Fast dictionary-based approach for mass spectrometry data analysis

Afef Cherni[1,3], Émilie Chouzenoux[1,2], Marc-André Delsuc [3]

[1] Université Paris-Est Marne-la-Vallée, LIGM, UMR CNRS 8049, Champs-sur-Marne, France.
[2] Centre pour la Vision Numérique, CentraleSupélec, INRIA Saclay, Gif-sur-Yvette, France.
[3] Université de Strasbourg, IGBMC, INSERM U596, UMR CNRS 7104, Illkirch-Graffenstaden, France.

15-20th April 2018 - Calgary, Alberta, Canada

Joint work with

E. CHOUZENOUX

M-A. DELSUC

Motivation

Mass Spectrometry (1/2)

- Mass Spectrometry :

• A technique used to measure the characteristics, the chemical composition and the structure of a sample or molecule.

• 1919, **Joseph John Thomson**.

- Utility area :

• Pharmaceutical : drug discovery, combinatorial chemistry, pharmacokinetics, drug metabolism.

• Clinical : neonatal screening, haemoglobin analysis, drug testing.

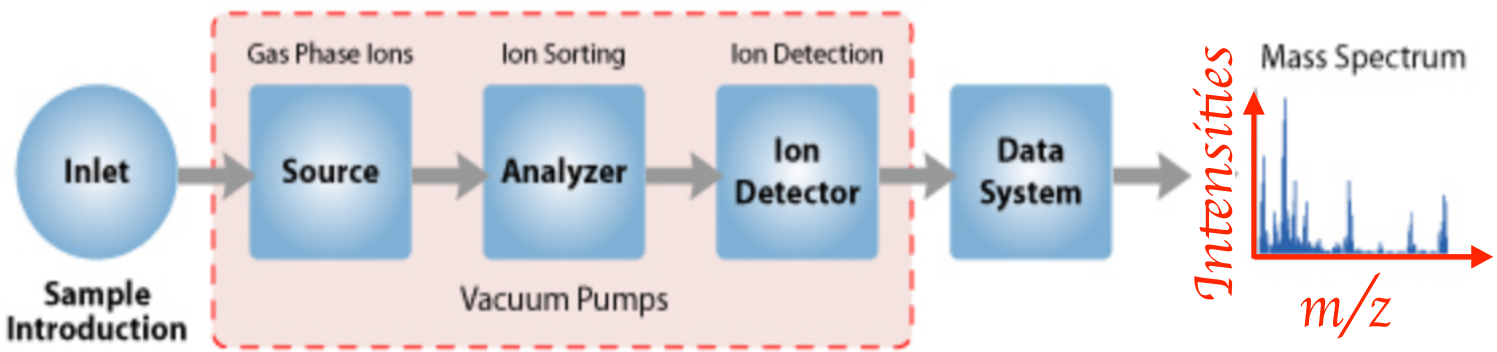
• Environmental : water quality, food contamination.

• Geological : oil composition.

• Biotechnology : analysis of proteins, peptides.

Mass Spectrometry (2/2)

- Principle of measure :



- Aim :

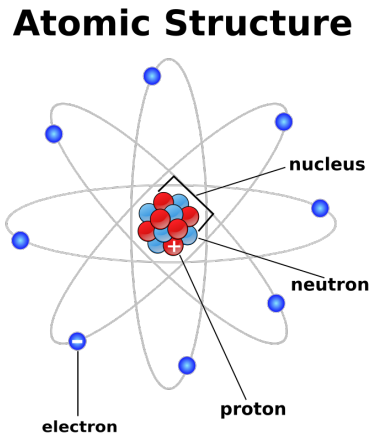
Analyse efficiently big MS data.

Outline

- Problem statement
- Observation model
  - Mathematical model
  - Dictionary-based strategy
  - Ill-posed problem
- Optimization strategy
  - Variational formulation
  - Primal dual algorithm
- Practical Implementation
  - Dictionary construction
  - Circulant approximation
- Application
  - Synthetic results
  - Experimental results
  - MS spectrum analysis

Problem statement

Some reminders on chemistry



- Atom ?
- Atomic number = Number of protons.
- Mass number = Number of protons + neutrons.

- Molecule ?

Set of atoms linked together.

- Isotopic pattern ?

An atom can be present under different forms with different numbers of neutron, called **isotopes**.
Each stable isotope is present in the nature with a specific **abundance**.

