New dictionary approach for the isotopic deconvolution

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

[1] Université Paris-Est, LIGM (UMR 8049), CNRS, ENPC, ESIEE Paris, UPEM, Marne-la-Vallée, France.
[2] Université de Strasbourg, IGBMC (UMR 7104), CNRS, INSERM U596, Illkirch-Graffenstaden, France.
[3] Centre pour la Vision Numérique, CentraleSupélec, INRIA Saclay, Châtenay-Malabry, France.

Abstract

In the context of proteomic analysis, the superposition of the isotopic patterns of different peptides, in different charge-states can produce MS spectra difficult to decipher, in particular in MS/MS top-down approaches. A robust analysis of such spectra is still an open question in the presence of crowded spectra presenting an important spectral overlap and poor resolution. The problem can be expressed as a convolution, however this is an "ill-posed" problem notoriously difficult to resolve. In addition, the large size of the MS data requires the usage of efficient data processing algorithms, able to handle efficiently the large data sets involved. The MaxEnt approach has long been used to solve this deconvolution problem. We present here a new approach, using a convex optimisation based on proximity operators. It implements the efficient ℓ_1 regularization which converges rapidly. Our method resolves the isotopic deconvolution problem using a dictionary concept via the averagine model. The size of the problem is circumvented by a Fourier Transform based convolution and the use of a Primal-Dual convex algorithm. An application of this method from MS/MS and 2D-FTICR spectra will be shown.

Index Terms: **Mass spectrometry, FTI-CR, Local deconvolution, Fourier Transform, PALMA, Primal Dual, Averagine model, Isotopes, Proteins.**

PROBLEM

☞ Mass Spectrometry is a powerful tool used to detect and quantify molecules in a complex mixture.
☞ Molecule presents a specific mass depending on the sum of its isotopes.
✓ Only the monoisotopic mass assures a non-ambiguous determination of the protein.
✗ With the large size of proteins, the signal of the monoisotopic mass can be vanishingly small.
✗ The direct detection of the monoisotopic mass is not possible.
✗ The characterization of the protein is difficult to decipher.

MEASUREMENT MODEL

For a given protein with P molecules, the signal $\mathbf{y}$ is related to the isotopic masses through the model:

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

$\mathbf{y} \in \mathbb{R}^N$: input signal, $a_p \in [0, +\infty[$: intensities of the isotopic pattern, $D(m_p^{\text{iso}}, z) \in [0, +\infty[^N$: mass distribution of the simplistic protein, m^{iso} isotopic mass, z_p: charge, $\mathbf{b} \in \mathbb{R}^N$: acquisition noise.
☞ **Goal: Reconstruct the set of coefficients $(a_p, m_p^{\text{iso}}, z_p)_{1\leq p\leq P}$ from y throw the following model:**

$$\mathbf{y} = \mathbf{K}\mathbf{x} + \mathbf{b} \qquad (2)$$

where $\mathbf{K} = (\mathbf{K}_i)_{1\leq i\leq N}$ and $\mathbf{K}_i = D(m_p^{\text{iso}}, z_p)$.
✓ Averagine model will be used to determine the isotopic distribution $D(m_p^{\text{iso}}, z)$ for m_p^{iso} and z [1].

CONVEX OPTIMIZATION

Penality approach:

$$\underset{\mathbf{x}\in[0,+\infty[^N}{\text{minimize}} \ \ell_1(\mathbf{x}) \quad \text{subject to} \quad \|\mathbf{K}\mathbf{x} - \mathbf{y}\| \leq \tau \qquad (3)$$

• τ: a parameter depending on the noise characteristics
• ℓ_1: the regularization function chosen according the sparsity of $\mathbf{x}$.

