METABOLOMICS

The X-windows High-Throughput Mass Spectroscopy Pipeline

Rui C. Martins*, Castro CC**, Teixeira JA**, Silva-Ferreira AC ***

***BioInformatics and Biophysics Lab**
Molecular Biology and Ecology Research Centre
University of Minho
Braga - Portugal

*****Chemiomics and Metabolomics Lab**
Biotechnology Research Center
Portuguese Catholic University
Porto - Portugal

****BioEngineering Lab**
Institute for BioEngineering and BioTechnology
University of Minho
Braga - Portugal

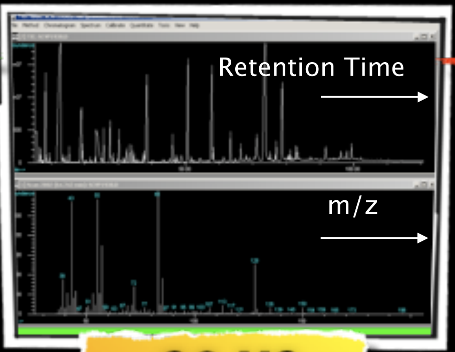
In This Talk:

1. Metabolomics: Target vs Non-Target Approaches
2. Mass Spectroscopy Signal Processing
3. The X – Metabolomics Architecture
4. Software Features
5. High-throughput examples
6. Software Demonstration

METABOLOMICS

High-Throughput Mass Spectroscopy

1. Metabolomics: Target vs Non-Target Approaches



GC-MS data

"Target"

Identification

Quantification

Classification

Kinetic parameters

"Non-Target"

Preprocessing:
- Baseline correction
- Scan Alignment

Classification

Relevant Scans

Identification

METABOLOMICS

High-Throughput Mass Spectroscopy

2. Mass Spectroscopy Signal Processing

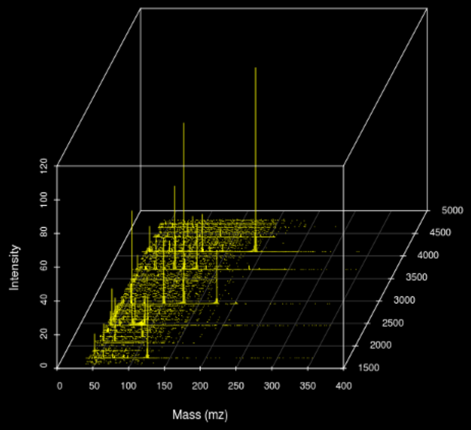
2.1. The MS Chromatographic Signal
2.2. Main Approaches: BioInformatics Vs Chemometrics
2.3. Pre-Processing
2.4. Feature Extraction
2.4. Chromatographic Alignment
2.5. Robust Peak Recognition
2.6. Identification and Composition
2.7. High-troughput MS BioInformatics

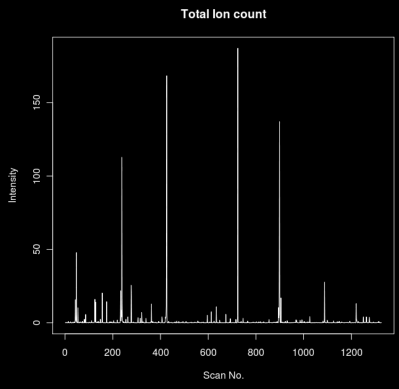
METABOLOMICS

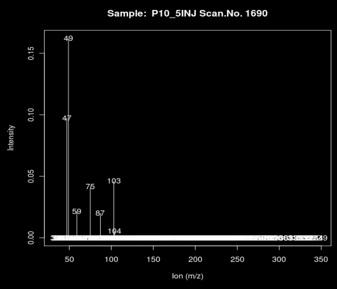
High-Throughput Mass Spectroscopy

2. Mass Spectroscopy Signal Processing

2.1. The MS Chromatographic Signals







Dynamic
Multi-Scale
Multivariate
Convoluted

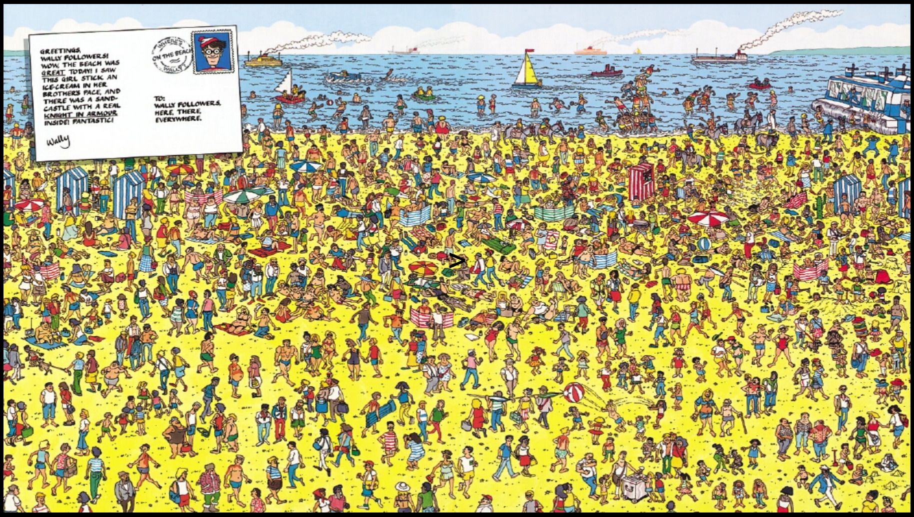
Complex Mathematics

METABOLOMICS

High-Throughput Mass Spectroscopy

2. Mass Spectroscopy Signal Processing

2.1. The MS Chromatographic Signals



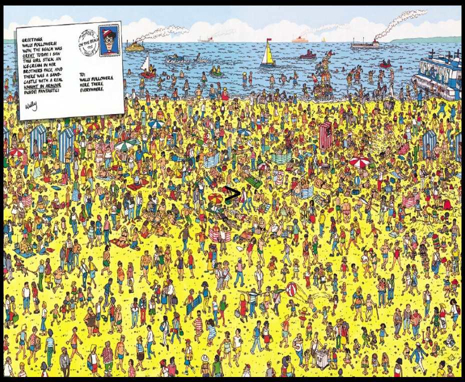
Where Is Wally?
or
How Can I Know Everybody?

METABOLOMICS

High-Throughput Mass Spectroscopy

2. Mass Spectroscopy Signal Processing

2.1. The MS Chromatographic Signals



Target Approach

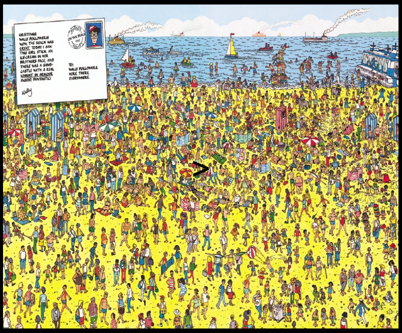
Holistic Approach

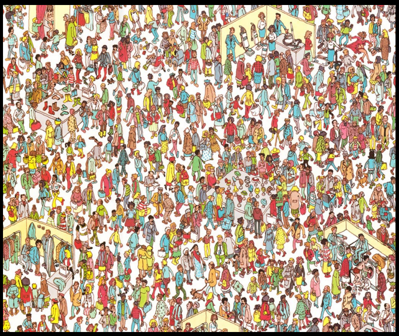
METABOLOMICS

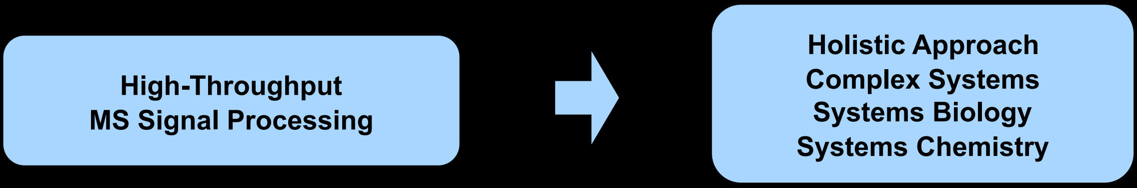
High-Throughput Mass Spectroscopy

2. Mass Spectroscopy Signal Processing

2.1. The MS Chromatographic Signals







High-Throughput MS Signal Processing

→

Holistic Approach
Complex Systems
Systems Biology
Systems Chemistry

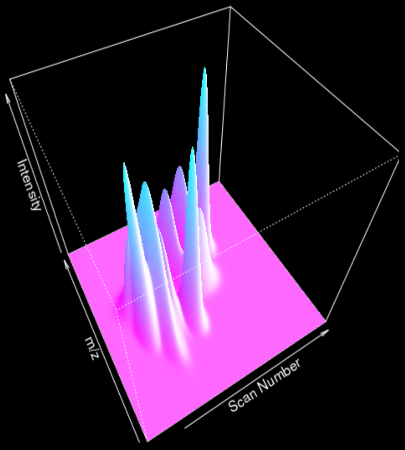
METABOLOMICS

High-Throughput Mass Spectroscopy

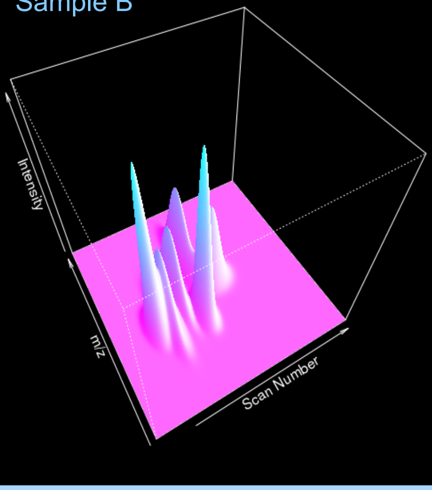
2. Mass Spectroscopy Signal Processing

2.1. The MS Chromatographic Signals

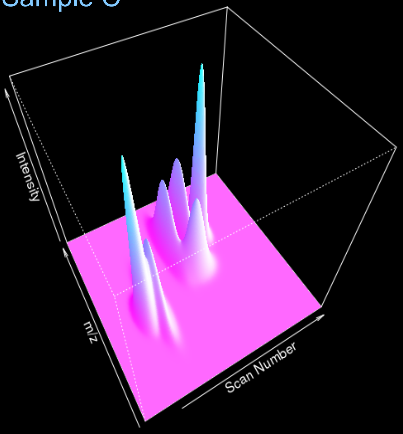
Sample A



Sample B



Sample C



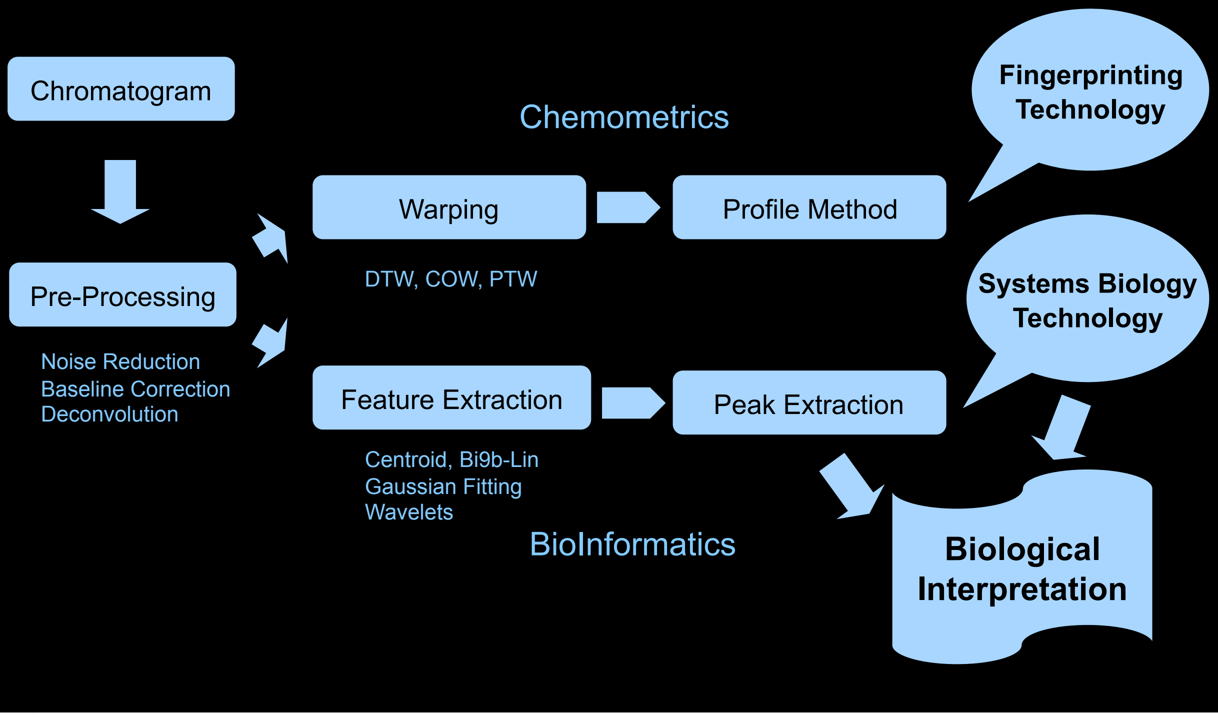
'In-Silico' Processing:
Pre-process, Extract, Recognize, Identify, Quantify, Analise