Protein class (1/2)

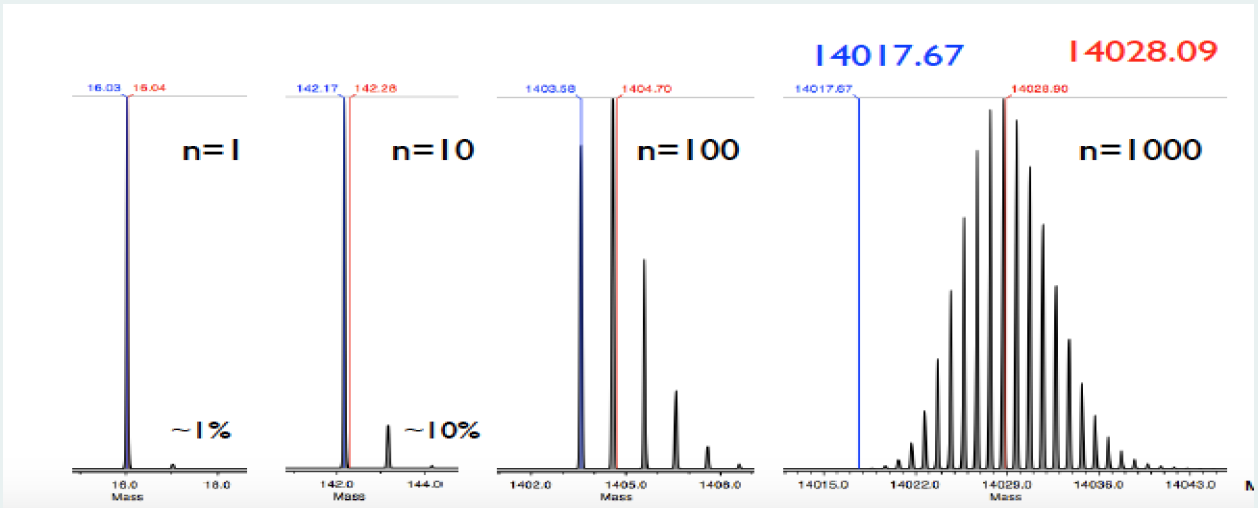
- Protein formula ($C_{N_C}H_{N_H}O_{N_O}N_{N_N}S_{N_S}$)

Atom Name	Atom Symbol	Mass (in Dalton)	Relative Abundance
Carbon	^{12}C	12 *(by definition)*	0.9893
	^{13}C	13.0033548378	0.0107
Hydrogen	^{1}H	1.00782503207	0.999885
	^{2}H	2.0141017778	0.000115
Oxygen	^{16}O	15.99491461956	0.99757
	^{17}O	16.99913170	0.00038
	^{18}O	17.9991610	0.00205
Nitrogen	^{14}N	14.0030740048	0.99636
	^{15}N	15.0001088982	0.00364
Sulfur	^{32}S	31.97207100	0.9499
	^{33}S	32.97145876	0.0075
	^{34}S	33.96786690	0.0425

Isotopic mass and natural abundance of atoms found in proteins (by definition)

Protein class (2/2)

- Example of Alkane molecule C_nH_{2n+2} $(n \in \mathbb{N}^*)$:



$\rightsquigarrow$ The larger the molecule, the larger is the number of different isotopes present, associated with specific probabilities of appearance.

How to find the position of the mono-isotopic peak from a large pattern distribution ?

Observation model

Mathematical model

For a given chemical sample, composed of P different molecules with monoisotopic mass $m_p^{\text{iso}} \in (0, +\infty)$, charge state $z_p \in \mathbb{N}^*$ and abundance $a_p \in (0, +\infty)$, for $p \in \{1, \dots, P\}$:

$$y = \sum_{p=1}^{P} a_p D(m_p^{\text{iso}}, z_p) + n \tag{1}$$

- y : acquired MS spectrum.
- $D(m_p^{\text{iso}}, z_p)$: mass distribution associated to monoisotopic mass m_p^{iso} and charge state z_p.
- n : acquisition noise.

- Discrete measurements on a grid with size M :

$$\mathbf{y} = \sum_{p=1}^{P} a_p \mathbf{d}(m_p^{\text{iso}}, z_p) + \mathbf{n} \tag{2}$$

with $\mathbf{y} \in \mathbb{R}^M$, $\mathbf{d}(m_p^{\text{iso}}, z_p) \in [0, +\infty[^M$ and $\mathbf{n} \in \mathbb{R}^M$.

Dictionary-based strategy

Protein molecule : Averagine model [Senko et al., 1994]

$$\begin{array}{rccc} \mathcal{A}: & \mathbb{R}_+ \times \mathbb{N}^* & \rightarrow & \mathbb{R}_+ \\ & (m,z) & \rightarrow & \mathbf{d}(m,z) \end{array} \tag{3}$$

For M candidate values of isotopic masses and Z candidate charge values, we define a grid with size $T = MZ$, and we define the dictionary $\mathbf{D} \in \mathbb{R}^{M \times T}$:

$$\mathbf{D} = [\mathbf{D}_{1 \leq j \leq M,1}, \mathbf{D}_{1 \leq j \leq M,2}, ..., \mathbf{D}_{1 \leq j \leq M,T}] \tag{4}$$

- i-th column of $\mathbf{D}$ determined by the averagine model at i-th grid position :

$$(\mathbf{D})_{1 \leq j \leq M,i} = \mathbf{d}(m_i, z_i)$$

Inverse problem

Problem

$$\mathbf{y} = \mathbf{D}\overline{\mathbf{x}} + \mathbf{n}' \tag{5}$$

where $\mathbf{y} \in \mathbb{R}^M$, $\mathbf{D} \in \mathbb{R}^{M \times T}$, $\overline{\mathbf{x}} \in \mathbb{R}_+^T$ and $\mathbf{n}' \in \mathbb{R}^M$.

