Alignment and Preprocessing for Data Analysis

- Preprocessing tools for chromatography
- Basics of alignment
- GC-FID (1D) data and issues
  - PCA
  - F-Ratios
- GC-MS (2D) data and issues
  - PCA
  - F-Ratios
  - PARAFAC
- Piecewise Alignment GUI (available online)
  - synoveclab.chem.washington.edu/Downloads.htm
  - Email for username/password
Tools for Analysis: Classification

- Principal Component Analysis (PCA)

- Degree of Class Separation (DCS)

\[
DCS = \frac{D_{A,B}}{\sqrt{s_A^2 + s_B^2}}
\]

\[
D_{A,B} = \sqrt{(X_A - X_B)^2 + (Y_A - Y_B)^2}
\]
Why Align?

• Reduction in classification
  – PCA

• Increase in uncertainty for quantification
  – PARAFAC

• Misalignment occurs frequently
  – Daily instrument variation causes misalignment
  – Correction is necessary to apply these methods
Retention Time Precision & PCA

10 Replicates with limited shifting

10 Replicates with large shifting

Signal
Retention Time

PCA
Loadings
Retention Time

Scores
Score PC1
DCS=32

Scores
Score PC1
DCS=3

Loadings
Retention Time

Scores
Score PC2
DCS=32

Scores
Score PC2
DCS=3
Basics of Alignment

• Types of alignment algorithms
  – Cross correlation coefficient
  – Correlation Optimized Warping (COW)
  – *Piecewise alignment

• Alignment Parameters
  – Window Size
  – Shift

• Target Selection
  – PCA
  – Correlation Coefficient
  – Windowed Target
Alignment Algorithms

• Cross correlation coefficient
  – Move the entire chromatogram to maximize correlation

• Correlation Optimized Warping (COW)
  – Separate the chromatogram into windows
  – Warp and move the windows to optimize the correlation
  – Find the best alignment path to correct the data

• *Piecewise alignment
  – Separate the chromatogram into windows
  – Shift the windows to optimize the correlation
  – Find the best alignment path to correct the data
## Alignment Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Effect Too Small</th>
<th>Effect Too Large</th>
<th>Determining correct values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Window Size, $W$</td>
<td>$\text{Window Size } \sim 1 \text{ pk width}$</td>
<td>Relative movement high, difficult to determine quality of alignment</td>
<td>Insufficient flexibility to correct peak to peak shifting</td>
<td>Alignment Metric</td>
</tr>
<tr>
<td>Shift, $L$</td>
<td>$0 &lt; \text{shift} \leq \text{maximum shift}$</td>
<td>Insufficient movement of segments</td>
<td>Increases time</td>
<td>Alignment Metric</td>
</tr>
</tbody>
</table>
Target Selection

• *Global Approach
  – All chromatograms are initially collected
  – *User chosen target
  – PCA optimized target
    • Scores are produced for every sample
    • The sample with the minimum distance from the center is the target
  – *Maximum correlation target
    • Calculate the product of each chromatograms correlation to the others
    • The maximum correlation is the target

• Online Approach
  – An initial target is set
  – Chromatograms are aligned as they are collected
  – The target changes as new chromatograms are collected
Target Selection (Maximum Correlation)

- Sample #9 is the target

Retention Time, min

Product Correlation

Signal

Retention Time, min

Sample #
Alignment of GC-FID (1D) Data and Issues

- Removal of artifacts and solvent peaks
- Baseline correction and normalization
- Alignment
- Improving PCA
  - F-Ratio
Preprocessing Tools for Chromatography

- Noise filtering
  - Median filter
- Baseline correction
- Normalization
- Alignment
Experimental

- GC-FID Separation
- 4 diesel sample types
- 9 minute separation
- 17 replicate injections over 5 days
Noise Filtering (Solvent, Spikes, and Outliers)

GC-FID Separation

Signal vs. Retention Time, min

Noise Spike

Noise Spike

Median Filtered Data (minimize spike)
Noise Filtering (Solvent, Spikes, and Outliers)

GC-FID Separation

Scores Plot

Solvent Dominated Info

Outlier (low signal)
Noise Filtering (Solvent, Spikes, and Outliers)

GC-FID Separation

Scores Plot

Outlier (Low Signal)
Noise Filtering (Solvent, Spikes, and Outliers)

GC-FID Separation

Scores Plot

No Outlier (Baseline / Normalization on PC1)
Noise Filtering (Solvent, Spikes, and Outliers)

Baseline Corrected and Normalized Data

Scores Plot

Tight Clustering in PCA

Retain Time, min

Normalized Signal

PC1

PC2
Experimental

- GC Separation
- 3 gasoline sample types
- 15 minute separation
- 6 replicate injections over 2 days
- Misalignment is due to day to day instrument variation
Alignment GUI

Initially blank to guide the user through alignment
Alignment GUI

Load from workspace

Load from file
Now save data to workspace or .mat file.

