Independent Component Analysis (ICA) for the extraction of protein profiles from MALDI-TOF MS spectra

Gopal Peddinti

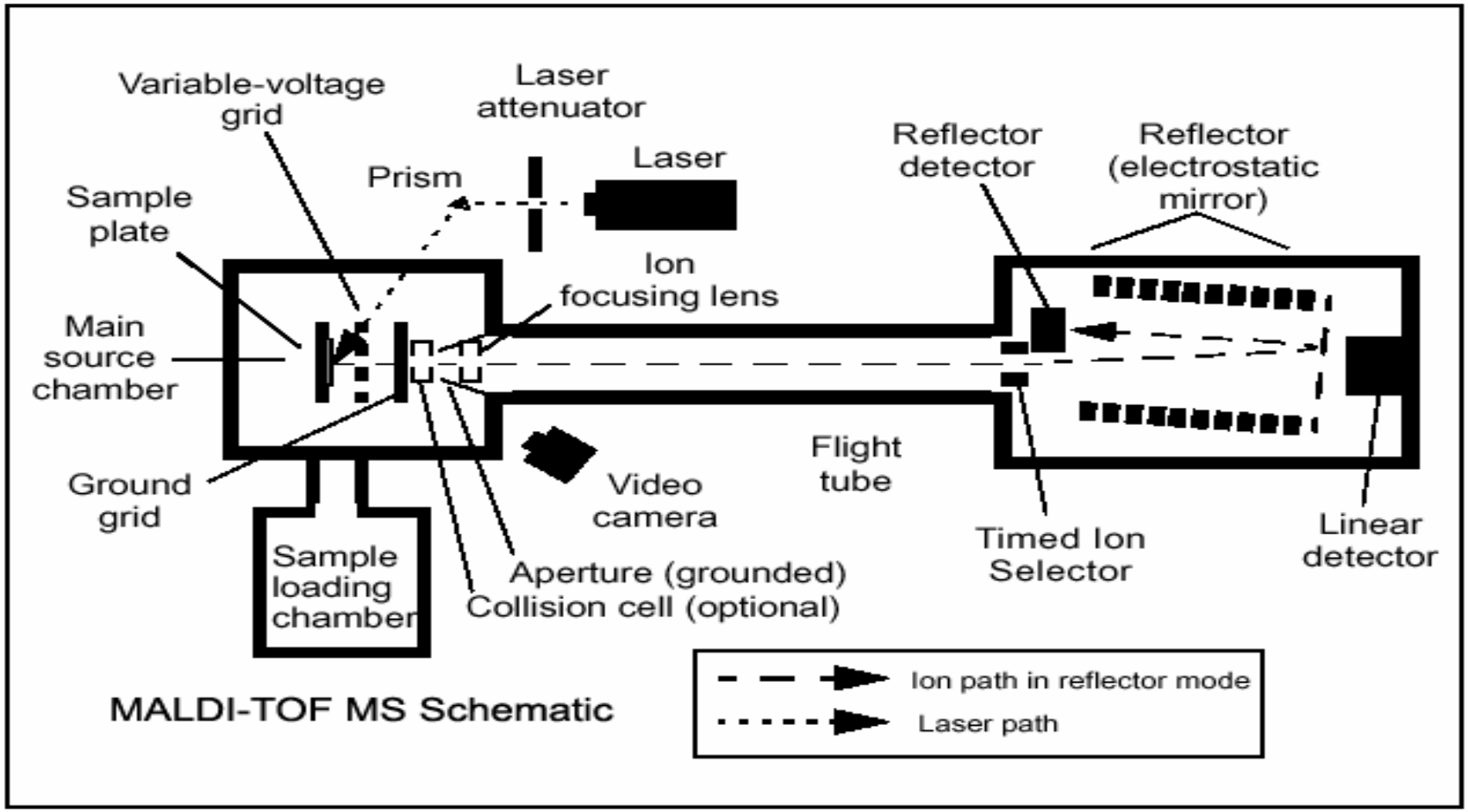
T.61-6070 Modeling of Proteomics Data

10.04.2008

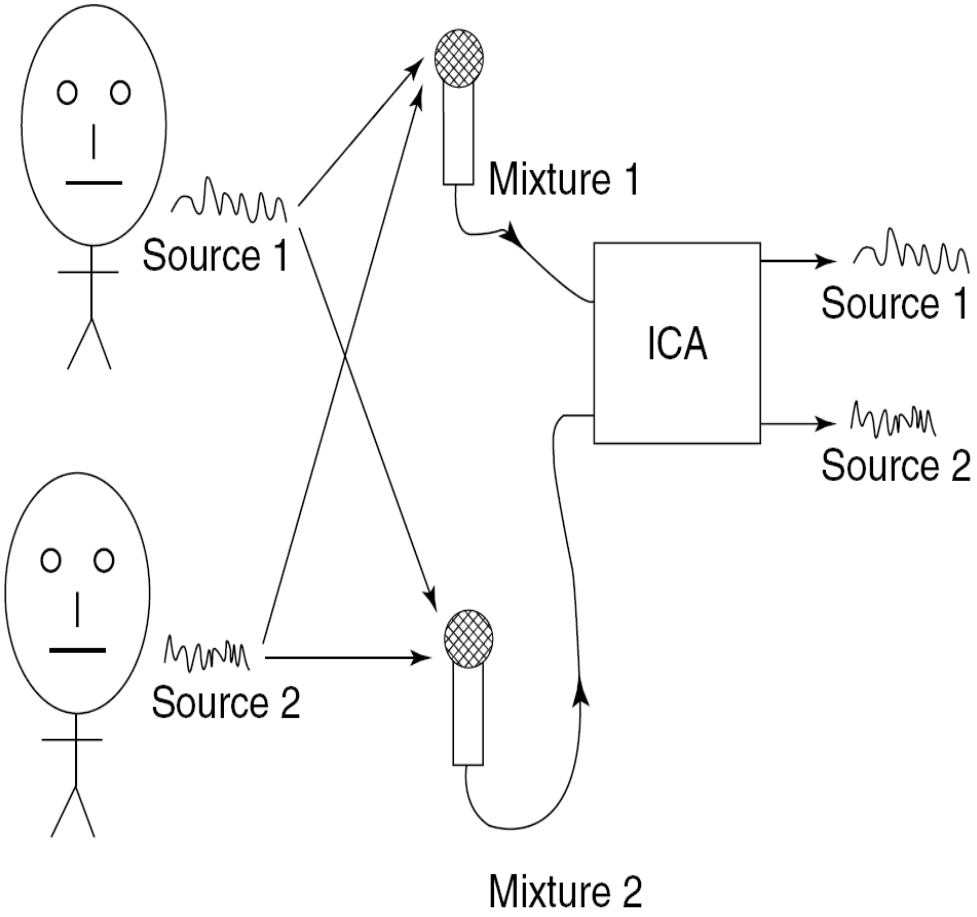
Overview

- MALDI-TOF Mass Spectrometer
- Independent Component Analysis
- MALDI-TOF MS Data
  - Simulated Data
  - Data from Inflammatory auto-immune disease patients
- Analysis of Spectra
  - Independent Components
  - Characterization of ICs
    - Smoothing
    - Baseline subtraction
    - Removal of residual noise
    - Peak picking
  - Biomarker identification
- Summary and Comments

MALDI-TOF Mass Spectrometer



ICA: Introduction



ICA Model

$$\text{ICA Model}: X = AS$$

$$\text{Matrix of observed signals}: X = [x_1, ..., x_n]^T$$

$$\text{Matrix of underlying signals}: S = [s_1, ..., s_m]^T$$

$$\text{Mixing matrix } A = [a_{ij}]_{n \times m}$$

- Generative model
  - Describes how observed data are generated by a process of mixing underlying signals s_i
  - s_i must be independent
  - s_i are called independent components
- When m ≤ n and mixing matrix (A) is full-column rank, one can determine unmixing matrix W such that $S = WX$