METABOLOMICS

High-Throughput Mass Spectroscopy

2. Mass Spectroscopy Signal Processing

2.2. Main Approaches: BioInformatics Vs Chemometrics



Chromatogram

Pre-Processing

Noise Reduction
Baseline Correction
Deconvolution

Chemometrics

Warping

DTW, COW, PTW

Profile Method

Fingerprinting Technology

Feature Extraction

Centroid, Bi9b-Lin
Gaussian Fitting
Wavelets

Peak Extraction

Systems Biology Technology

BioInformatics

Biological Interpretation

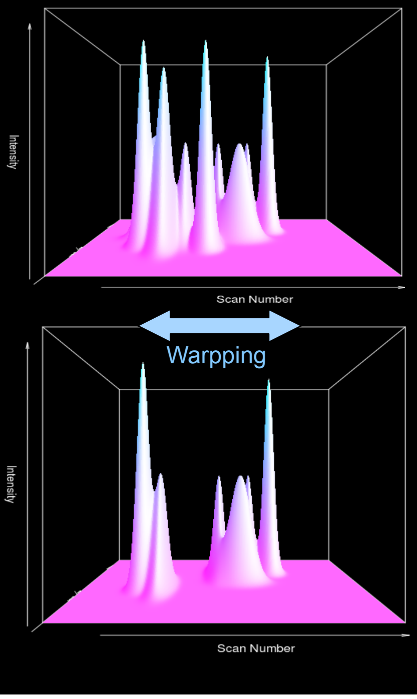
METABOLOMICS

High-Throughput Mass Spectroscopy

2. Mass Spectroscopy Signal Processing

2.2. Main Approaches: Chemometrics

Profile Method



Warp the chromatogram repecting an objective function:

1.Dynamic Time Warpping: $j=\text{argmin}\ (yi - yr)^2$

2. Correlation Time Warping: $j=\text{argmax}\ \text{corr}(Yi/Yr)$

e.g. Which Reference Chromatogram to use with complex biological samples?

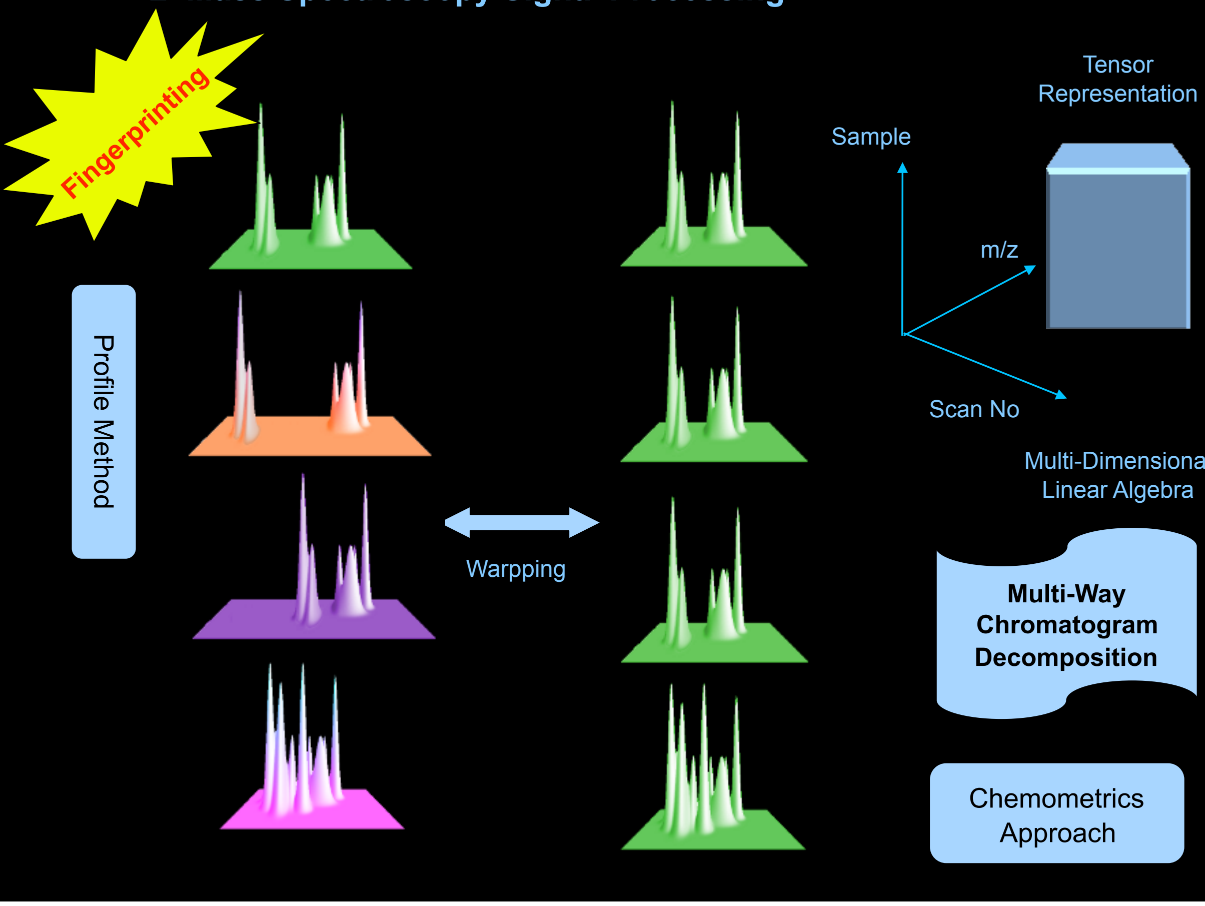
Only robust with low-rank Chromatography – Low complexity of samples → Not for Metabolomics!

Not For Complex Chromatograms!

METABOLOMICS

High-Throughput Mass Spectroscopy

2. Mass Spectroscopy Signal Processing



Fingerprinting

Profile Method

Warping

Tensor Representation

Multi-Dimensional Linear Algebra

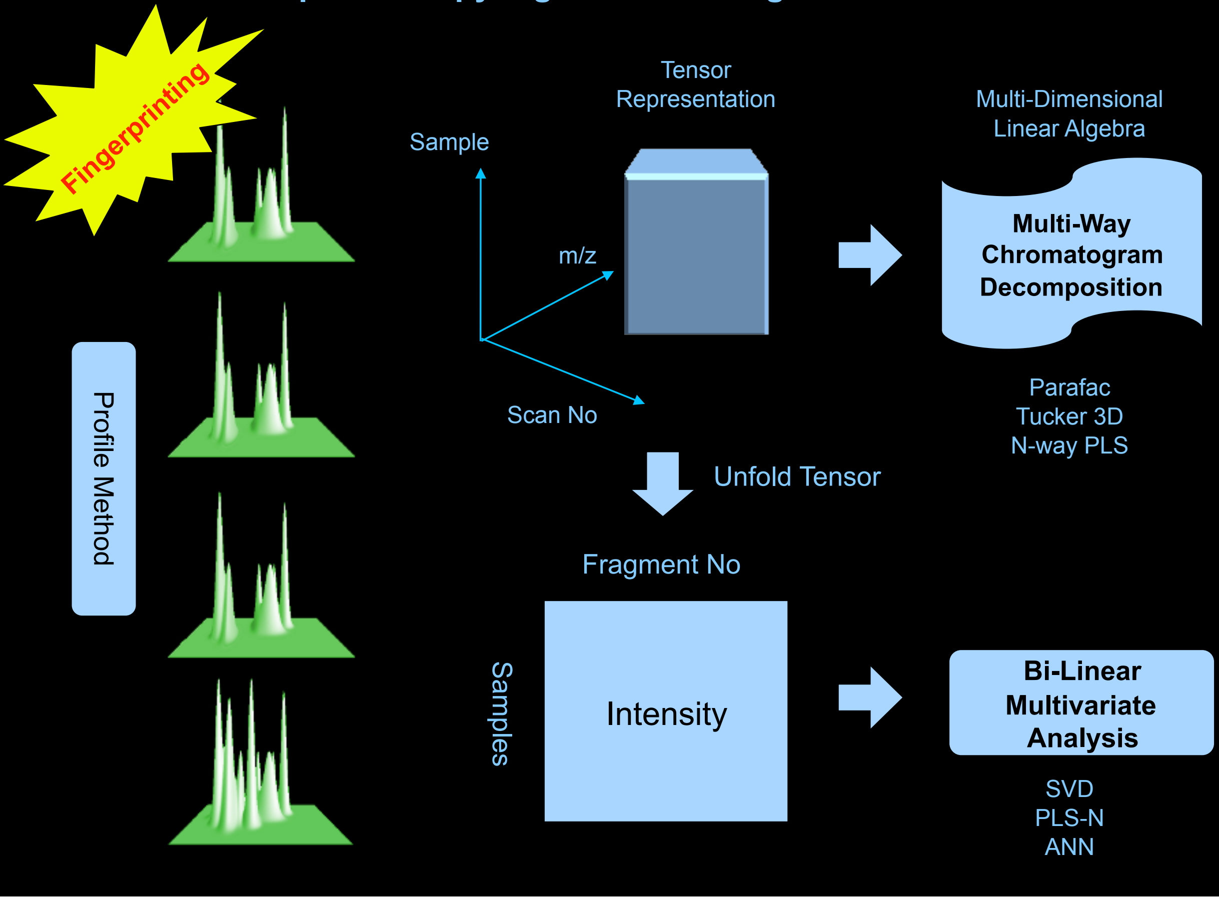
Multi-Way Chromatogram Decomposition

Chemometrics Approach

METABOLOMICS

High-Throughput Mass Spectroscopy

2. Mass Spectroscopy Signal Processing



Fingerprinting

Profile Method

Sample

m/z

Scan No

Tensor Representation

Multi-Dimensional Linear Algebra

Multi-Way Chromatogram Decomposition

Parafac
Tucker 3D
N-way PLS

Unfold Tensor

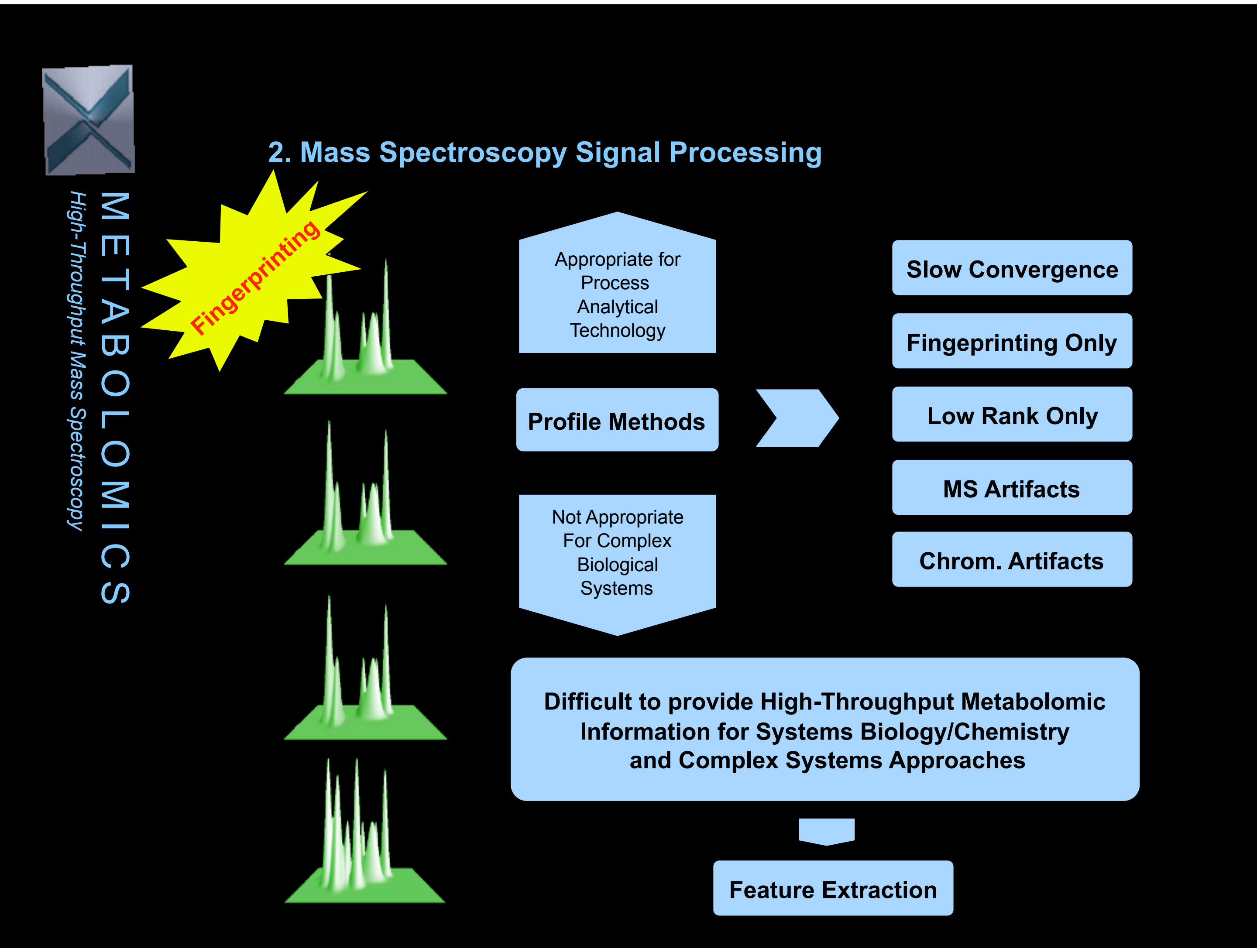
Fragment No

Samples

Intensity

Bi-Linear Multivariate Analysis

SVD
PLS-N
ANN



METABOLOMICS

High-Throughput Mass Spectroscopy

2. Mass Spectroscopy Signal Processing

Fingerprinting

Appropriate for Process Analytical Technology

Profile Methods

Not Appropriate For Complex Biological Systems

Slow Convergence

Fingeprinting Only

Low Rank Only

MS Artifacts

Chrom. Artifacts

Difficult to provide High-Throughput Metabolomic Information for Systems Biology/Chemistry and Complex Systems Approaches