Dictionary-based approach:
We build the dictionary $\mathbf{K}$ where:

$$\mathbf{K} = (\mathbf{K}_i)_{1\leq i\leq N} \quad \text{and} \quad \mathbf{K}_i = D(m_p^{\text{iso}}, z) \qquad (4)$$

Functional approximation:
We propose to approximate the dictionary $\mathbf{K}$ by a function $\mathcal{K} : [0, +\infty[^N \mapsto \mathbb{R}^N$ based on Fourier approximation of the linear function $\mathbf{x} \mapsto \mathbf{K}\mathbf{x}$ such that:

$$\mathcal{K}(x) = \sum_{j=1}^{N/L} \mathcal{F}^{-1}\left(\mathcal{F}(D(\overline{m}_j, z_j))\mathcal{F}(x_j)\right) \qquad (5)$$

• $\mathcal{F}^{-1}$: the inverse Fourier Transform, $\overline{m}_j$: the average mass on the j-th window on the mass axis, z_j: its charge and x_j: the j-th window of $\mathbf{x}$.

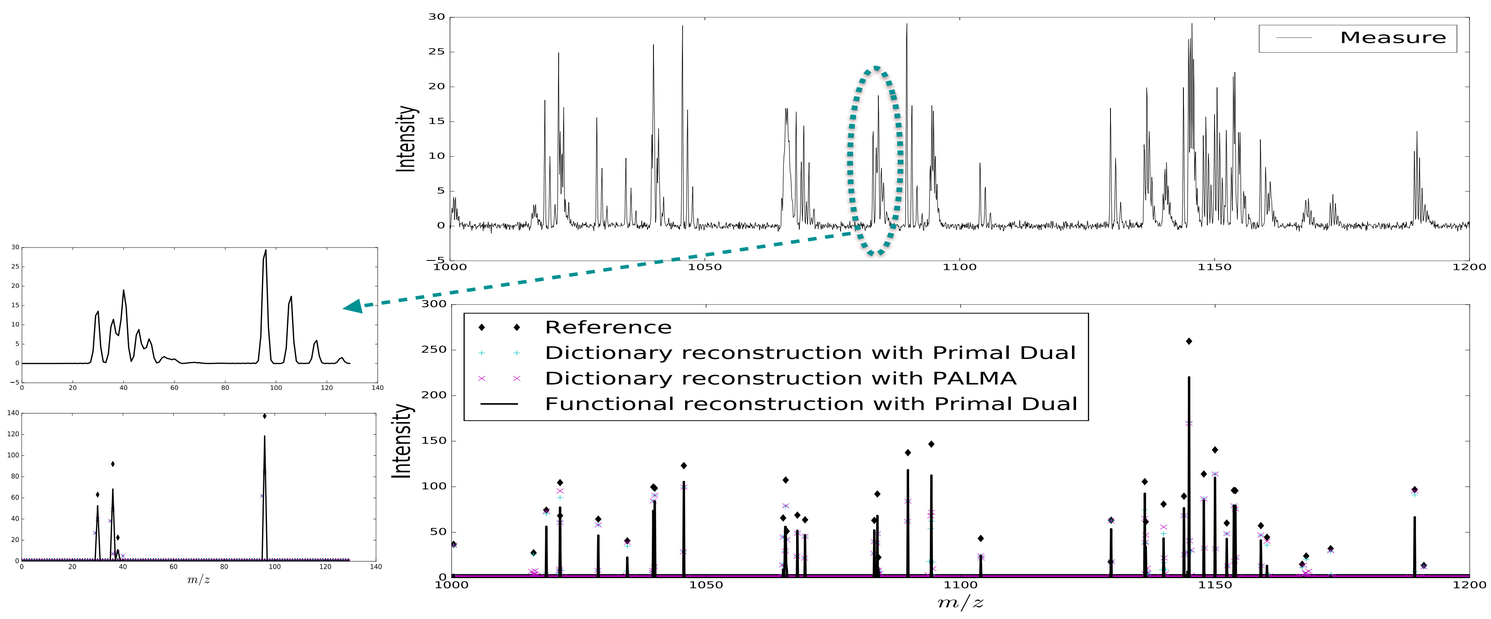
OPTIMIZATION STRATEGY

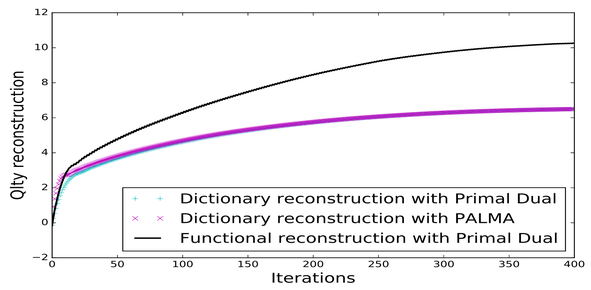
PALMA algorithm[2]:
✓ Good recovery of 1D spectrum.
✗ The treatment of large MS measures (2D and multicharge spectra) is not feasible.