ICA: Preprocessing for ICA

- **Centering** and **Whitening**
- Centering: Make each observation zero-mean
- Whitening: Linear transformation which makes observations uncorrelated and with unit variance
  - Common approach: Eigen value decomposition of the covariance matrix

$$E\{XX^T\} = EDE^T,$$

where E : orthogonal matrix of eigenvectors of $E\{XX^T\}$

$D : diag(d_1, d_2, ..., d_n)$

d_i : eigen values

Whitened matrix : $\overline{X} = ED^{-1/2}E^T X$

where $D^{-1/2} = diag(d_1^{-1/2}, d_2^{-1/2}, ..., d_n^{-1/2})$

$\Rightarrow S = WX = WE^T D^{1/2} E\overline{X} = \overline{WX},$

$\overline{W}$ is orthogonal (lower no. of degrees of freedom)

Finally the required unmixing matrix : $W = \overline{W}ED^{-1/2}E^T$

FastICA Algorithm

- ICA model can be estimated *iff* ICs are non-Gaussian
- Estimation principle: ICs are maximally non-Gaussian components
- Kurtosis (4th order cumulant) is a measure for non-gaussianity

$$y : \text{zero - mean random variable}$$

$$kurt(y) = E\{y^4\} - 3(E\{y^2\})^2$$

- Basic optimization technique: Gradient method
- Fixed-point algorithm for optimization
  - Find maxima of non-gaussianity using the absolute value of kurtosis