- ✗ y : measure.
- ✗ $\mathbf{D}$: large scale ill-conditioned matrix.
- ✗ $T = MZ$: very large size.
- ✗ $\bar{x}$: unknown sparse vector with positive entries.
- ✗ $\mathbf{n}'$: high noise level.

$\Rightarrow$ **Ill-posed inverse problem in high dimensions.**

Aim

$$\text{Find } \hat{\mathbf{x}} \text{ from } \mathbf{y} \text{ and } \mathbf{D} \text{ such that } \hat{\mathbf{x}} \simeq \overline{\mathbf{x}} \tag{6}$$

Optimization strategy

Variational formulation

$$\underset{\mathbf{x}\in\mathbb{R}^T}{\text{minimize}}\ \Phi(\mathbf{x}) \quad \text{subject to} \quad \|\mathbf{D}\mathbf{x}-\mathbf{y}\| \leq \eta \tag{7}$$

- $\Phi : \mathbb{R}^T \mapsto]-\infty, +\infty]$ is a proper, lower semicontinuous (lsc), **convex** regularization function used to enforce **positivity** and **sparsity** on the solution.
- $\eta > 0$ is a parameter that depends on the noise characteristics.

Proximity operator

Let $\Phi : \mathbb{R} \to]-\infty, +\infty]$ a lsc proper and convex function. The proximity operator of Φ is defined as *[Moreau, 1965]* `http://proximity-operator.net/` [Cherni et al., 2017] :

$$\text{prox}_{\Phi}(x) : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$$

$$x \to \underset{y\in\mathbb{R}^{\mathbb{N}}}{\text{argmin}} \left(\Phi(y) + \frac{1}{2}\|y-x\|^2 \right)$$

Primal Dual algorithm [Chambolle and Pock, 2011]

Initialisation

$\mathbf{u}^{(0)} \in \mathbb{R}^M$, $\mathbf{x}^{(0)} \in \mathbb{R}^T$

$0 < \sigma < \|\mathbf{D}\|^2/\tau$

$\rho \in (0,2)$, $\tau > 0$

Minimisation

For $k = 0, 1, \ldots$

$$
\begin{aligned}
\tilde{\mathbf{x}}^{(k)} &= \text{prox}_{\tau\Phi}(\mathbf{x}^{(k-1)} - \tau\mathbf{D}^\top(\mathbf{u}^{(k-1)})) \\
\mathbf{v}^{(k)} &= \mathbf{u}^{(k-1)} + \sigma\mathbf{D}(2\tilde{\mathbf{x}}^{(k)} - \mathbf{x}^{(k-1)}) \\
\tilde{\mathbf{u}}^{(k)} &= \mathbf{v}^{(k)} - \sigma\text{proj}_{\|\cdot-\mathbf{y}\|\leq\eta}(\mathbf{v}^{(k)}/\sigma) \\
\mathbf{x}^{(k)} &= \mathbf{x}^{(k-1)} + \rho(\tilde{\mathbf{x}}^{(k)} - \mathbf{x}^{(k-1)}) \\
\mathbf{u}^{(k)} &= \mathbf{u}^{(k-1)} + \rho(\tilde{\mathbf{u}}^{(k)} - \mathbf{u}^{(k-1)})
\end{aligned}
$$

For every $(\mathbf{y}, \mathbf{v}) \in (\mathbb{R}^N)^2$:

$$\text{proj}_{\|\cdot-\mathbf{y}\|\leq\eta}(\mathbf{v}) = \mathbf{v} + (\mathbf{v} - \mathbf{y}) \min\left(\frac{\eta}{\|\mathbf{v} - \mathbf{y}\|}, 1\right) - \mathbf{y}. \tag{8}$$

- The convergence of the iterates $(\mathbf{x}^{(k)})_{(k\in N)}$ is ensured [Chambolle and Pock, 2011].

Practical Implementation

Dictionary construction

Given a range of masses $[m_{\min}, m_{\max}]$ and charges $[z_{\min}, z_{\max}]$, we define a regular grid :

$$(\forall i \in \{1, \ldots, T\}) \quad m_i = m_{\min} + (j-1)m_{\max}, \tag{9}$$

$$z_i = z_{\min} + (\ell - 1)z_{\max}, \tag{10}$$

with the convention $i = \ell M + j$, $j \in \{1, \ldots, M\}$ and $\ell \in \{1, \ldots, Z\}$.

