</td>
<td>3.18919882676511e-07</td>
</tr>
<tr>
<td>B9.335</td>
<td>6.97039262783117e-08</td>
</tr>
<tr>
<td>B9.326</td>
<td>9.097755263609e-08</td>
</tr>
<tr>
<td>B9.315</td>
<td>9.023052343262307e-08</td>
</tr>
<tr>
<td>B9.305</td>
<td>8.7575895596097e-08</td>
</tr>
<tr>
<td>B9.296</td>
<td>5.2755895596097e-08</td>
</tr>
<tr>
<td>B9.285</td>
<td>1.58729789475568e-08</td>
</tr>
<tr>
<td>B9.275</td>
<td>3.42945322620259e-09</td>
</tr>
<tr>
<td>B9.266</td>
<td>4.91970642500408e-09</td>
</tr>
<tr>
<td>B9.255</td>
<td>3.297390160542284e-09</td>
</tr>
<tr>
<td>B9.246</td>
<td>5.3389935720707e-09</td>
</tr>
<tr>
<td>B9.236</td>
<td>5.2983433098654e-09</td>
</tr>
<tr>
<td>B9.225</td>
<td>6.354065887863e-09</td>
</tr>
<tr>
<td>B9.216</td>
<td>4.87516605424845e-09</td>
</tr>
<tr>
<td>B9.205</td>
<td>8.1993936139859e-09</td>
</tr>
</tbody>
</table>

### Features

- BPA_c21_ag_126_BPA25ng
- BPA_c21_ag_133_BPA25ng
- BPA_c21_ag_135_BPA25ng
- BPA_c21_ag_136_BPA25ng
- BPA_c21_ag_137_BPA25ng

### Samples

- B9.495
- B9.486
- B9.475
- B9.465
- B9.456
- B9.445
- B9.436
- B9.426
- B9.415
- B9.406
- B9.396
- B9.385
- B9.376
- B9.366
- B9.355
- B9.345
- B9.335
- B9.326
- B9.315
- B9.305
- B9.296
- B9.285
- B9.275
- B9.266
- B9.255
- B9.246
- B9.236
- B9.225
- B9.216

### Diagram

The diagram shows a comparison of features across samples, with a histogram illustrating the distribution of values for each sample.
Results: spectra representation
Results: spectra representation
PREPARING DATA
Format of data files

• Three tables gathering all the information
  • data matrix: intensities / expression (ions, buckets, proteins, …)
  • sample metadata file: information regarding samples (individuals)
  • variable metadata file: information regarding variables (ions, buckets, proteins, …)

• Tabulated files
  • TSV / TXT with tabulation as column separator
  • “.” as decimal separator

• Convention for identifiers and column names
  • Should not contain any duplicate
  • Rather use only alphanumeronic characters, and points (.) and underscores (_).
Format of data files: dataMatrix

- **variables**
  - **intensities**
    - **first column = Variables’ ID**
    - **ONLY intensities (no other information)**
    - Note: missing values should be coding NA

- **samples**
  - **first row = Samples’ ID**
Format of data files: sampleMetadata

<table>
<thead>
<tr>
<th>samples</th>
<th>Column names</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>intensities</td>
</tr>
</tbody>
</table>

- **first row** = treatment, diet, age ...
- Information about your samples
  - Can add to this table as many columns as you want or need
- **first column** = samples’ ID
  - must absolutely match those in the dataMatrix file

```
<table>
<thead>
<tr>
<th></th>
<th>sampleMetadata</th>
<th>sampleType</th>
<th>injectionOrder</th>
<th>mode</th>
<th>batch</th>
<th>age</th>
<th>bmi</th>
<th>gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sampleMetadata</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>HU_neg_017</td>
<td>sample</td>
<td>17</td>
<td>neg</td>
<td>ne1</td>
<td>41</td>
<td>23.03</td>
<td>M</td>
</tr>
<tr>
<td>3</td>
<td>HU_neg_028</td>
<td>sample</td>
<td>23</td>
<td>neg</td>
<td>ne1</td>
<td>41</td>
<td>23.92</td>
<td>F</td>
</tr>
<tr>
<td>4</td>
<td>HU_neg_034</td>
<td>sample</td>
<td>26</td>
<td>neg</td>
<td>ne1</td>
<td>52</td>
<td>23.37</td>
<td>M</td>
</tr>
<tr>
<td>5</td>
<td>HU_neg_051</td>
<td>sample</td>
<td>45</td>
<td>neg</td>
<td>ne1</td>
<td>24</td>
<td>23.23</td>
<td>F</td>
</tr>
</tbody>
</table>
```
Format of data files: variableMetadata

- **First column** = variables’ ID must absolutely match those in the dataMatrix file.
- **First row** = retention time, molar mass, ...

Information about your variables:
Can add to this table as many columns as you want or need.

---

### Example Table

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>variableMetadata</td>
<td>mz/rt</td>
<td>fold</td>
<td>tstat</td>
<td>pvalue</td>
<td>mzmed</td>
<td>mzmin</td>
<td>mzmmax</td>
</tr>
<tr>
<td>M97T61</td>
<td>47</td>
<td>69.2762</td>
<td>47.6615</td>
<td>585</td>
<td>96.9598</td>
<td>96.9544</td>
<td>96.9605</td>
</tr>
<tr>
<td>M99T61</td>
<td>52</td>
<td>390.1176</td>
<td>18.1537</td>
<td>251</td>
<td>98.9556</td>
<td>98.9554</td>
<td>98.9556</td>
</tr>
<tr>
<td>M135T54</td>
<td>179</td>
<td>394.0080</td>
<td>18.5812</td>
<td>9475</td>
<td>135.0296</td>
<td>135.0295</td>
<td>135.0295</td>
</tr>
<tr>
<td>M136T54</td>
<td>183</td>
<td>Inf</td>
<td>-17.6102</td>
<td>2175</td>
<td>136.0329</td>
<td>136.0328</td>
<td>136.0328</td>
</tr>
<tr>
<td>M189T52</td>
<td>504</td>
<td>Inf</td>
<td>-18.7528</td>
<td>2935</td>
<td>189.0346</td>
<td>189.0346</td>
<td>189.0346</td>
</tr>
</tbody>
</table>
EXERCISE – MTBLS1 DATA
EXERCISE

- Do the spectra preprocessing using different parameters:
  - Apodization exponential: LB=0; 0.3 or 1
  - Zero order phase correction: rms or max
  - Shift referencing: nearvalue; all; or window
  - Baseline correction: use different values for lambda, and for p

- Do the spectra alignment using different segment size, reference spectrum and intensity threshold

- Do the spectra bucketing using different bin sizes

- Do the spectra normalization using different methods

- Compare results