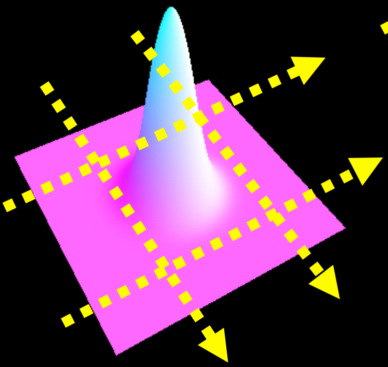
Feature Extraction

METABOLOMICS

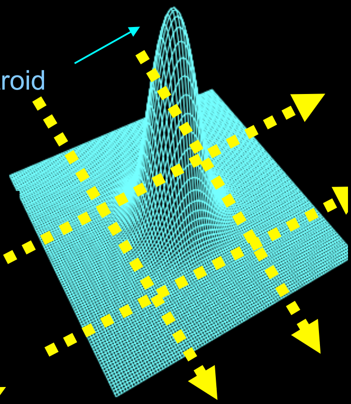
High-Throughput Mass Spectroscopy

2. Mass Spectroscopy Signal Processing

2.4. Feature Extraction

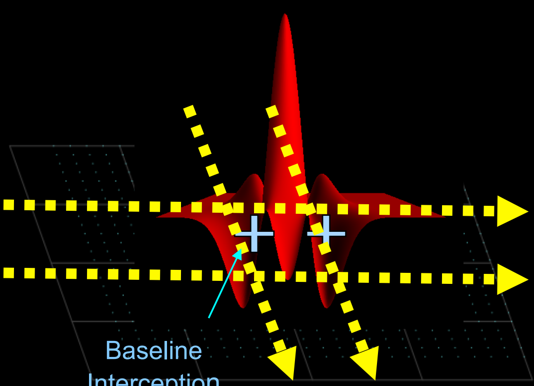


Peak Bining



Centroid

Gaussian Curve Peak Extraction



Centroid

Baseline Interception

Wavelet Peak Extraction

Bin-Lin/Centroid Methods

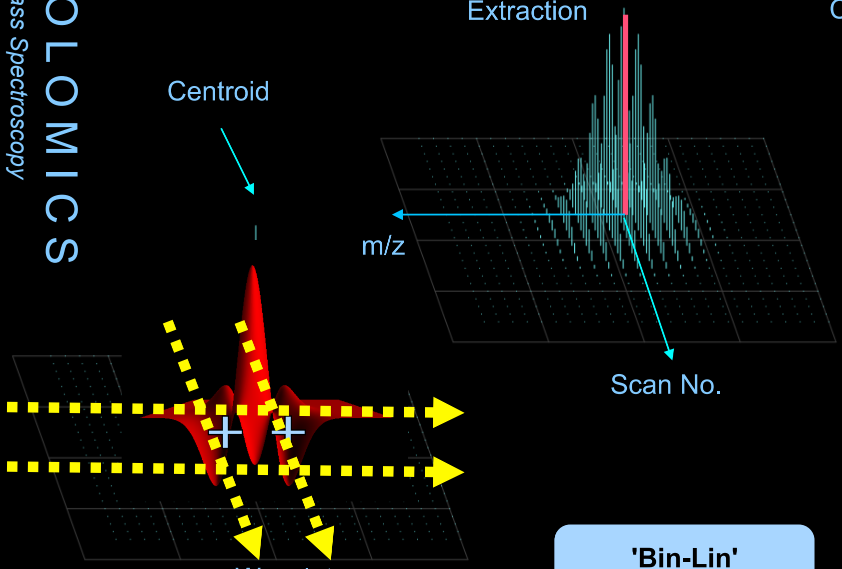
METABOLOMICS

High-Throughput Mass Spectroscopy

2. Mass Spectroscopy Signal Processing

2.4. Feature Extraction

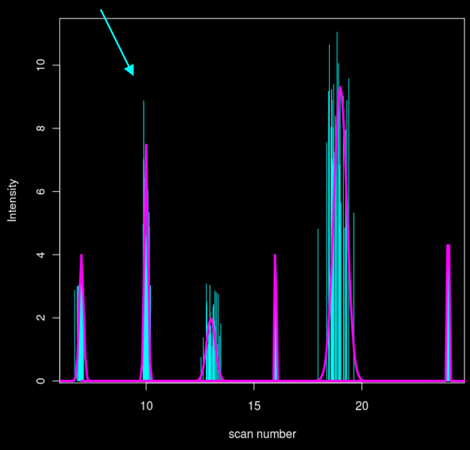
Centroid Extraction



'Bin-Lin' Method

'Equal' Sample Centroids

Peak Grouping



Assuming Centroid Scan No is Normally Distributed!

METABOLOMICS

High-Throughput Mass Spectroscopy

2. Mass Spectroscopy Signal Processing

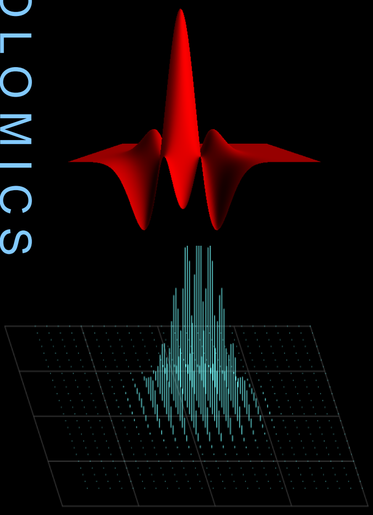
2.4. Feature Extraction

Wavelets:

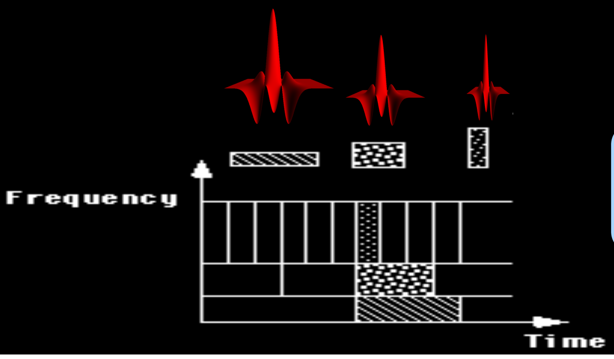
- Representation of a signal by a new orthonormal space basis given by non-stationary oscillating waveforms;

- Discontinuities and sharp peaks;

- The Mexican Hat Wavelet: $\psi(t) = \frac{1}{\sqrt{2\pi}\sigma^3}\left(1 - \frac{t^2}{\sigma^2}\right) e^{\frac{-t^2}{2\sigma^2}}$



Wavelet
Peak Extraction



Multi-Scale Chromatogram Decomposition

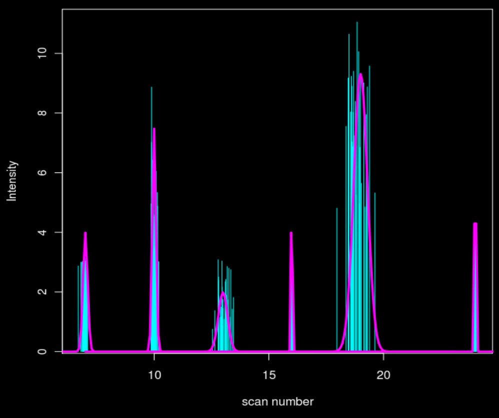
METABOLOMICS

High-Throughput Mass Spectroscopy

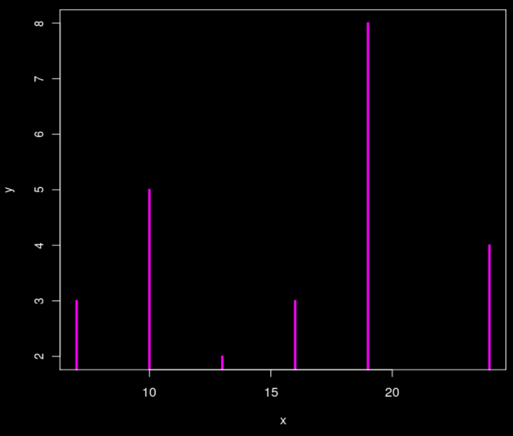
2. Mass Spectroscopy Signal Processing

2.4. Feature Extraction

Synchronization



Centroid Retention Time Correction Among Samples



Interpolation functions To correct deviation in 'non-common' centroids

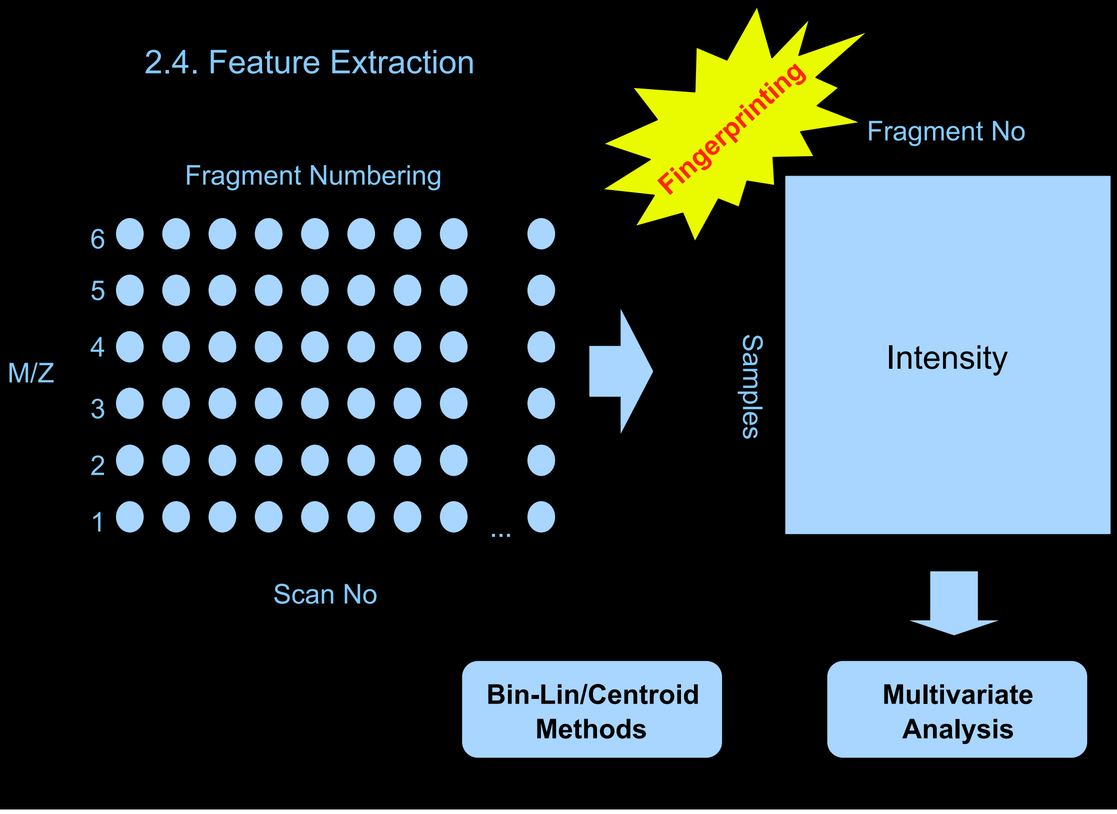
Bin-Lin/Centroid Methods

METABOLOMICS

High-Throughput Mass Spectroscopy

2. Mass Spectroscopy Signal Processing

2.4. Feature Extraction



Fragment Numbering

M/Z

6

5

4

3

2

1

...