ICA: Post processing

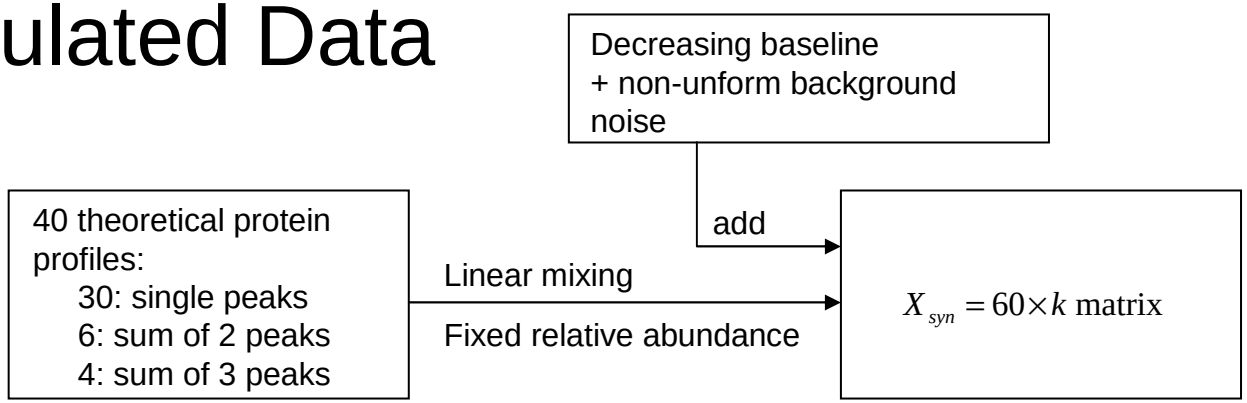
- After ICA decomposition A can be obtained from W as follows:

$$A = (W^T W)^{-1} W^T$$

- Power of i^{th} IC can be computed as follows:

$$p_i = \sum_{j=1}^{n} a_{ij}^2$$

Simulated Data



$$x(z) = \frac{A_0}{\tau} \exp\left(\frac{\sigma_p^{\,2}}{2\tau^2} - \frac{z - z_p}{\tau} \right) \int_{-\infty}^{h} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2} \right) dt$$

z : m/z value

A_0 : area of the peak $\leftarrow [180,700]$

τ : time constant of the exponential decay $\leftarrow 0.0172$

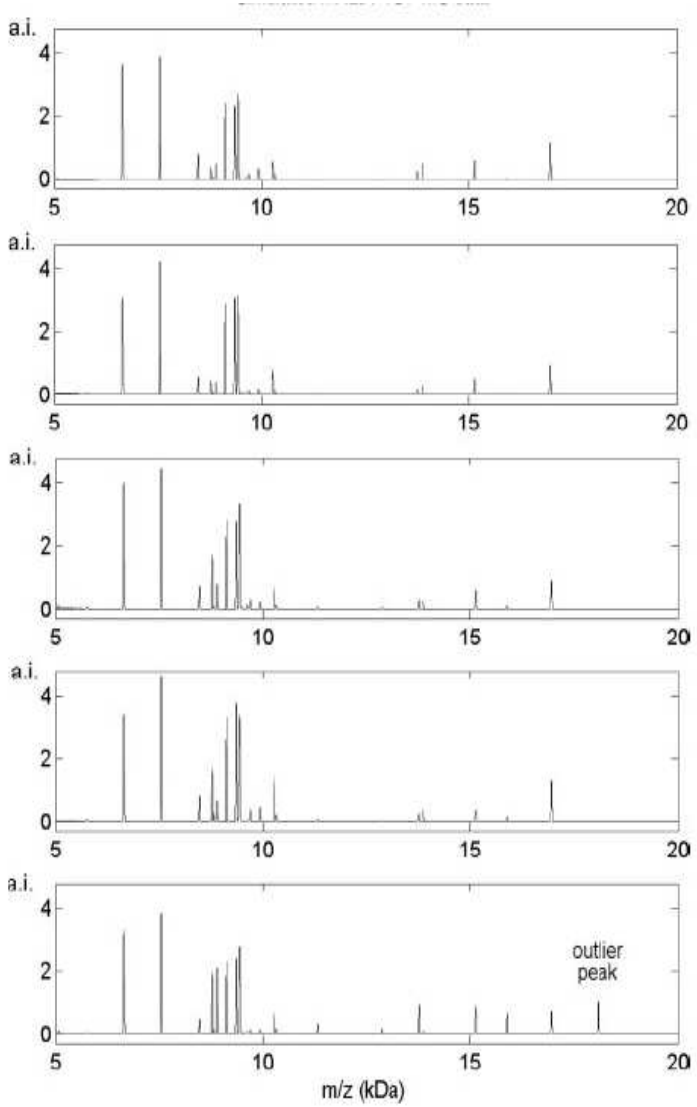
σ_p : controls the tailing of the peak $\leftarrow 0.0189$