- The i-th column of $\mathbf{D}$ is taken as $\mathbf{d}(m_i, z_i)$.
- $\mathbf{d}(m_i, z_i)$ corresponds to a sampled version of $D(m, z)$ on the mass grid with size M.

Circulant approximation

- Difficulty

✗ Very large value for MZ.
$\Rightarrow$ Large dictionary $\mathbf{D}$ which presents a computational challenge and large memory resources.

- Assumptions

• Similar isotopic mass patterns mainly differ by a sample shift of peaks positions.
• Isotopic patterns are sparse with positives entries.

Proposal

$\rightsquigarrow$ Decompose the mass axis into windows with width $L \leq M$.

Circulant approximation

- **Initial dictionary**

$$\mathbf{D} = \left[\mathbf{D}_{1\leq j\leq M,1}, \mathbf{D}_{1\leq j\leq M,2}, \ldots, \mathbf{D}_{1\leq j\leq M,T}\right] \quad T = MZ \tag{11}$$

$$\mathbf{D} = \left[\mathbf{D}_1 | \ldots | \mathbf{D}_\ell | \ldots | \mathbf{D}_Z\right]_{1\leq \ell\leq Z} \tag{12}$$

- **Circulant model**

Each $\mathbf{D}_\ell$ is approximated by the following block diagonal (BDiag) matrix made of M/L blocks assumed to be circulant (Circ) matrices with first line $\overline{\mathbf{d}}_{s,\ell}$, $s \in \{1, \ldots, M/L\}$:

$$\overline{\mathbf{D}}_\ell = \mathrm{BDiag}\left(\left[\mathrm{Circ}\left(\overline{\mathbf{d}}_{s,\ell}\right)\right]_{1\leq s\leq M/L}\right). \tag{13}$$

✓ Fast computation of products ($\mathbf{D}$, $\mathbf{D}^T$) using fast Fourier tools.

✓ High reduction of the memory requirements.

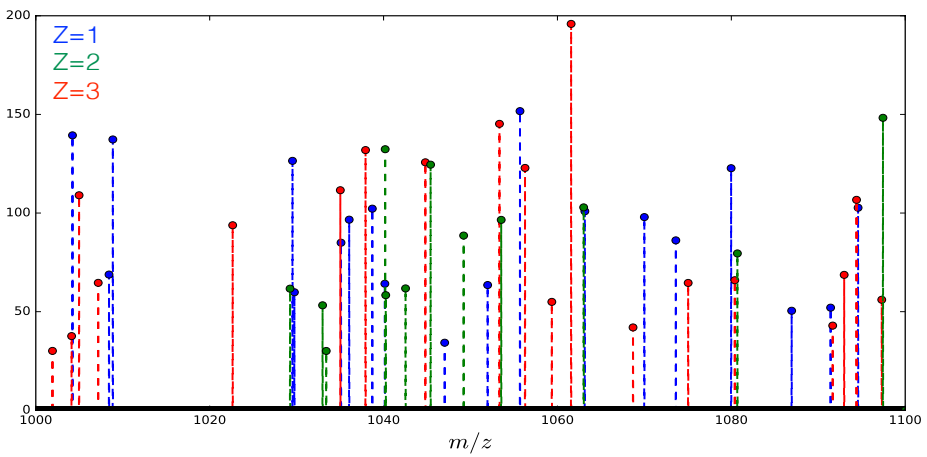
$$\overline{\mathbf{D}} = \left[\overline{\mathbf{D}}_1 | \ldots | \overline{\mathbf{D}}_\ell | \ldots | \overline{\mathbf{D}}_Z\right]_{1\leq \ell\leq Z} \tag{14}$$

Application

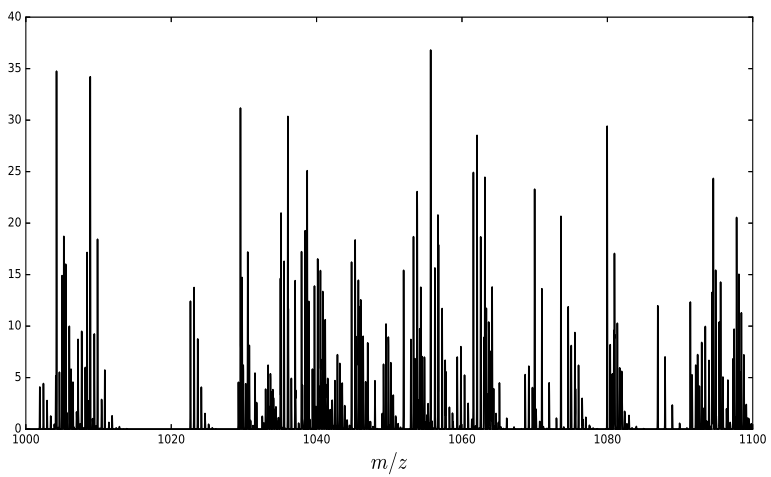
Synthetic results (1/4)