Scan No

Fingerprinting

Fragment No

Samples

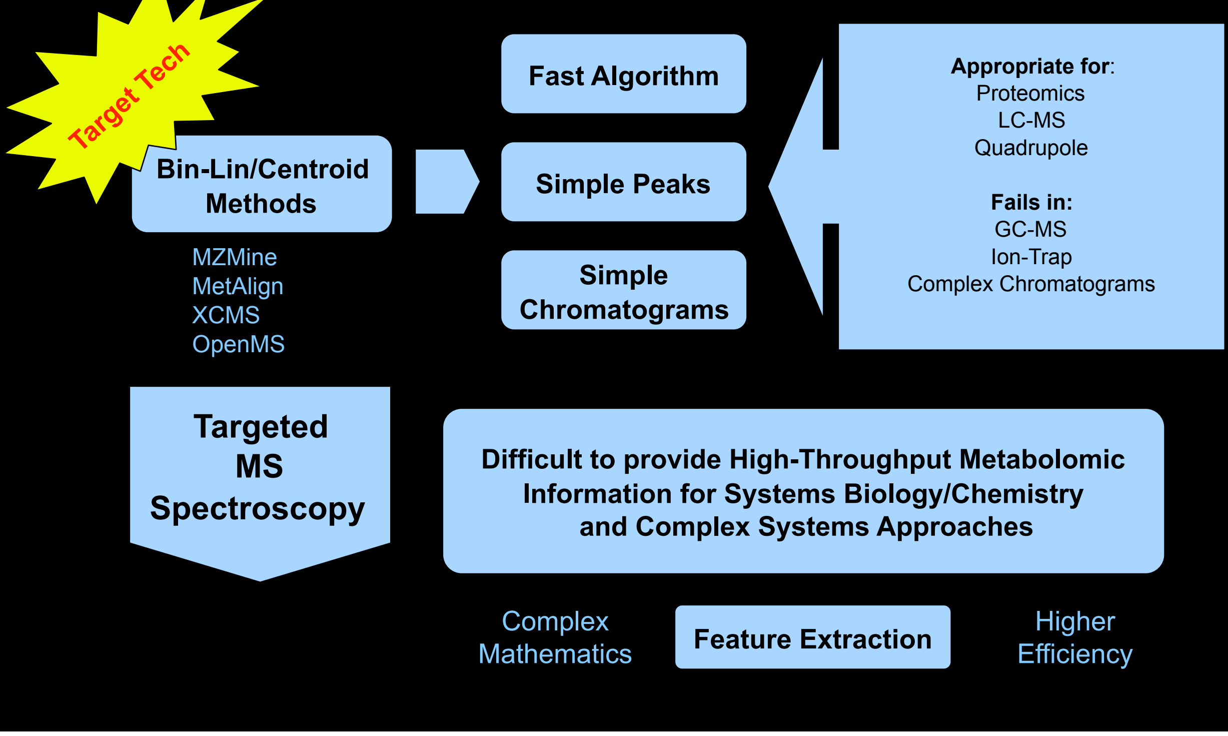
Intensity

Bin-Lin/Centroid Methods

Multivariate Analysis

2. Mass Spectroscopy Signal Processing

2.4. Feature Extraction



Target Tech

Bin-Lin/Centroid Methods

MZMine
MetAlign
XCMS
OpenMS

Fast Algorithm

Simple Peaks

Simple Chromatograms

Appropriate for:
Proteomics
LC-MS
Quadrupole

Fails in:
GC-MS
Ion-Trap
Complex Chromatograms

Targeted MS Spectroscopy

Difficult to provide High-Throughput Metabolomic Information for Systems Biology/Chemistry and Complex Systems Approaches

Complex Mathematics

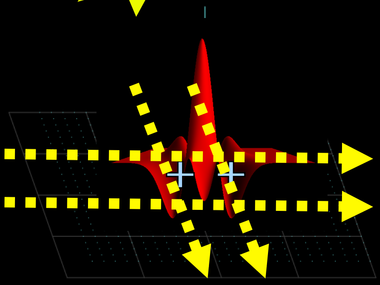
Feature Extraction

Higher Efficiency

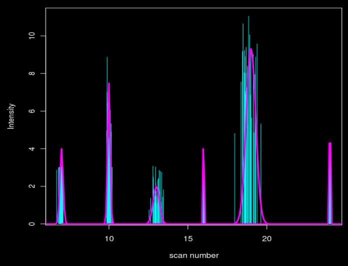
2. Mass Spectroscopy Signal Processing

2.4. Feature Extraction

Target Tech

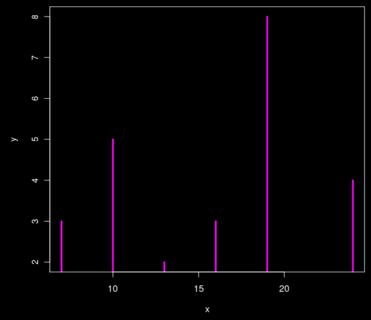


Failure in Multiscale Extraction
(e.g. loss of small peaks)



Failure
In Compound Grouping
(e.g. Loss of Compounds)

Bin-Lin/Centroid Methods



Failure
In Chromatrogram
Reconstruction
(e.g. Colapse of adjacent Scans, loss of fragments)

METABOLOMICS

High-Throughput Mass Spectroscopy

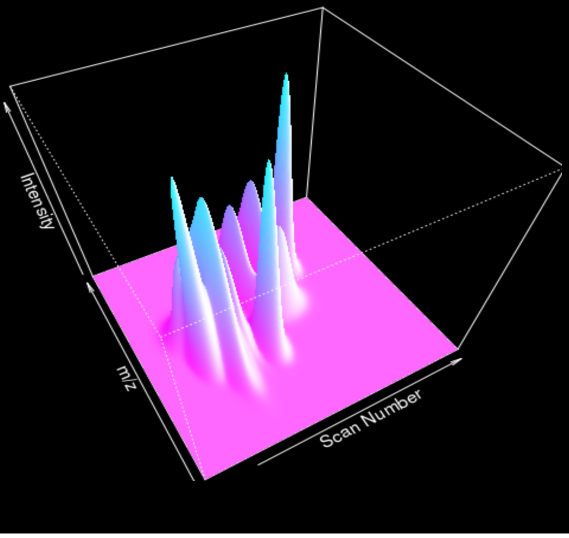
2. Mass Spectroscopy Signal Processing

2.4. Feature Extraction

How to provide High-Throughput Metabolomic Information?

Signal Processing Feature Extraction

Standards for 'In-silico' High-Throughput Chromatography In Metabolomics



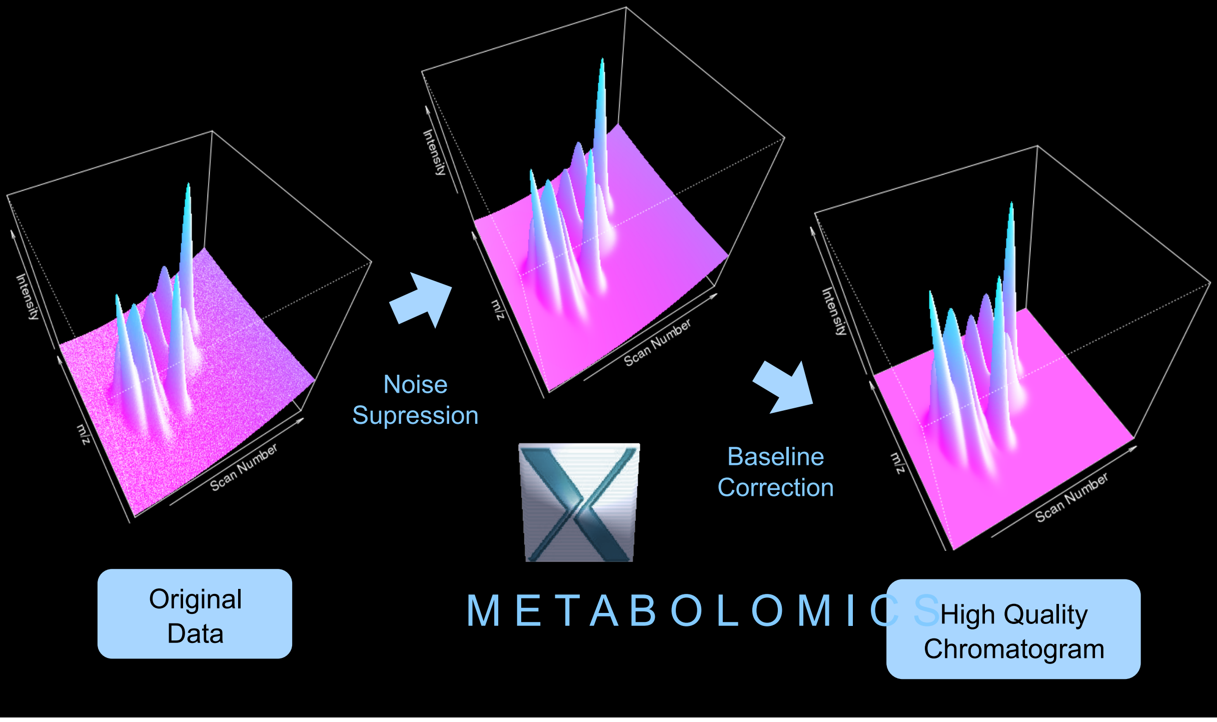
High Quality Signals
Peaks High Definition
High Peak Concentration
Filters
Deconvolution
Recognize,
Identify, Quantify,
Analise, Quality Control
Biological Interpretation

METABOLOMICS

High-Throughput Mass Spectroscopy

2. Mass Spectroscopy Signal Processing

2.3. Pre-Processing



Original Data

Noise Supression

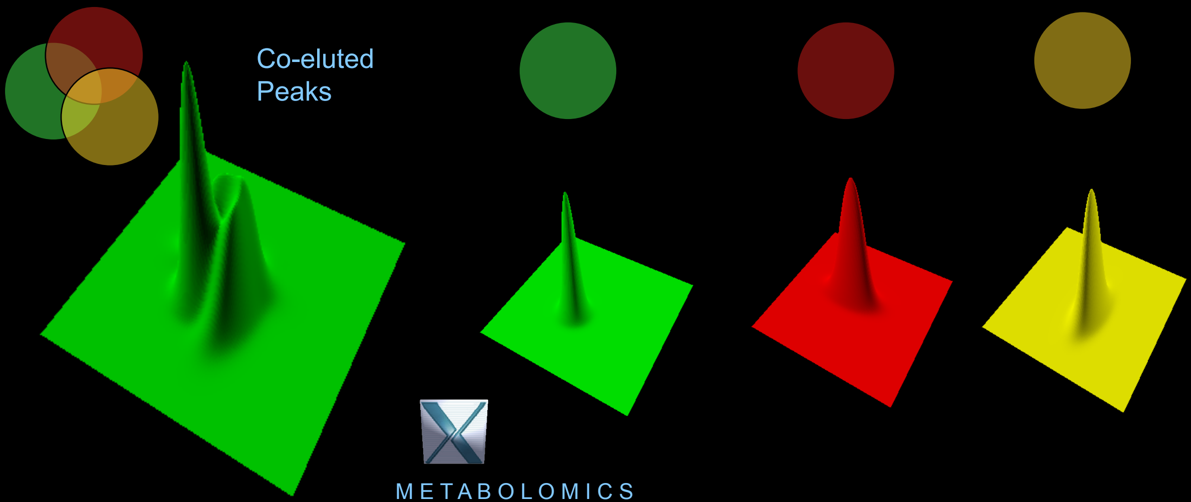
Baseline Correction

High Quality Chromatogram

METABOLOMICS

2. Mass Spectroscopy Signal Processing

2.3. Pre-Processing



Convolution Detection

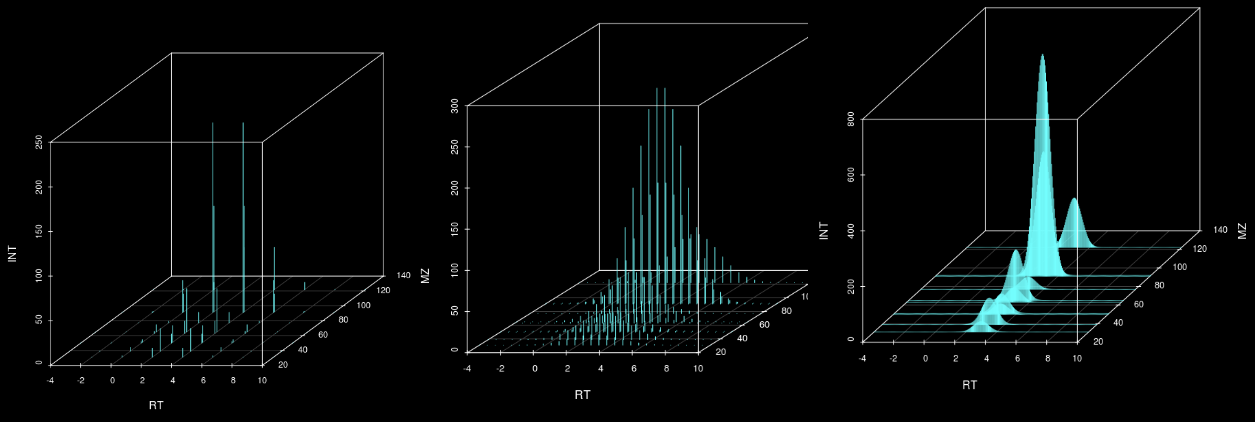
'In-Silico' Deconvolution for **High-Resolution Feature Extraction**