Estimated target and parameters:

Data before alignment:

Data after alignment:

Alignment GUI
We want better separation of groups.
Fisher Ratio Method (F-Ratio)

For each mass channel calculate Fisher Ratio at each point in 2D space,

\[
\text{Fisher Ratio} = \frac{\sigma^2_{\text{between class}}}{\sigma^2_{\text{within class}}}
\]

- Works well for samples that have large amounts of within class variance
- Works best when comparing a small number of sample classes
Fisher Ratio Method (F-Ratio)

Selected Chromatographic Region

F-Ratio

Retention Time, min

Signal

Retention Time, min

F-Ratio

Retention Time, min

No alignment

After alignment

Select Top Ten
Fisher Ratio Method (F-Ratio)

Regions of interest

Scores Plot

Tight clustering of sample types
Summary

• Removal of solvents or artifacts is essential
• Baseline correction is an important step
• Alignment is essential for improving classification
• The F-Ratio algorithm can further improve classification
Alignment of GC-MS (2D) Data and Issues

- Removal of artifacts and solvent peaks
- Baseline correction and normalization
- Alignment
- Improving PCA
  - F-Ratio
- PARAFAC
Experimental

- GC-MS Separation
- 3 gasoline sample types
- 15 minute separation
- 6 replicate injections over 2 days
- Misalignment is due to day to day instrument variation
Alignment GUI

Initially blank to guide the user through alignment
Alignment GUI

Load from workspace

Load from file
Alignment GUI

Load & View Data Sizes

Choose or Estimate Parameters

Choose or Estimate Target

Align Data

Zoom & Pan
Now save data to workspace or .mat file.
Alignment

![Alignment Diagram]

TIC

Retention Time

Sample Target

Shift

Retention Time

Total Ion Current Shift Function (TIC-SF)

Alignment
Alignment

Total Ion Current Shift Function (TIC-SF)

Retention Time

Sample
Target

GC-MS Data

Apply alignment

Retention Time

Retention Time

GC-MS Data
Alignment

Retention Time, min

TIC Signal

Retention Time, min

TIC Signal

Align
Classification of Full MS Data

Scores Plot

PCA w/ full MS

TIC Signal

Retention Time, min
Fisher Ratio Method

For each mass channel calculate Fisher Ratio at each point in 2D space,

\[
\text{Fisher Ratio} = \frac{\sigma^2_{btw class}}{\sigma^2_{within class}}
\]

- Works well for samples that have large amounts of within class variance
- Works best when comparing a small number of sample classes
Simulated Example Using F-Ratios for PCA

Align

TIC

Scores

PCA

F-Ratios

m/z

2D F-Ratios

Retention Time

W = 400
L = 1000
Align

Retention
Time

m/z

Scores 2D

F-Ratios Shows
differences
between all
samples

TICTIC

F-RatiosGC-MS Data
Procedure

• Select masses using mass spectral information
• Select time regions using F-Ratios
• Combine to reduce the data set
• Improve results of PCA
Classification of Full MS Data

Regions of interest (with MS values)

Scores Plot of Nearest Samples

Clear Separation on 1st PC
Quantification of GC-MS Data

• Use aligned, baseline corrected and normalized data

• Use PARAFAC of small regions for analysis
  – Match values to mass spectra
  – Peak Sums
  – Peak Profiles
Quantification

- Target analyte Parallel Factor Analysis (PARAFAC) isolates the pure component peak and mass spectral information from overlapping peaks and background for both *identification* and *quantification*.

Data Matrix = Analyte 1 + Analyte 2 + ... Noise

PARAFAC Results
- Compound: malate, TMS
- # of factors: 3
- Chosen comp.: #3
- Match value: 948
- Peak sum: 4413882
- Peak height: 376188

Col. 1 Time

Intensity

Col. 2 Time

Intensity
PARAFAC for GC-MS Data

Stack Samples

Run PARAFAC

Peak Profile

Mass Spectrum

Sample 1

Sample 2

Retention Time

Sample

Analyte Concentration
Experimental

- GC-MS Separation
- 3 gasoline sample types
- 15 minute separation
- 6 replicate injections over 2 days
- Misalignment is due to day to day instrument variation
- Looking for isobutyl benzene in the gasoline
Isobutyl Benzene Spectrum
PARAFAC Results for GC-MS Data

Selected Analysis Region for PARAFAC

Retention Time, min

TIC Signal

- Peak Profile
- Qualitative Info
- Quantitative Info
- Qualitative Info
- Mass Spectrum
PARAFAC Results for GC-MS Data

Component 1

\[ \text{MV}_{\text{IBB}} = 403 \]

Component 2

\[ \text{MV}_{\text{IBB}} = 788 \]

Component 3

\[ \text{MV}_{\text{IBB}} = 910 \]

Best match to isobutyl benzene
Summary

• Mass spectral chromatographic data should be aligned
• An available alignment algorithm is able to align 1-D and 2-D data
• F-Ratio methods can be used to improve classification
• PARAFAC can be effectively used to separate overlapped analytes