z_p : determines the position of the peak on m/z axis $\leftarrow [6000,18000]$

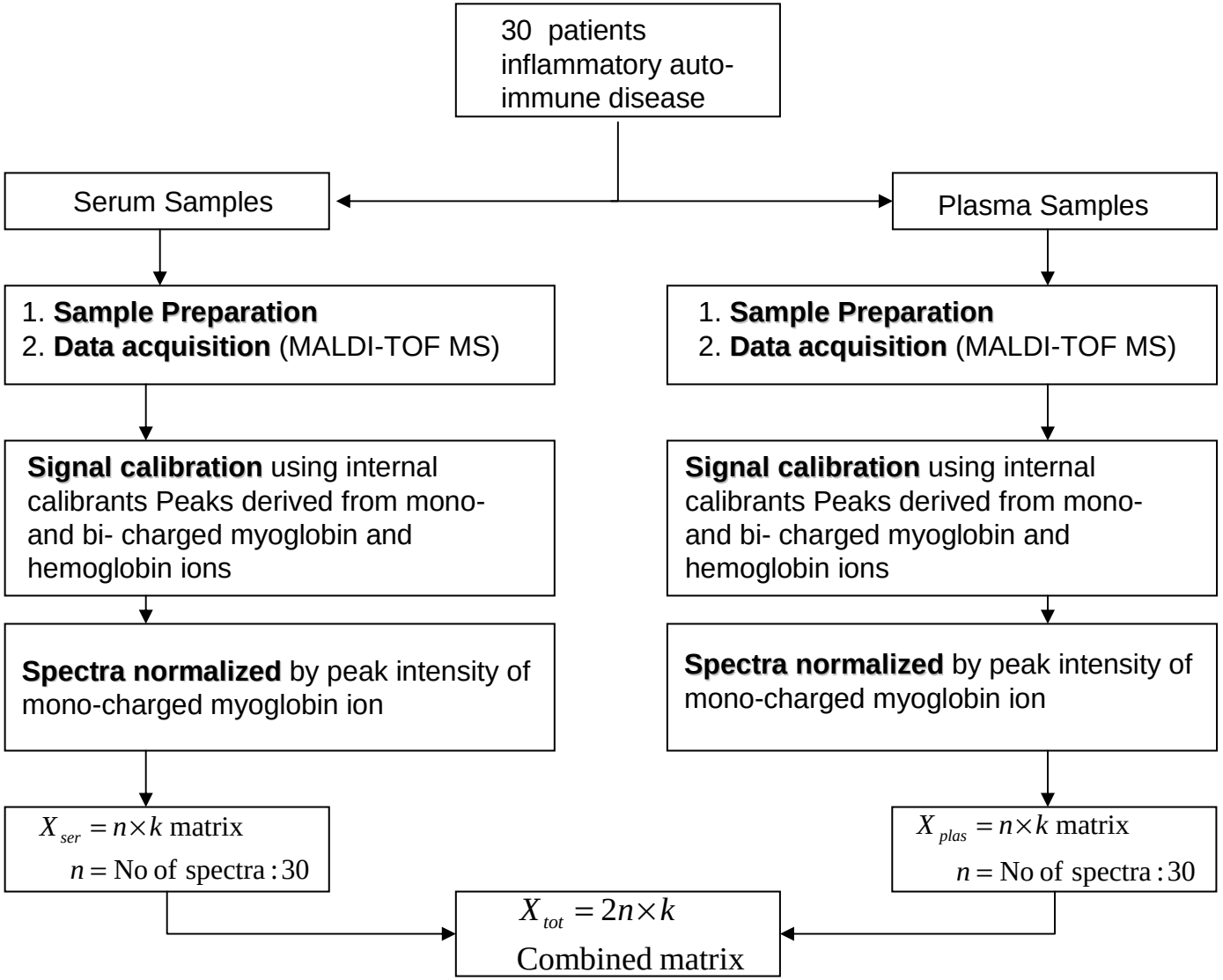
τ / σ_p : measure of asymmetry

$$h = \frac{z - z_p}{\sigma_p} - \frac{\sigma_p}{\tau}$$

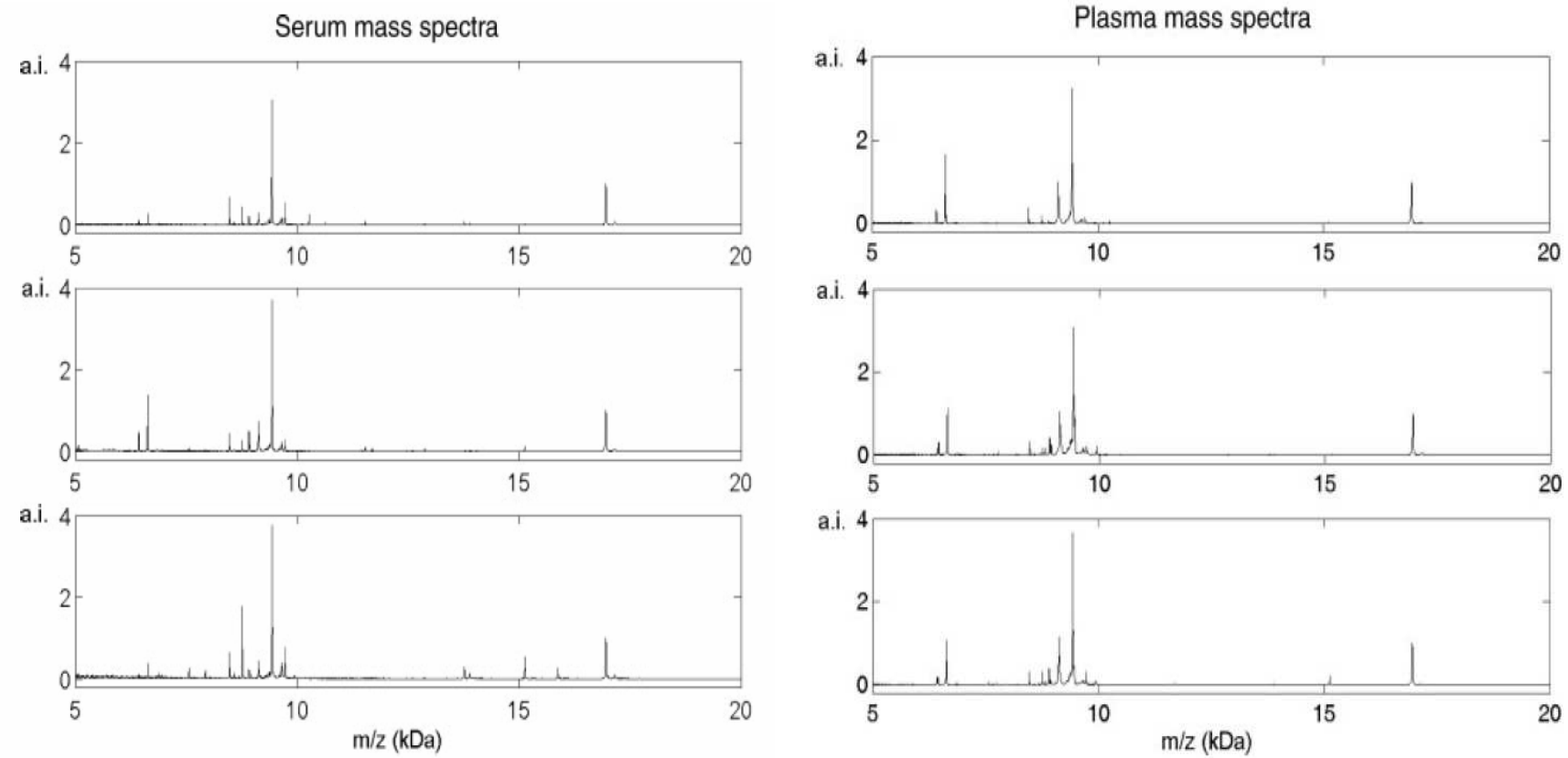
Simulated MALDI-TOF MS Data



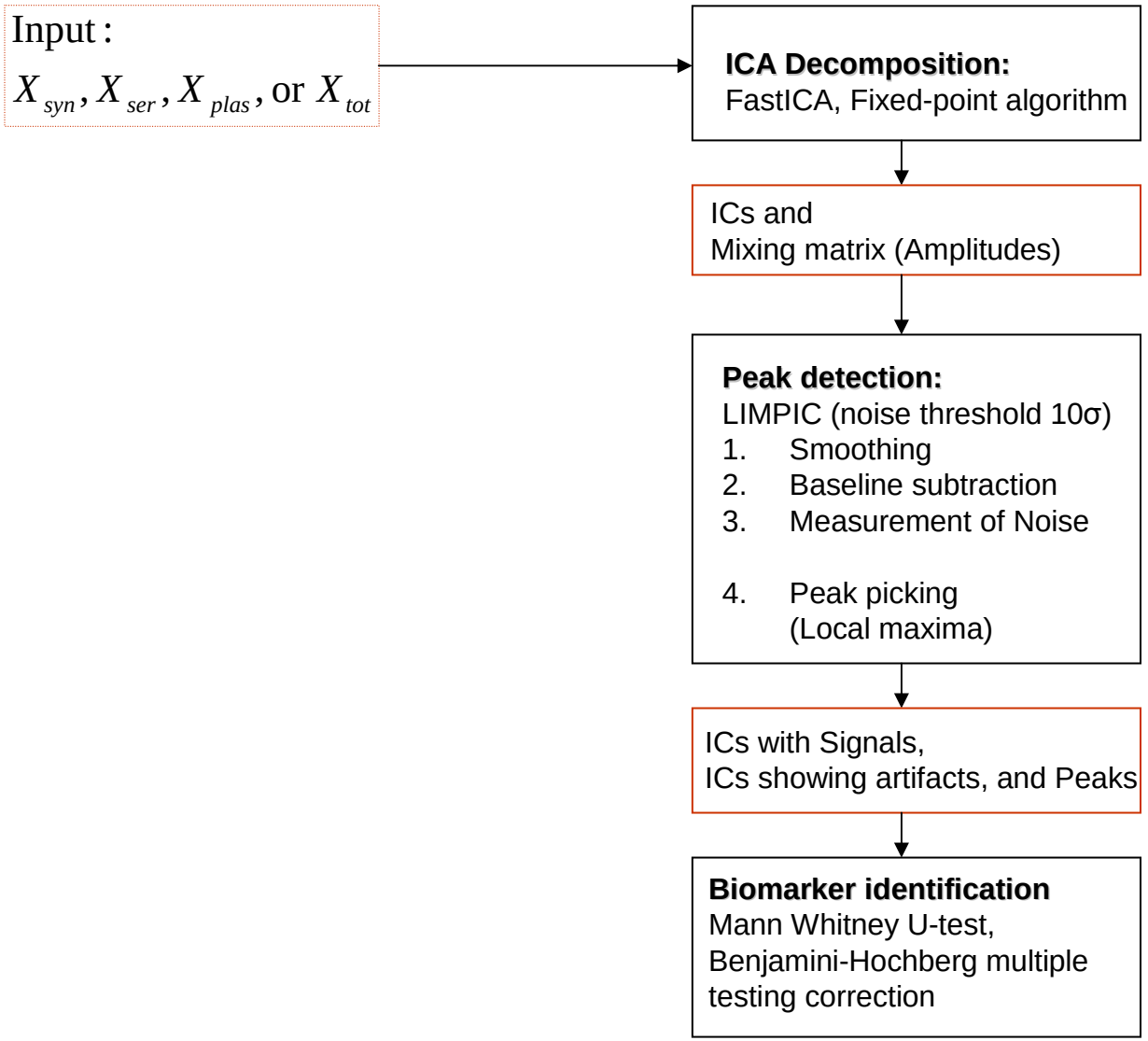
MALDI-TOF MS Experiments



MALDI-TOF MS Data from Experiments



Analysis of Spectra



Peak detection: Characterization of ICs.
1. Smoothing

- Signal enhancement
- Reduction of chemical and electronic noise
- Smoothing performed using Kaiser filter with smoothing factor *p* set to cover a range of 5 Da.

Peak detection: Characterization of ICs. 2. Baseline subtraction

- Baseline drift *c* locally estimated from signal blocks having width of 150 Da
  - For each of them, average intensity (a.i.) was calculated so that a vector *w* of amplitude values was generated
  - *w* was associated to the vector *b* of *m/z* values corresponding to the central point of each interval
  - Components of *w* with rapid intensity variations were considered to be out of the baseline. They were discarded
  - Baseline drift calculated from the remaining (b_i, w_i) by linear interpolation. Then removed from the spectrum