METABOLOMICS

High-Throughput Mass Spectroscopy

2. Mass Spectroscopy Signal Processing

2.3. Pre-Processing



Scan Rate Influence on High-Resolution Feature Extraction

Minimum of 5 to 8 scans

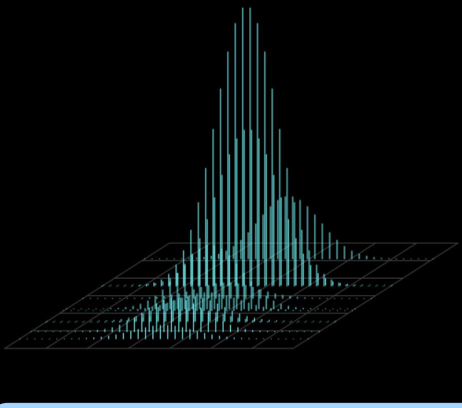
METABOLOMICS

High-Throughput Mass Spectroscopy

2. Mass Spectroscopy Signal Processing

2.3. Pre-Processing

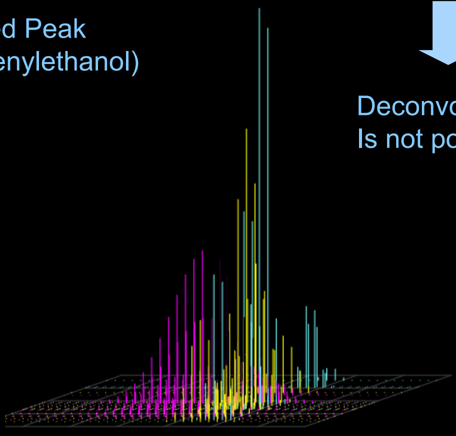
Regular Peak
(e.g. phenylethanol)



Saturated Peak
(e.g. phenylethanol)

Feature Randomization
During on-Trap
Saturation

Deconvolution
Is not possible



Saturation

Saturation Filter

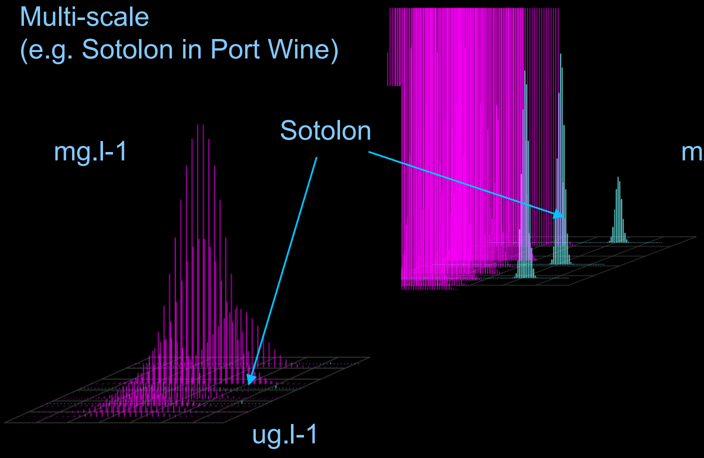
METABOLOMICS

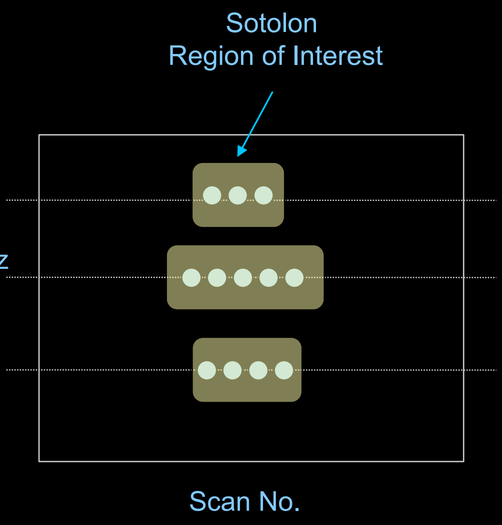
High-Throughput Mass Spectroscopy

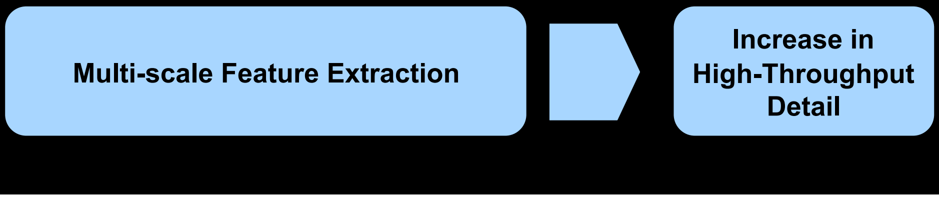
2. Mass Spectroscopy Signal Processing

2.3. Pre-Processing

Multi-scale
(e.g. Sotolon in Port Wine)







Multi-scale Feature Extraction

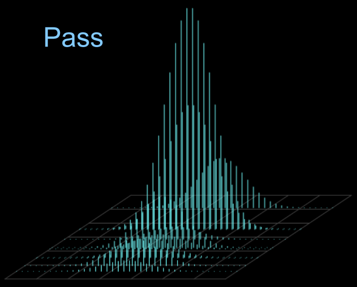
Increase in High-Throughput Detail

METABOLOMICS

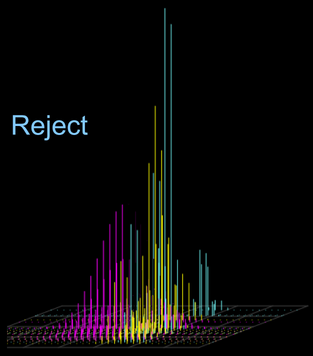
High-Throughput Mass Spectroscopy

2. Mass Spectroscopy Signal Processing

2.3. Pre-Processing



Pass



Reject

Fuzzy Filtering using Sinkhorn Factorization!!!

Peak = D_1AD_2 (perform until convergence)

- Inside a sample and Between samples!!!

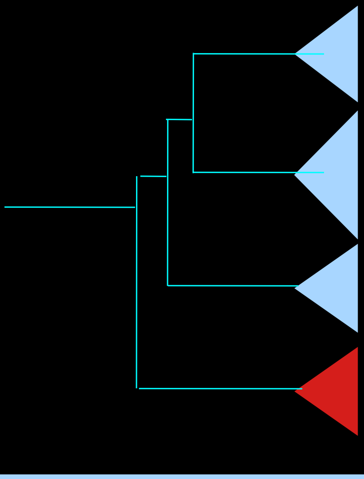
Feature Self-Consistency Test → **Only Robust Peaks Pass**

METABOLOMICS

High-Throughput Mass Spectroscopy

2. Mass Spectroscopy Signal Processing

2.3. Pre-Processing



Non-biological components

Correlation Between fragments to obtain The clusters of Non-biological components

Check each cluster According to Tikunov et al (2005)

Clean Non-Biological Components → **Robust Databases**

METABOLOMICS

High-Throughput Mass Spectroscopy

3. The X-Metabolomics Architecture

3.1. Operating Systems
3.2. Software Compatibility
3.3. File Formats
3.4. Supported Equipments
3.5. Software Architecture in a Nutshell

In: Castro, CC, Teixeira, JA, Silva-Ferreira, AC, Martins, RC. 2009. X-Metabolomics: A high-throughput GC-MS metabolomics pipeline for Saccharomyces cerevisiae. BMC BioInformatics, Submitted.

3. The X-Metabolomics Architecture

3.1. Operating Systems - UNIX Like Platforms

Currently Available for:

Ubuntu Linux:
8.04 (Discont.)
8.10 (Stable)
9.04 (Stable)
9.10 (Testing)

Mac OS X
Snow Leopard

3. The X-Metabolomics Architecture

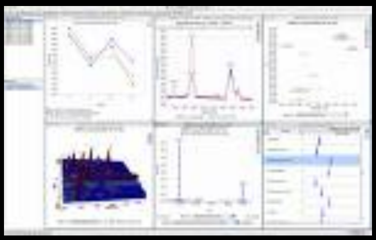
3.2. Software Compatibility

Xorg

R-project

BioInformatics Platforms

OpenPAT
Plug-In



MZmine

MetAlign

Markup Languages
and DataBases

Mass Spectroscopy
Processing Software

METABOLOMICS

High-Throughput Mass Spectroscopy

3. The X-Metabolomics Architecture

3.3. File Formats

ASCII Text Files

MZ XML (Proteomics)

NetCDF (Preferencial!!!)
(Use Mass Transit for conversion)

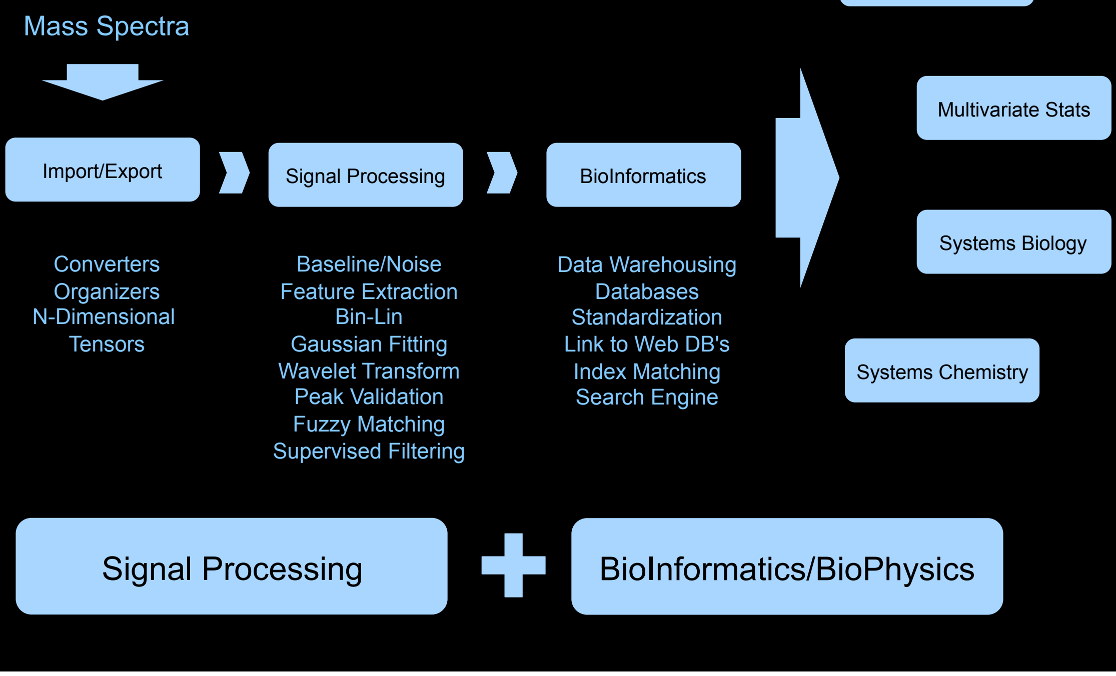
NetCDF (network Common Data Form)
Multi-Dimensional Array-oriented Scientific Data
Useful for Complex MS/MS datasets

METABOLOMICS

High-Throughput Mass Spectroscopy

3. The X-Metabolomics Architecture

3.5. Architecture in a nutshell



Mass Spectra

Import/Export

Converters
Organizers
N-Dimensional
Tensors

Signal Processing

Baseline/Noise
Feature Extraction
Bin-Lin
Gaussian Fitting
Wavelet Transform
Peak Validation
Fuzzy Matching
Supervised Filtering

BioInformatics

Data Warehousing
Databases
Standardization
Link to Web DB's
Index Matching
Search Engine