Simulated data

- Signal A :
  - M = 5000
  - Z = 3, $z_{\text{min}} = 1$, $z_{\text{max}} = 3$
  - P = 50 proteins
  - Mass axis = $[1000, 1100]$



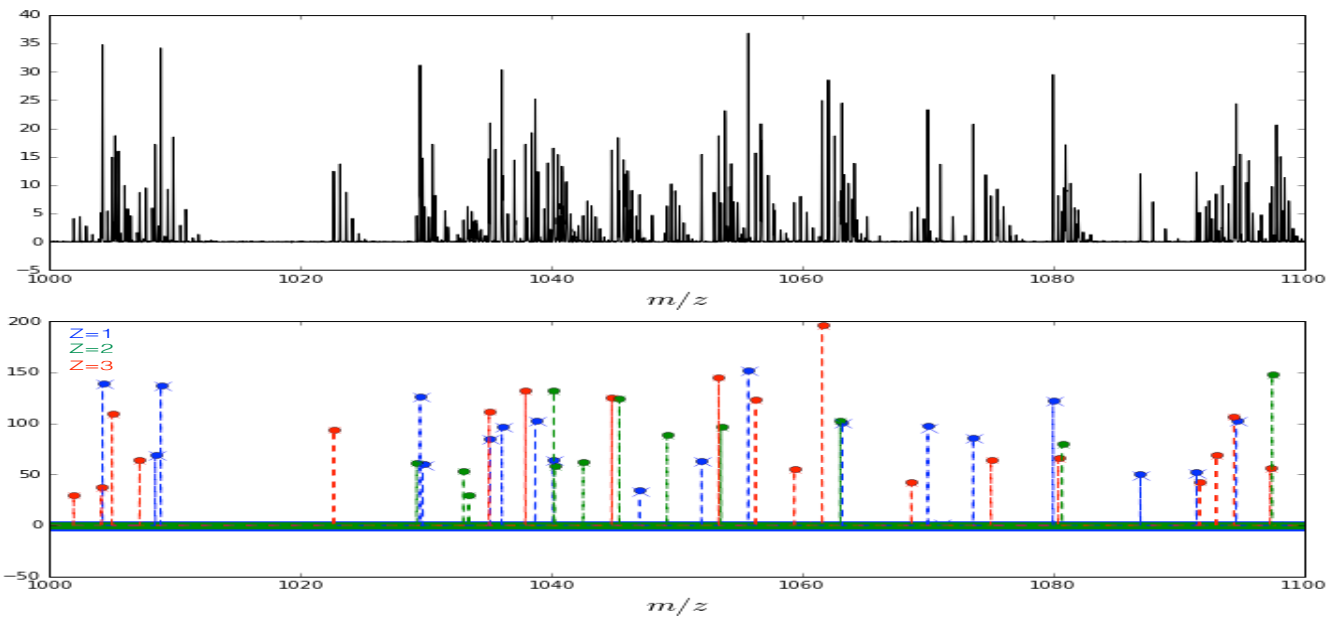
Original signal



Measured data

Synthetic results (2/4)

- Parameters dataset : Gaussian noise, iid with std $\sigma = 10^{-2}$, $\tau = \theta\sigma\sqrt{M}$, $\theta \simeq 1$.
- Parameters algorithm : $\rho = 1.9$, $\tau = \|\mathbf{D}\|^{-1}$, $\tau = 0.9\sigma$, maximum iterations number=1000.



Reconstruction results of the signal A : (top) input data y, (bottom) exact spectrum (dots), restored spectrum with exact dictionary (dashed line), and with its block-circulant approximation for L = 10 (asterisks).

Synthetic results (3/4)

- Impact of noise ? Memory storage ?