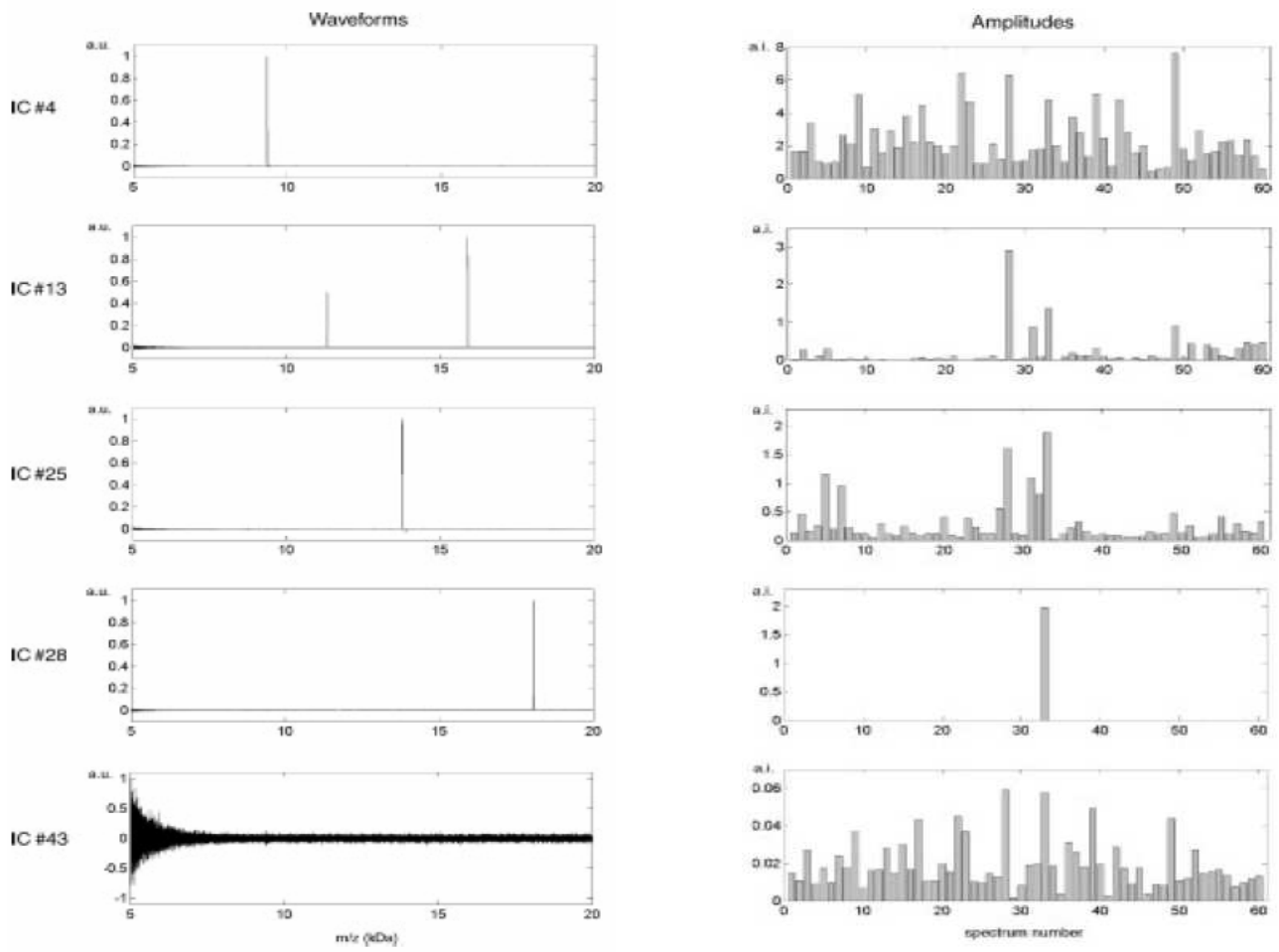
Peak detection: Characterization of ICs. 3. Removal of residual noise

- Residual noise level σ
  - Calculate SD of the values included in the blocks (width 150 Da). Call them g_k
  - Now calculate σ by polynomial interpolation of the points (bk,gk)

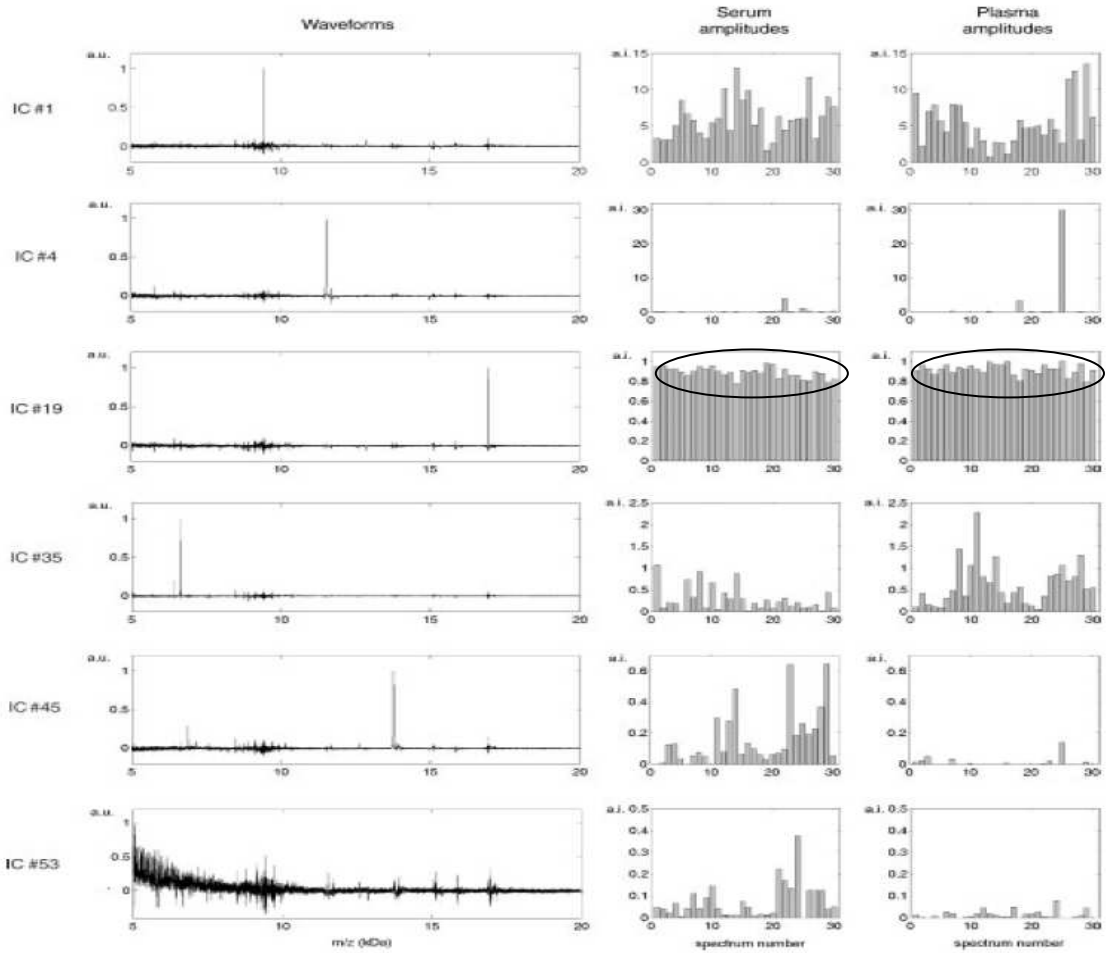
Peak detection: Characterization of ICs
4. Peak picking