Process Analytical Technology

Multivariate Stats

Systems Biology

Systems Chemistry

Signal Processing + BioInformatics/BioPhysics

METABOLOMICS

High-Throughput Mass Spectroscopy

4. Software Features

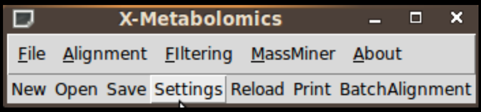
4.1. Configuration
4.2. File Import/Export
4.3. Pre-processing
4.4. Filtering
4.4. Fingerprinting Diagnostics
4.5. MS Quality Control Charts
4.6. Unsupervised Metabolomics
4.7. Identification and Composition Table
4.8. Co-Expression Pathway Analysis
4.9. Time Course Metabolomic Analysis

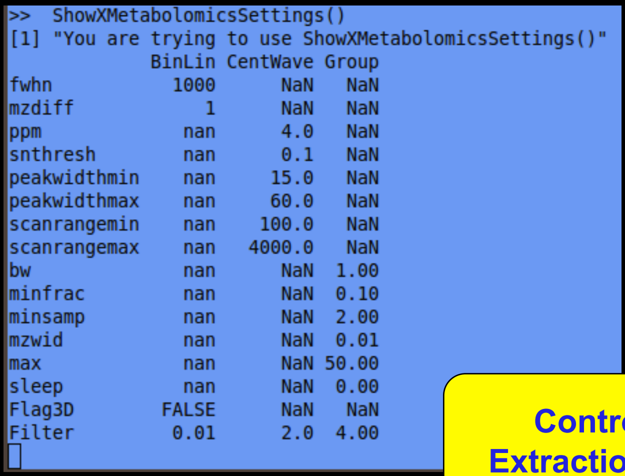
METABOLOMICS

High-Throughput Mass Spectroscopy

4. Software Features

4.1. Configuration





```
>>  ShowXMetabolomicsSettings()
[1] "You are trying to use ShowXMetabolomicsSettings()"
             BinLin CentWave Group
fwhn           1000      NaN   NaN
mzdiff            1      NaN   NaN
ppm             nan      4.0   NaN
snthresh        nan      0.1   NaN
peakwidthmin    nan     15.0   NaN
peakwidthmax    nan     60.0   NaN
scanrangemin    nan    100.0   NaN
scanrangemax    nan   4000.0   NaN
bw              nan      NaN  1.00
minfrac         nan      NaN  0.10
minsamp         nan      NaN  2.00
mzwid           nan      NaN  0.01
max             nan      NaN 50.00
sleep           nan      NaN  0.00
Flag3D        FALSE      NaN   NaN
Filter         0.01      2.0  4.00
```

Settings: Unix Style

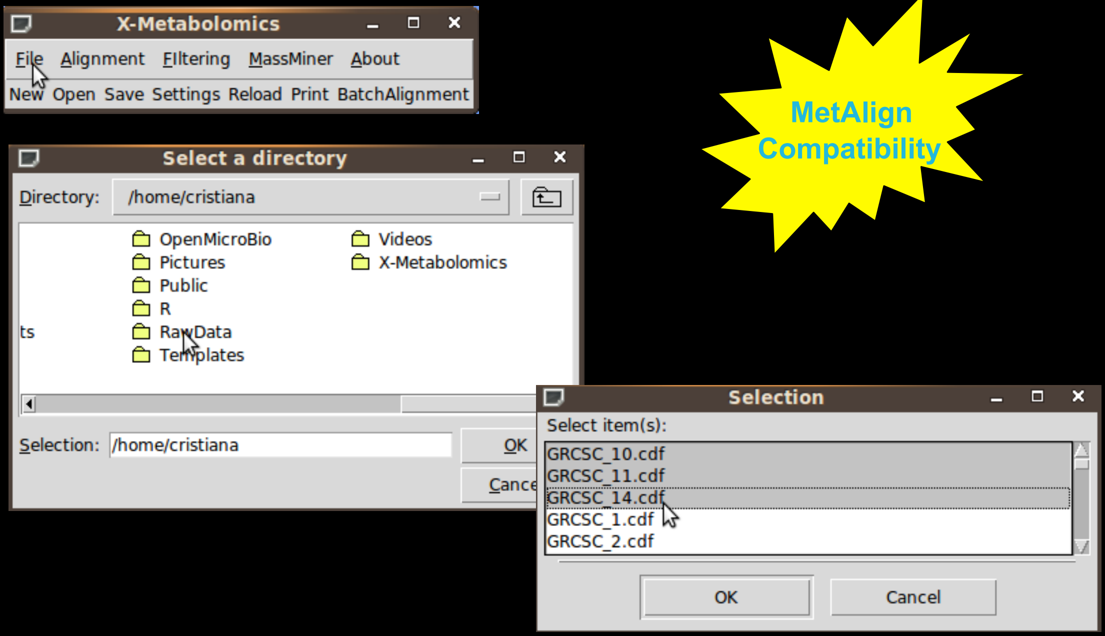
Control Feature Extraction and Filters

METABOLOMICS

High-Throughput Mass Spectroscopy

4. Software Features

4.2. File Import



Directory: /home/cristiana

Selection: /home/cristiana

GRCSC_10.cdf
GRCSC_11.cdf
GRCSC_14.cdf
GRCSC_1.cdf
GRCSC_2.cdf

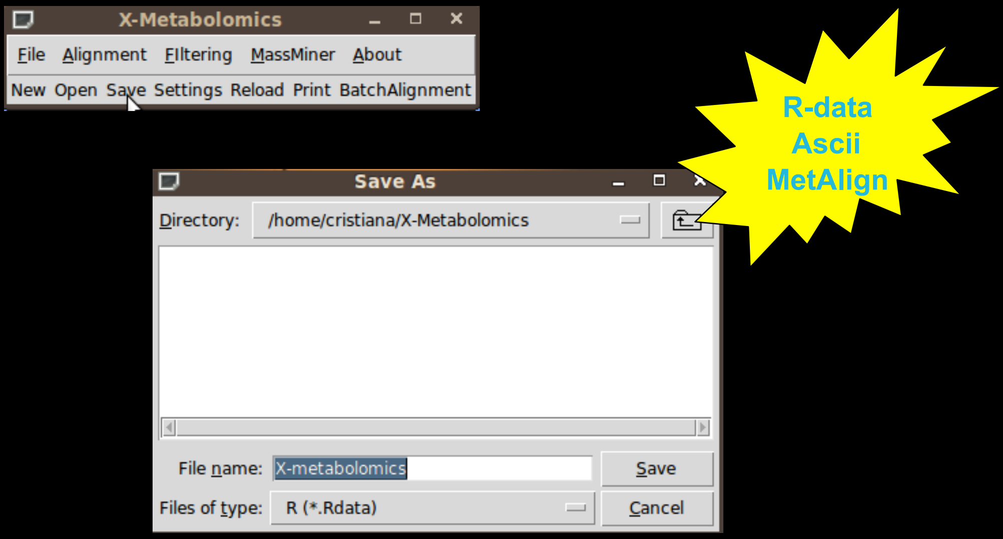
MetAlign Compatibility

METABOLOMICS

High-Throughput Mass Spectroscopy

4. Software Features

4.2. File Export



Directory: /home/cristiana/X-Metabolomics

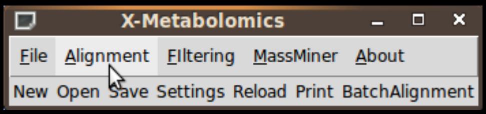
File name: X-metabolomics

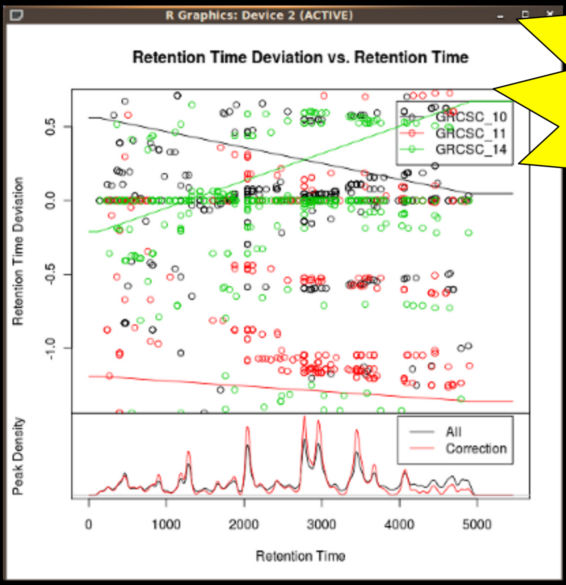
Files of type: R (*.Rdata)

R-data
Ascii
MetAlign

4. Software Features

4.3. Pre-processing


X-Metabolomics
File Alignment FIltering MassMiner About
New Open Save Settings Reload Print BatchAlignment



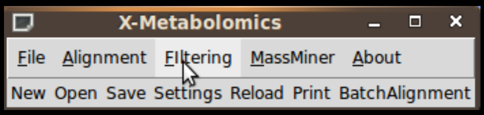
Synchronization

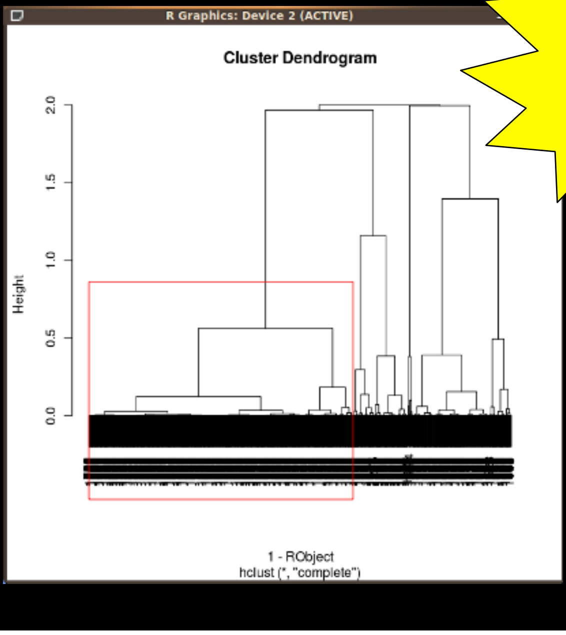
METABOLOMICS

High-Throughput Mass Spectroscopy

4. Software Features

4.4. Filtering


X-Metabolomics
File Alignment Filtering MassMiner About
New Open Save Settings Reload Print BatchAlignment



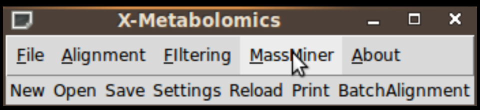
HCA Filtering

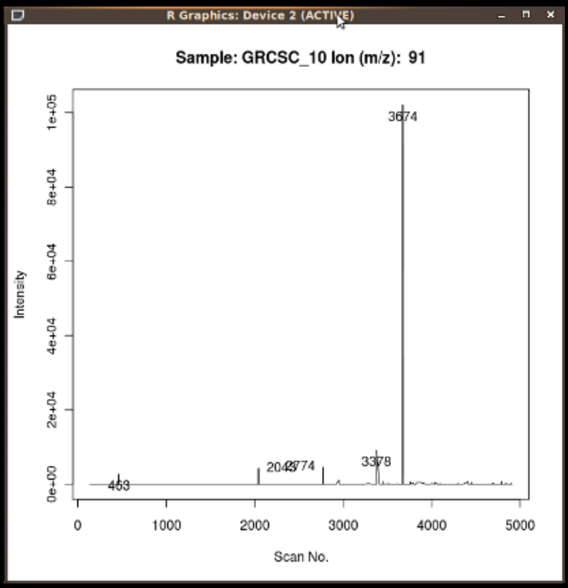
METABOLOMICS

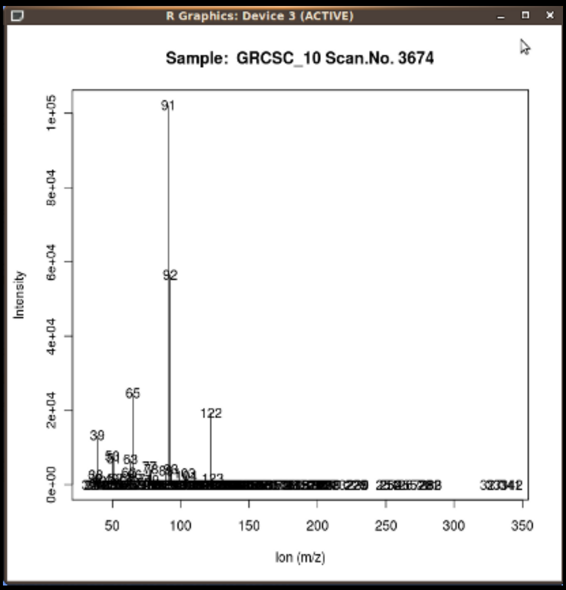
High-Throughput Mass Spectroscopy

4. Software Features

4.5. Fingerprinting Diagnostics


X-Metabolomics
File Alignment FIltering MassMiner About
New Open Save Settings Reload Print BatchAlignment