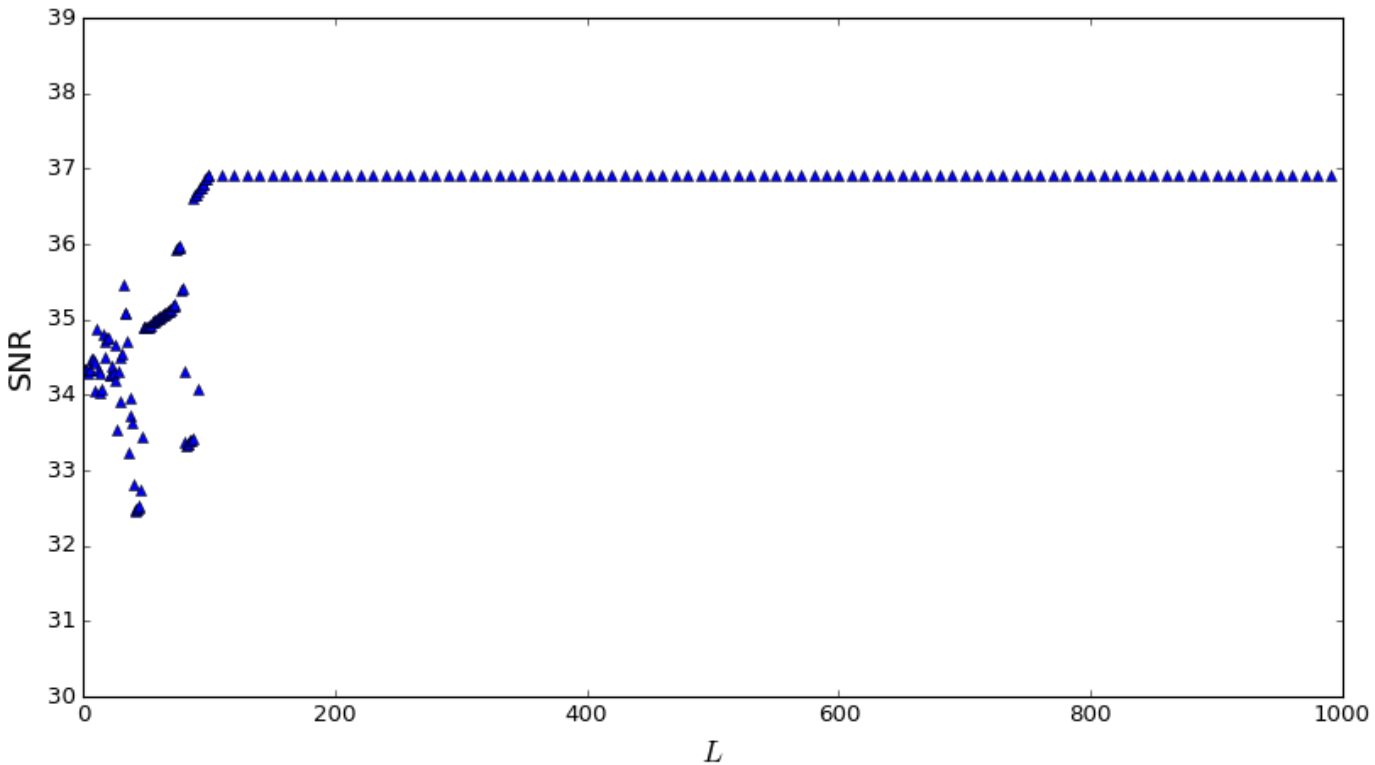
σ	Exact dictionary approach			Block-circulant approximation		
	SNR	Time	Memory storage	SNR	Time	Memory
1	16.18	303.33	572	15.57	127.85	0.53
0.1	35.73	206.84		35.43	44.48	
0.01	39.56	377.80		38.38	290.56	

SNR (in dB), computation time (in s) and memory storage (in MB) for the restoration of signal A for various values of noise level. Block-circulant approximation $\overline{\mathbf{D}}$ is tested for $L = 10$.

- Good quality reconstruction even with high noise level.
- Block-circulant approximation is faster than exact dictionary approach with limited deterioration of the results quality.

Synthetic results (4/4)

- Influence of L ?



SNR of the restored signal A with $\sigma = 10^{-2}$ using $\overline{\mathbf{D}}$ for various L values

- Reconstruction quality stable to the value of L.

