NMR Data Pre-processing

UAB Metabolomics Training Course

June 14-18, 2015

Wimal Pathmasiri, Rodney Snyder

NIH Eastern Regional Comprehensive Metabolomics Resource Core

(RTI RCMRC)
After NMR data acquisition, the result is a set of spectra for all samples.

For each spectrum, quality of the spectra should be assessed.
- Line shape
- Phase
- Baseline

Spectra should be referenced
- Compounds commonly used: DSS, TSP, Formate

Variations of pH, ionic strength of samples has effects on chemical shift
- Peak alignment
- Bucket integration

Remove unwanted regions
Quality Control Steps

- Quality of metabolomics analysis depends on data quality

- Typical problems
  - Water peak (suppression issues)
  - Baseline (not set at zero and not a flat line)
  - Alignment of peaks (chemical shift, due to pH variation)
  - Variation in concentration (e.g., Urine)

- High quality of data is needed for best results
Water Suppression Effects and Other Artifacts

- If water is not correctly suppressed or removed there will be effects on normalization
- Need to remove other artifacts
- Remove drug or drug metabolites

Same Serum sample

Poor water suppression

Good water suppression
Before

Phase
Baseline to be corrected

After

Reference,
Line shape
Chemical shift variability
- pH
- ionic strength
- metal concentration

Methods to overcome this problem
- Use a buffer when preparing samples
- Binning (Bucketing)
  - Fixed binning
  - Intelligent binning
  - Optimized binning
- Available data alignment tools
  - Recursive Segment-wise Peak Alignment (RSPA)
  - Icosift
  - speaq

http://www.chenomx.com/software/software.php
Vu, T. N. et al., BMC Bioinformatics 2011, 12:405
Example

**icoshift**

One of the Citrate peaks

- **Before**: Graph showing multiple overlapping peaks before alignment.
- **After**: Graph showing the peaks aligned, with clearer and more distinct separation.

Example

speaq

Vu, T. N. et al., *BMC Bioinformatics* 2011, **12**:405
NMR Binning

- A form of quantification that consists of segmenting a spectrum into small areas (bins/buckets) and attaining an integral value for that segment

- Binning attempts to minimize effects from variations in peak positions caused by pH, ionic strength, and other factors.

- Two main types of binning
  - Fixed binning
  - Flexible binning
The entire NMR spectrum is split into evenly spaced integral regions with a spectral window of typically 0.04 ppm.

The major drawback of fixed binning is the non-flexibility of the boundaries.

If a peak crosses the border between two bins it can significantly influence your data analysis.

Peak shift can cause the same peak across multiple samples to fall into different bins.

Signals for citrate are split into multiple bins.
Signals for citrate are properly captured.

Signals for citrate are split into multiple bins.
Remove regions

Downfield region

Urea

Water

DSS and up field

Chemical Shift (ppm)

Normalized Intensity

Urea Water

DSS and up field
- Integrate bins (0.04 ppm bin size)
- Normalize integral of each bin to the total integral of each spectrum
- Merge metadata
- Result is a spreadsheet ready for further multivariate data analysis and other statistical analysis
Data Normalization, Transformation, and Scaling
Normalization reduces the sample to sample variability due to differences in sample concentrations—particularly important when the matrix is urine.

- Normalization to total intensity is the most common method
  - For each sample, divide the individual bin integral by the total integrated intensity

- Other Methods
  - Normalize to a peak that is always present in the same concentration, for example normalizing to creatinine
  - Probabilistic quotient normalization
  - Quantile and cubic spline normalization
### Centering, Scaling, and Transformations

<table>
<thead>
<tr>
<th>I</th>
<th>Centering</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\tilde{x}<em>{ij} = x</em>{ij} - \bar{x}_i$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>II</th>
<th>Autoscaling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Range scaling</td>
<td>$\tilde{x}<em>{ij} = \frac{x</em>{ij} - \bar{x}_i}{s_i}$</td>
</tr>
<tr>
<td>Pareto scaling</td>
<td>$\tilde{x}<em>{ij} = \frac{x</em>{ij} - \bar{x}_i}{\sqrt{s_i}}$</td>
</tr>
<tr>
<td>Vast scaling</td>
<td>$\tilde{x}<em>{ij} = \frac{(x</em>{ij} - \bar{x}_i)}{s_i} \cdot \frac{\bar{x}_i}{\bar{x}_i}$</td>
</tr>
<tr>
<td>Level scaling</td>
<td>$\tilde{x}<em>{ij} = \frac{x</em>{ij} - \bar{x}_i}{\bar{x}_i}$</td>
</tr>
</tbody>
</table>

| III | Log transformation               |
|     | $\tilde{x}_{ij} = 10 \log(x_{ij})$ |
|     | $\tilde{x}_{ij} = \tilde{x}_{ij} - \bar{x}_i$ |

| III | Power transformation             |
|     | $\tilde{x}_{ij} = \sqrt(x_{ij})$ |
|     | $\tilde{x}_{ij} = \tilde{x}_{ij} - \bar{x}_i$ |

Analysis results vary depending on the scaling/ transformation methods used.

Van den Berg et al 1006, BMC Genomics, 7, 142
Before transformation
- skew distribution

After log-transformation
- More close to normal distribution

Susan Wicklund, Multivariate data analysis for omics, Sept 2-3 2008, Umetrics training
- Unit variance (autoscaling) divides the bin intensity by the standard deviation
  - May increase your baseline noise
  - Dimensionless value after scaling

- Pareto scaling divides the bin intensity by the square root of the standard deviation
  - Not dimensionless after scaling

- For NMR data, centering with pareto scaling is commonly used
Multivariate Data Analysis and Other Statistical Analyses

- Mean centered and scaled data
- Non-supervised analysis
  - Principal component analysis (PCA)
- Supervised analysis
  - PLS-DA and OPLS-DA
- Loadings plots and VIP Plots to identify discriminatory bins
- p-Value, fold change
ACKNOWLEDGEMENTS

Director RTI RCMRC
Susan Sumner, PhD

Program Coordinator
Jason Burgess, PhD

NIH Scientific Officer
David Balshaw, PhD, NIH/NIEHS

Internship Program
Stella Lam, BS

Feasibility Studies
Susan Sumner, PhD
Susan McRitchie, MS
Executive Committee

Website
Roger Austin, MS

Biochemistry and Molecular Biology
Timothy Fennell, PhD
Ninell Mortensen, PhD
Delisha Stewart, PhD

Biorepository
Brian Thomas, PhD
Mike McCleary, BS

Interns
Tammy Cavallo
Aastha Ghimire
Zachery Acuff

LC-MS Metabolomics
Suraj Dhungana, PhD
Brian Thomas, PhD
James Carlson, MS
Alex Kovach, BS
Rodney Snyder, MS
Moses Darko, BS

GC-MS Metabolomics
Wimal Pathmasiri, PhD
Jocelin Deese-Spruill, BS
Keith Levine, PhD
James Harrington, PhD
William Studabaker, PhD

NMR Metabolomics
Wimal Pathmasiri, PhD
Kelly Mercier, PhD
Rodney Snyder, MS
Tammy Cavallo, BS
Kevin Knagge, PhD, DHMRI
Jason Winnike, PhD, DHMRI

Statistics, Bioinformatics, and Computing
Susan McRitchie, MS
Robert Clark, PhD
Andrew Novokhatny, BS

Advisors: Imperial College, UK
Jeremy Nicholson, PhD
Elaine Holmes, PhD
Ian Wilson, PhD