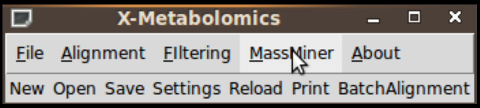


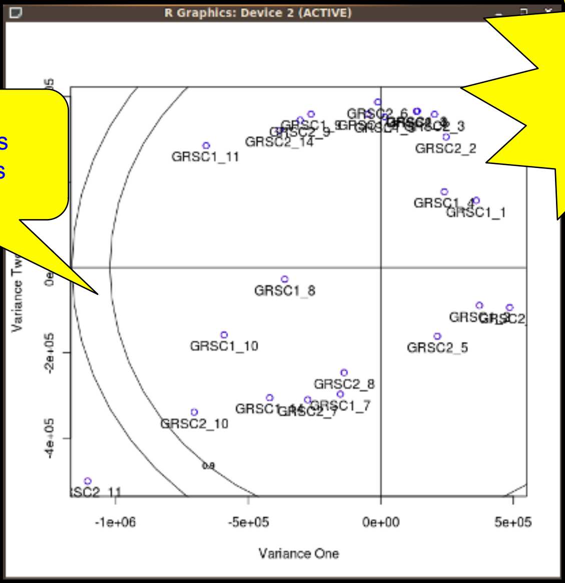
METABOLOMICS

High-Throughput Mass Spectroscopy

4. Software Features

4.6. MS Quality Control Charts





Control Methods and Equipments

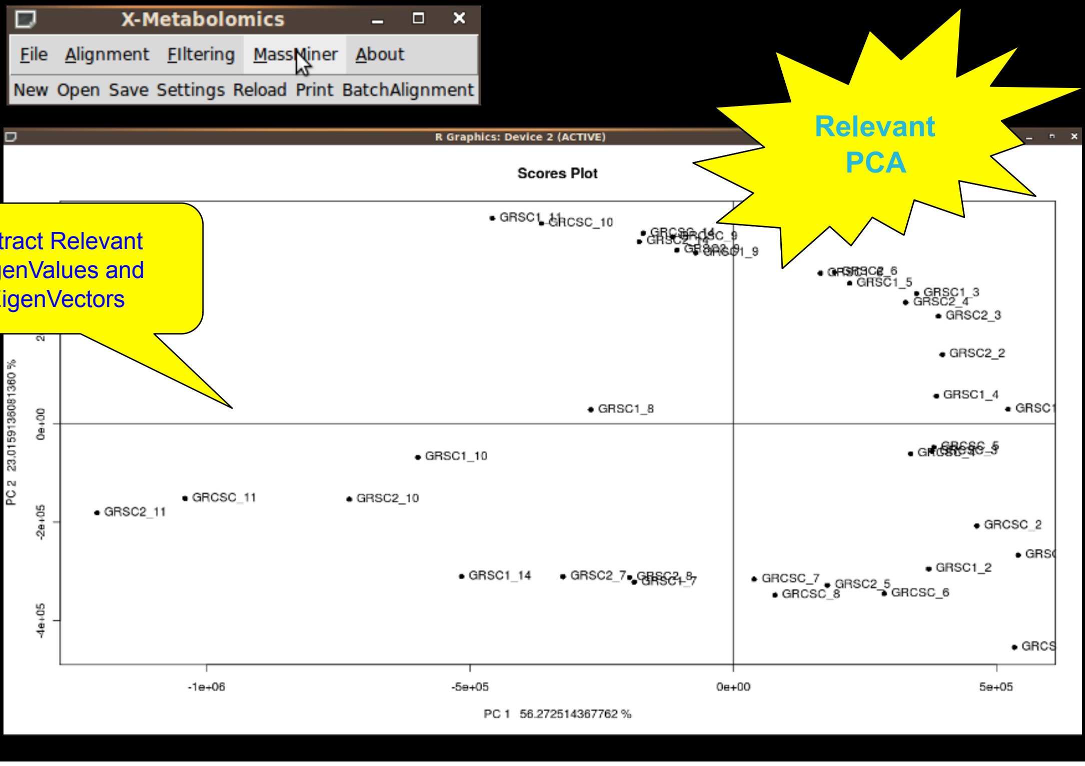
Multivariate Chart Control

METABOLOMICS

High-Throughput Mass Spectroscopy

4. Software Features

4.7. Unsupervised Metabolomics



Extract Relevant EigenValues and EigenVectors

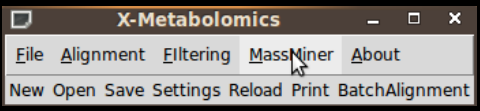
Relevant PCA

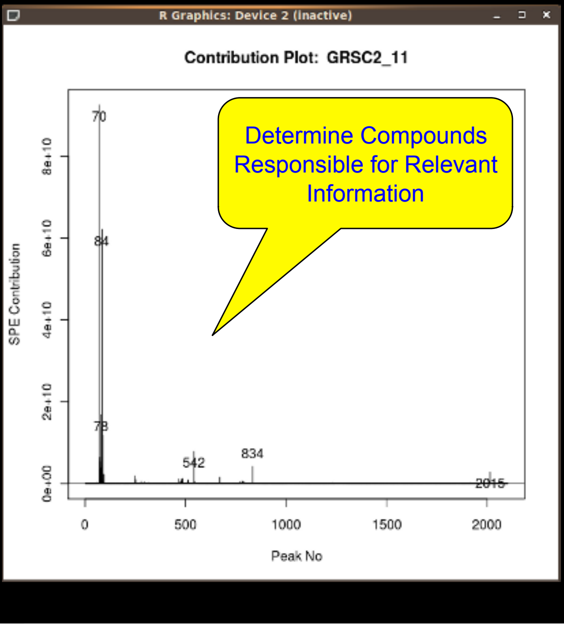
METABOLOMICS

High-Throughput Mass Spectroscopy

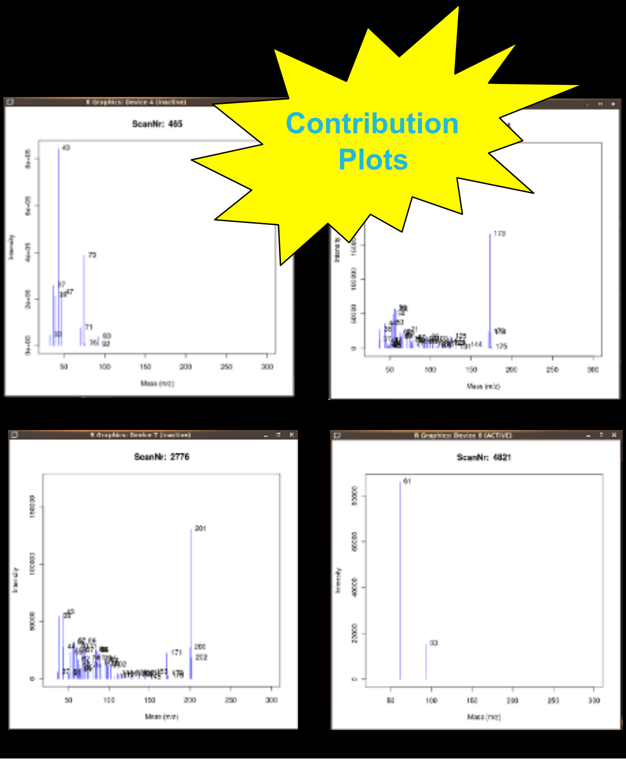
4. Software Features

4.7. Unsupervised Metabolomics


X-Metabolomics
File Alignment FIltering MassMiner About
New Open Save Settings Reload Print BatchAlignment



Determine Compounds Responsible for Relevant Information



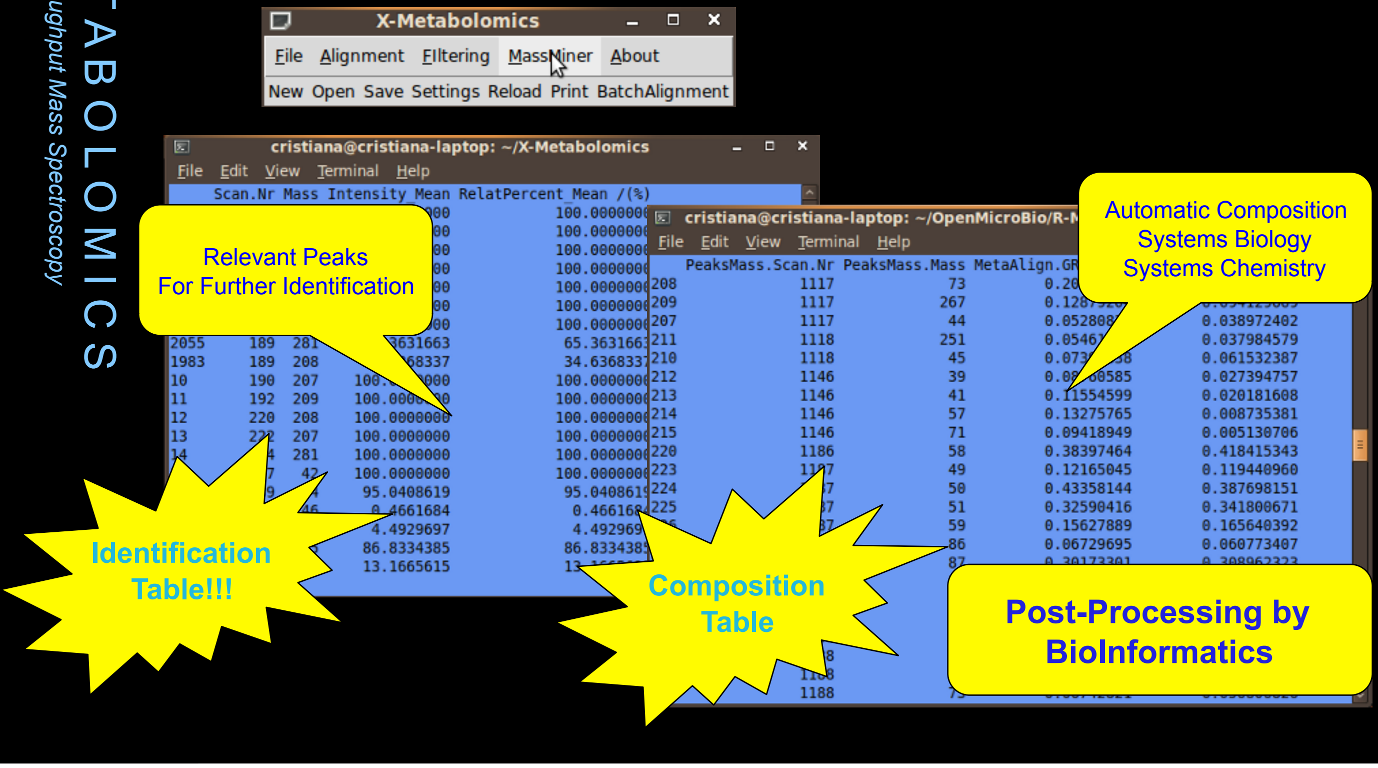
Contribution Plots

METABOLOMICS

High-Throughput Mass Spectroscopy

4. Software Features

4.8. Identification and Composition Table


X-Metabolomics
File Alignment FIltering MassMiner About
New Open Save Settings Reload Print BatchAlignment
cristiana@cristiana-laptop: ~/X-Metabolomics
File Edit View Terminal Help
cristiana@cristiana-laptop: ~/OpenMicroBio/R-M
File Edit View Terminal Help

```
      Scan.Nr Mass Intensity_Mean RelatPercent_Mean /(%)
                            000           100.0000000
                             00           100.0000000
                             00           100.0000000
                             00           100.0000000
                             00           100.0000000
                             00           100.0000000
                             00           100.0000000
2055      189  281         3631663         65.3631663
1983      189  208          68337          34.6368337
10        190  207    100.    000         100.0000000
11        192  209    100.0000  0         100.0000000
12        220  208    100.0000000         100.0000000
13        222  207    100.0000000         100.0000000
14          4  281    100.0000000         100.0000000
            7   42    100.0000000         100.0000000
            9    4     95.0408619          95.0408619
                46      0.4661684           0.4661684
                        4.4929697           4.492969
                       86.8334385          86.8334385
                       13.1665615          13
```

```
    PeaksMass.Scan.Nr PeaksMass.Mass MetaAlign.GR
208              1117             73        0.20
209              1117            267        0.1287
207              1117             44        0.052808             0.038972402
211              1118            251        0.05461              0.037984579
210              1118             45        0.073    38          0.061532387
212              1146             39        0.08  60585          0.027394757
213              1146             41        0.11554599           0.020181608
214              1146             57        0.13275765           0.008735381
215              1146             71        0.09418949           0.005130706
220              1186             58        0.38397464           0.418415343
223              1187             49        0.12165045           0.119440960
224                 7             50        0.43358144           0.387698151
225                 7             51        0.32590416           0.341800671
                   87             59        0.15627889           0.165640392
                                  86        0.06729695           0.060773407
                                  87        0.30173301           0.308962323
                    8
                 1188
                 1188
```

Relevant Peaks
For Further Identification

Identification
Table!!!