Experimental results

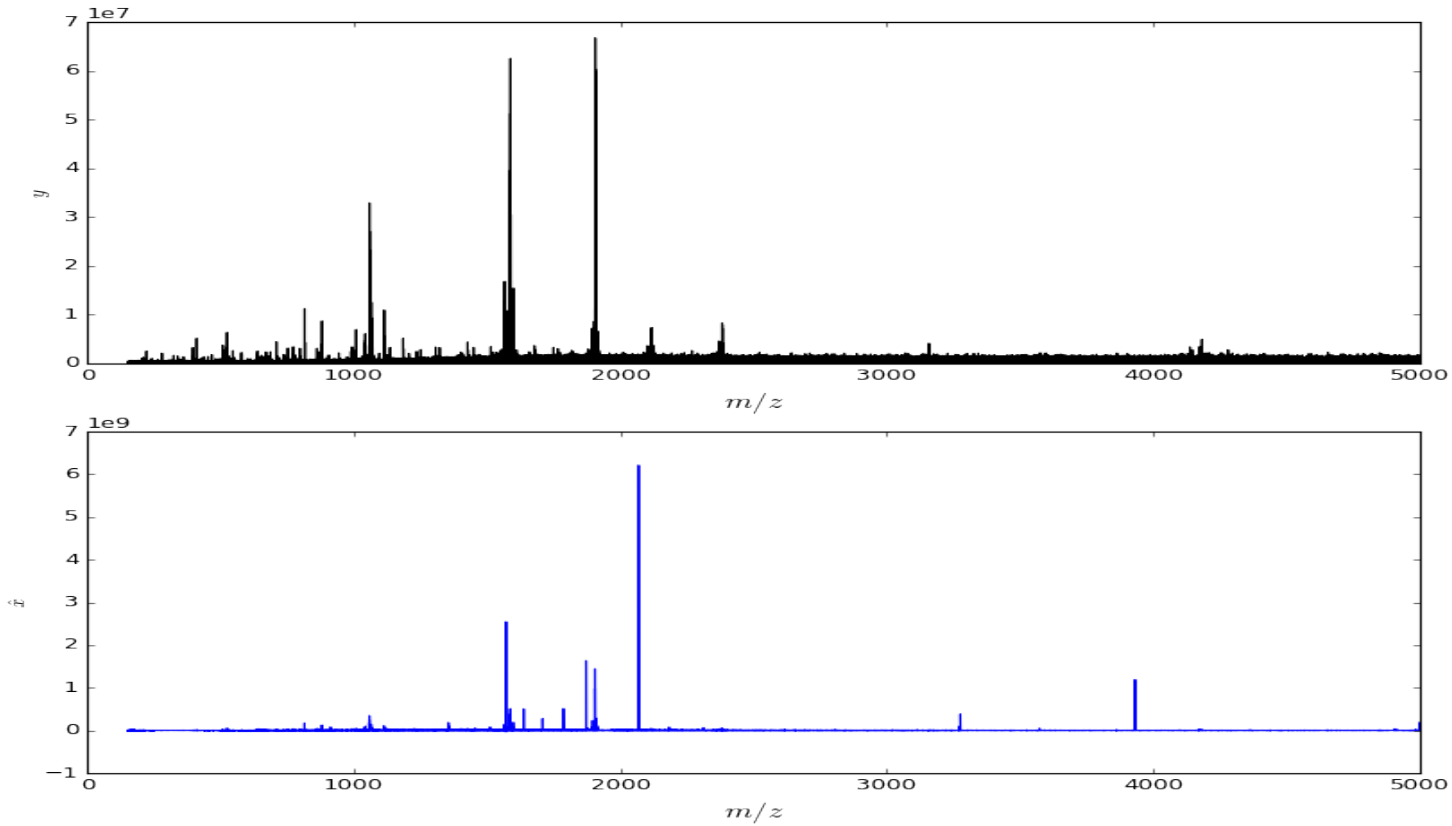
Parameters dataset

- peptide `EVEALEKKVAALESKVQALEKKVEALEHG-NH2`
- 3 μM ($C_{140}H_{240}N_{38}O_{45}$).
- Trimer form within 50 mM of NH_4OAc
- BRUKER Solarix 15 T, FT-ICR instrument, ESI source.
- $N = 8000000$