- Local maxima: point of highest intensity among the ±*f* nearest points is the peak in that neighbourhood
- *f* = 2. Covers a range of 0.5 Da
- Peaks with intensity lower than 10σ are eliminated from the peak list

IC waveforms of simulated data



IC waveforms of experiment data



IC #1:
Component with the largest power

IC #4:
Signal. Outlier peak

IC #19:
Myoglobin protein
(16952.25 Da)

IC #35, #45:
- Double peak components
- Differentially expressed between Plasma and serum ($P < 0.05$)

IC #53:
- Biological artifact (no peak above the noise level detected)
- Amplitudes significantly different ($P < 0.001$)

Biomarker identification

IC label	m/z (Da)	Serum intensity	Plasma intensity	P
IC #13	9139	0.503 ± 0.370	0.863 ± 713	0.048
IC #17	9715	0.986 ± 0.695	0.534 ± 0.658	0.042
IC #20	6434, 6633	0.566 ± 0.413	1.298 ± 1.038	0.047
IC #22	8917	0.634 ± 1.029	0.225 ± 0.338	0.008
IC #23	9127	0.331 ± 0.271	0.552 ± 0.416	0.024
IC #25	9629	0.367 ± 0.222	0.231 ± 0.202	0.044
IC #30	6439, 6636	0.246 ± 0.218	0.695 ± 0.621	0.007
IC #35	6430, 6629	0.281 ± 0.287	0.614 ± 0.494	0.029
IC #42	6451, 6648	0.018 ± 0.027	0.152 ± 0.247	0.038
IC #45	6881, 13762	0.158 ± 0.177	0.004 ± 0.024	<0.001
IC #51	6941, 13882	0.101 ± 0.086	0.023 ± 0.047	<0.001
IC #59	5601, 5757	0.078 ± 0.107	0.035 ± 0.046	0.049
IC #60	5069	0.092 ± 0.073	0.043 ± 0.031	0.002

Performance comparison of peak identification algorithms

Hit-rate: *Ratio between number of peaks using multi-subject data and the average number of peaks detected in the single spectra*
Hit-rate = 1 means no false positives

		ICA + LIMPIC	LIMPIC	APEX	CENTROID
Serum	Peaks	52	67	93	84
	hit-rate	1	0.42	0.32	0.30
Plasma	Peaks	49	84	121	113
	hit-rate	1	0.40	0.30	0.25
Serum and plasma	Peaks	89	88	143	128
	hit-rate	1	0.47	0.41	0.35

Summary

- MALDI-TOF Mass spectra are contaminated by biological and physical artifacts
- ICA extracted protein signals from calibrated and normalized spectra
- Background noise and outlier peaks could be identified
- Real protein signals showed same peaks contained in mass spectra with increased signal-to-noise ratio
- Can be integrated with existing peak detection methods to enhance their effectiveness
- ICA does not need any parameter tuning for separating protein peaks from noise

Comments

- Optimal number of independent signals is unknown
  - ≤ number of mass spectra according to typical ICA model
  - Prior dimensionality reduction can perhaps help
- False positives (hit-rate: does that make sense?)
  - They indeed *assume* the absence of false positives… and the paper states it!!!
  - Why not directly count them for synthetic data?
- Biomarkers: *as such* are they meaningful here?

References

- Mantini D, Petrucci F, Boccio PD, Pieragostino D, Nicola MD, Lugaresi A, Federici G, Sacchetta P, Ilio CD, Urbani A (2007) Independent component analysis for the extraction of reliable protein signal profiles from MALDI-TOF mass spectra. *Bioinformatics***:** btm533.

- Mantini D, Petrucci F, Pieragostino D, Del Boccio P, Di Nicola M, Di Ilio C, Federici G, Sacchetta P, Comani S, Urbani A (2007) LIMPIC: a computational method for the separation of protein MALDI-TOF-MS signals from noise. *BMC Bioinformatics* **8:** 101.