Composition
Table

Automatic Composition
Systems Biology
Systems Chemistry

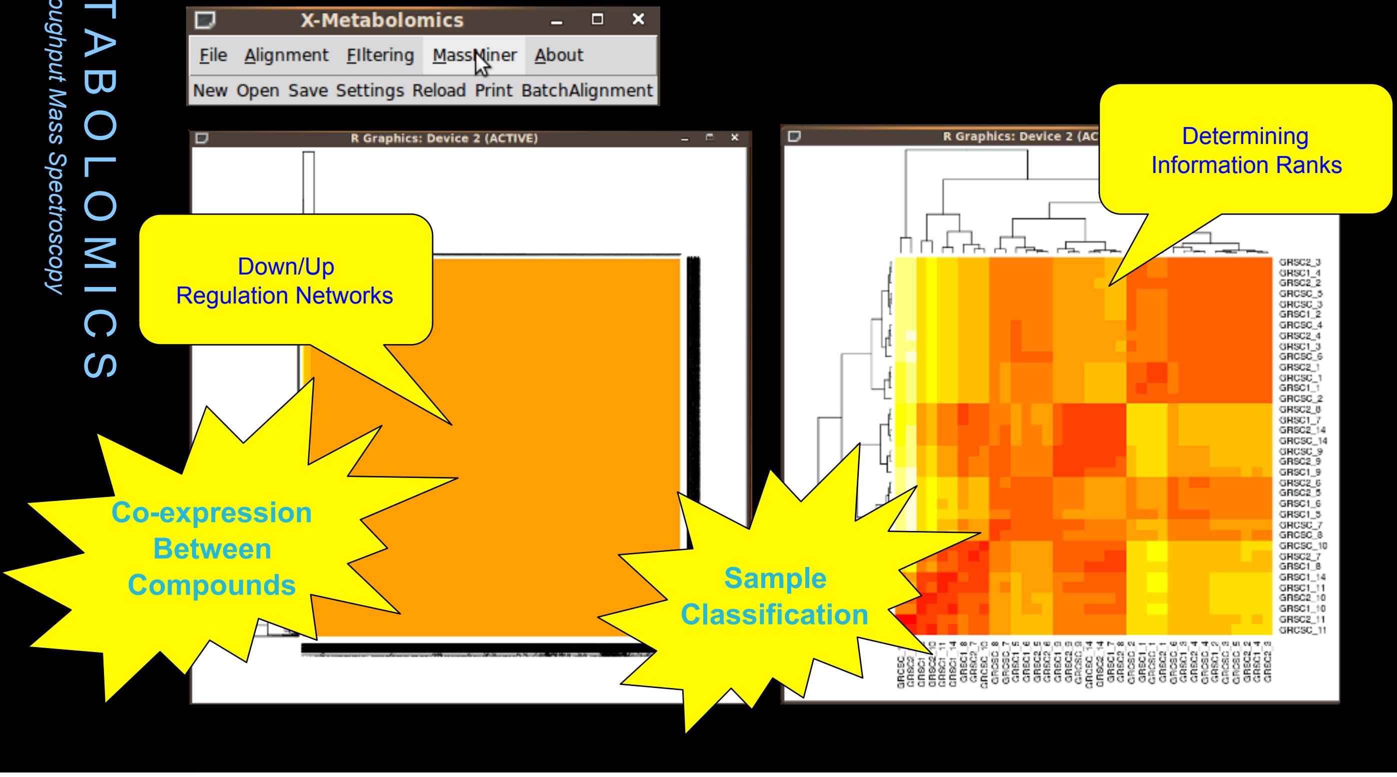
Post-Processing by
BioInformatics

METABOLOMICS

High-Throughput Mass Spectroscopy

4. Software Features

4.9. Co-Expression Pathway Analysis


X-Metabolomics
File Alignment Filtering MassMiner About
New Open Save Settings Reload Print BatchAlignment
R Graphics: Device 2 (ACTIVE)
R Graphics: Device 2 (AC

Determining Information Ranks

Down/Up Regulation Networks

Co-expression Between Compounds

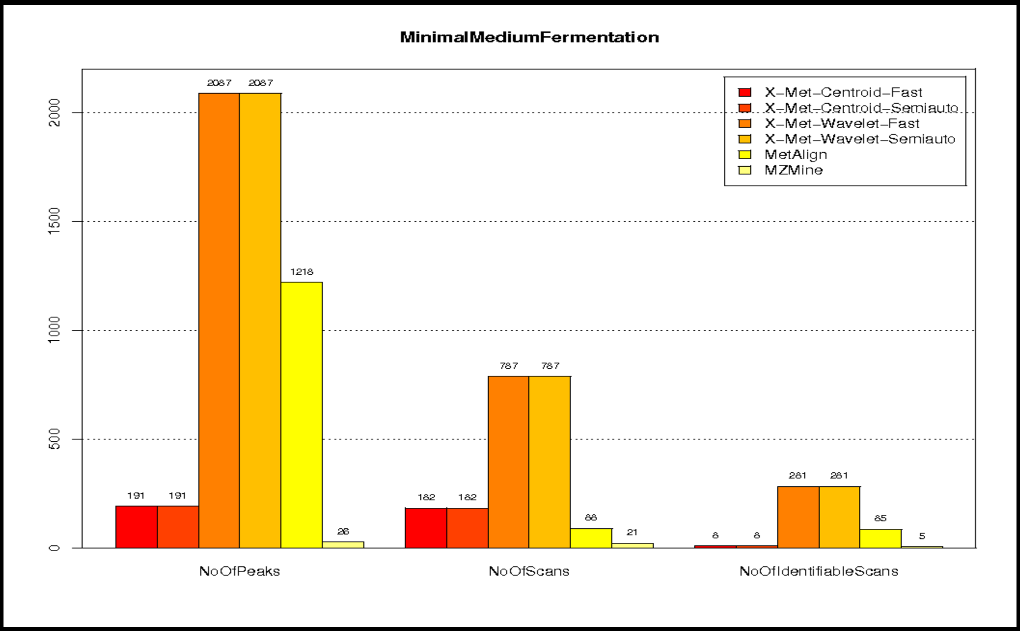
Sample Classification

METABOLOMICS

High-Throughput Mass Spectroscopy

5. Benchmarks

Minimum Media *Saccharomyces Cerevisiae* Growth



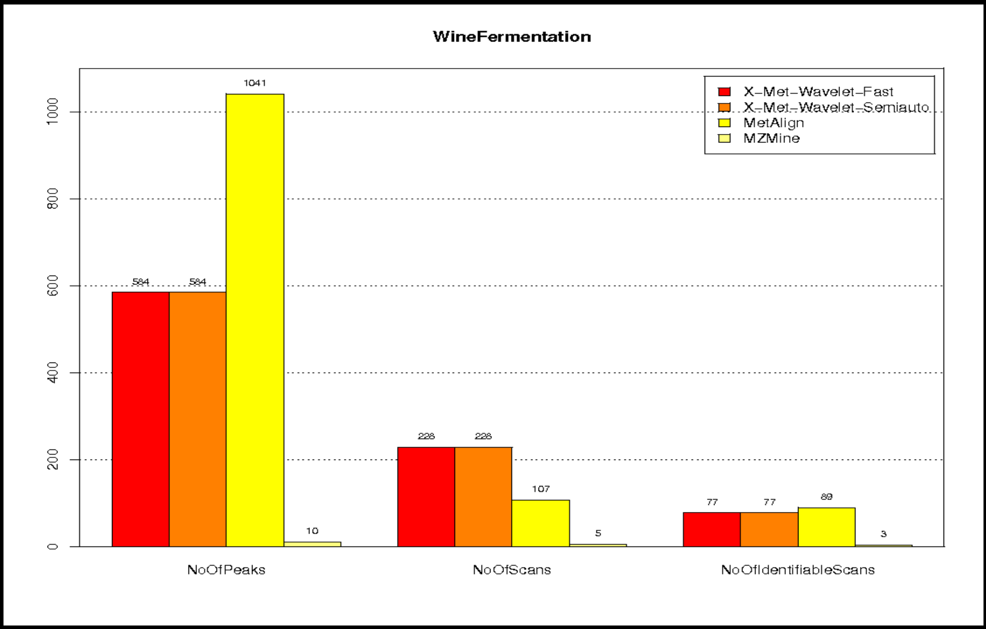
In: Castro, CC, Silva-Ferreira, AC,,Teixeira, JA, Martins, RC. 2009. X-Metabolomics: A High-throughput GC-MS metabolomics pipeline for Saccharomyces cerevisiae . BMC BioInformatics, Submitted.

METABOLOMICS

High-Throughput Mass Spectroscopy

5. Benchmarks

Wine Fermentation



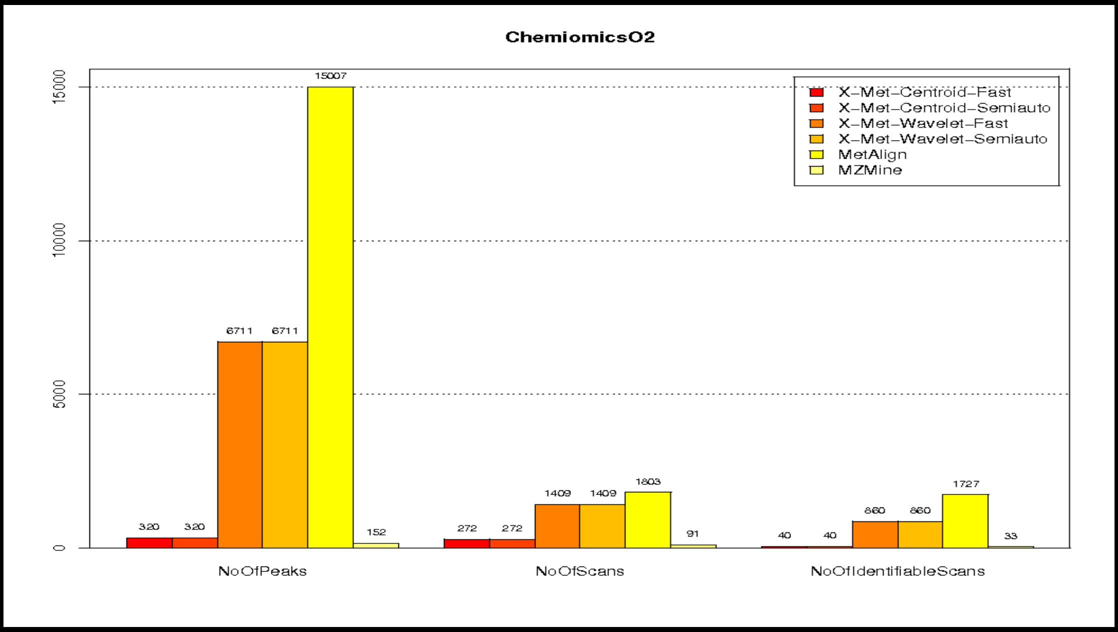
In: Castro, CC, Silva-Ferreira, AC,,Teixeira, JA, Martins, RC. 2009. X-Metabolomics: A High-throughput GC-MS metabolomics pipeline for Saccharomyces cerevisiae . BMC BioInformatics, Submitted.

METABOLOMICS

High-Throughput Mass Spectroscopy

5. Benchmarks

Madeira Wine GC-MS Data



In: Castro, CC, Silva-Ferreira, AC,,Teixeira, JA, Martins, RC. 2009. X-Metabolomics: A High-throughput GC-MS metabolomics pipeline for Saccharomyces cerevisiae . BMC BioInformatics, Submitted.

METABOLOMICS

High-Throughput Mass Spectroscopy

5. Benchmarks

Conclusions:

X-Metabolomics and **MetAlign** were developed for Metabolomics and Metabonomics research – **Out-perform all other software**, which were mainly **developed for Proteomics**

Performance of **X-Metabolomics** is **in most cases similar** to **MetAlign with well defined peaks**

X-Metabolomics out-performs MetAlign when chromatograms exhibit **ion-trap artifacts** (due to X-metabolomics filters)

Further developments in «In-Silico» Chromatography are only possible by the development of new feature extraction methods!

METABOLOMICS

High-Throughput Mass Spectroscopy

6. Software Demonstration

Minimum Media Fermentation Time Course Analysis

In: Silva-Ferreira, AC, Gunning, C., Castro, CC, Teixeira, JA,, Martins, RC. 2009. A non-target approach for time-course *Saccharomyces cerevisiae* oxidative response by GC-MS and Cyclic Voltammetry. Metabolomics, Submitted.