Parameters algorithm

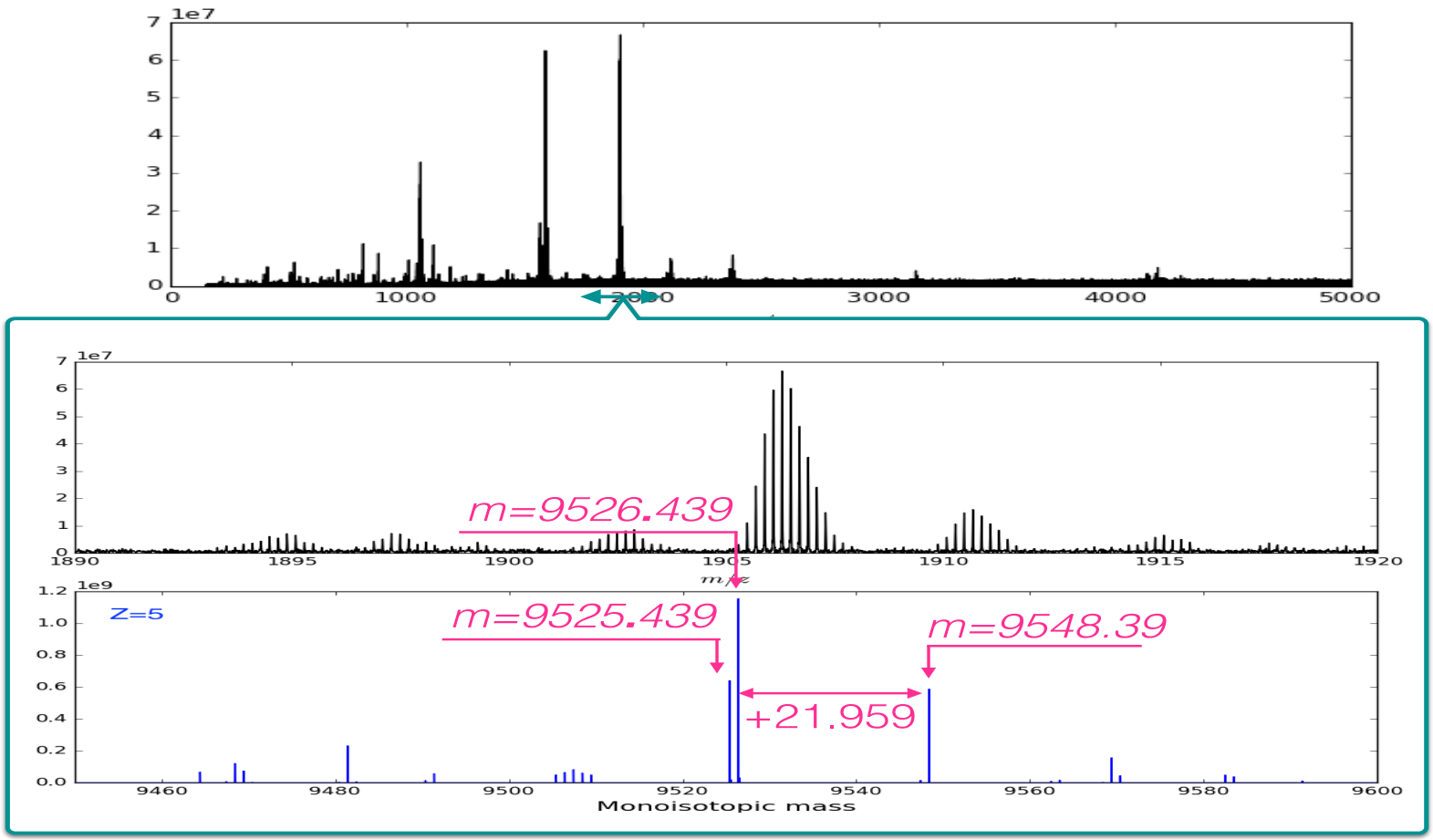
- σ : estimated from an empty frame of the measured signal.
- maximum iterations number : 2000

Experimental results



Analysis of the real FT-ICR MS spectrum of peptide in trimer form : (top) acquired data, (bottom) recovered spectrum $z = 5$ using block-circulant approximated dictionary with $L = 10$

MS spectrum analysis



- Theoretical monoisotopic mass of the peptide : $m = 9526.337$ Daltons.
- Theoretical position of the Sodium : $m = +21.982$ Daltons.

Conclusion & perspectives

Conclusion & perspectives

$\surd$ New dictionary-based approach for MS data analysis based on proteomic averagine model.

$\surd$ Penalized cost function promoting sparsity and positivity, minimized with efficient primal-dual scheme with sought convergence guarantees.

$\surd$ Block-circulant approximation of the dictionary-based approach.

$\Rightarrow$ Efficient analysis of synthetic and real MS spectra.

$\Rightarrow$ Fast approach devoted to the big data scale.

$\Rightarrow$ Limited memory resources required.

Conclusion & perspectives

$\surd$ New dictionary-based approach for MS data analysis based on proteomic averagine model.

$\surd$ Penalized cost function promoting sparsity and positivity, minimized with efficient primal-dual scheme with sought convergence guarantees.

$\surd$ Block-circulant approximation of the dictionary-based approach.

$\Rightarrow$ Efficient analysis of synthetic and real MS spectra.

$\Rightarrow$ Fast approach devoted to the big data scale.

$\Rightarrow$ Limited memory resources required.

- Extend this approach to the processing of multidimensional MS spectra.

References

Chambolle, A. and Pock, T. (2011).

A first-order primal-dual algorithm for convex problems with applications to imaging.
Journal of Mathematical Imaging and Vision, 40(1) :120–145.

Cherni, A., Chouzenoux, E., and Delsuc, M.-A. (2017).

Palma, an improved algorithm for dosy signal processing.
Analyst, 142(5) :772–779.

Moreau, J.-J. (1965).

Proximité et dualité dans un espace hilbertien.
Bulletin de la Société mathématique de France, 93 :273–299.

Senko, M.-W., Speir, J.-P., and McLafferty, F.-W. (1994).

Collisional activation of large multiply charged ions using Fourier transform mass spectrometry.
Analytical Chemistry, 66(18) :2801–2808.

Thank you for your attention !

Fast dictionary-based approach for mass spectrometry data analysis

A. Cherni[1], E. Chouzenoux[2], M.-A. Delsuc [3]

[1] cherniafef@hotmail.fr
[2] emilie.chouzenoux@univ-mlv.fr
[3] madelsuc@unistra.fr

ICASSP 2018 - Calgary, 15-20th April