☞ **Primal Dual algorithm**[3]

1D MS SPECTRUM RECOVERY

• **1D Simulated signal**: $N = 2000$, $M_{min} = 1000$, $M_{max} = 1200$, $P = 40$, $Z \in (1, 2, 3)$, noise level= 1%, iterations= 200.

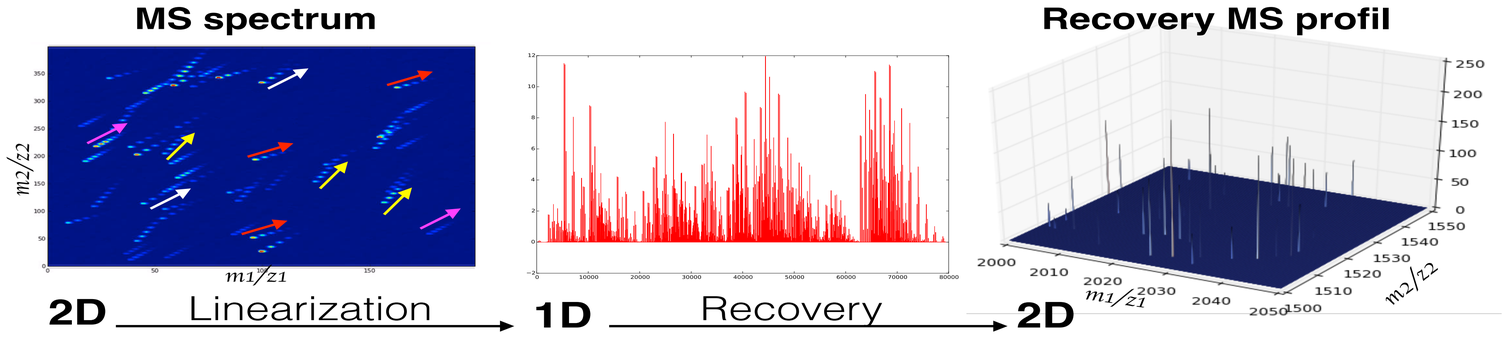




	Computation time(s)	Qlty recover(dB)	Memory Storage(Mb)
Dictionary reconstruction with Primal Dual	15.19	7.73	91
Dictionary reconstruction with PALMA	33.11	7.62	365
Functional reconstruction with Primal Dual	82.26	10.25	✓

Results with iterations number equal to 400.

• **2D Simulated signal**: $N1 = 400$, $N2 = 200$, $M1_{min} = 2000$, $M1_{max} = 2050$, $M2_{min} = 1500$, $M2_{max} = 1550$, $P = 30$, $Z_1 \in (1, 2)$, $Z_2 \in (1, 2)$, noise level= 1%, iterations= 200



CONCLUSION & PERSPECTIVES

✓ 1D & 2D deconvolution of isotopic masses of proteins ☞ Generalization of our approach to process any complex mixture.

REFERENCES

[1] Michael W Senko, J Paul Speir, and Fred W McLafferty, "*Collisional activation of large multiply charged ions using fourier transform mass spectrometry*", Analytical chemistry, vol. 60, no. 18, pp. 2801-2808,1994.

[2] Afef Cherni, Emilie Chouzenoux, and Marc-André Delsuc, "*Palma, an improved algorithm for dosy signal processing*", Analyst, vol. 142, no. 5, pp. 772-779, 2017.

[3] Antonin Chambolle and Thomas Pock, "*A first-order primal-dual algorithm for convex problems with applications to imaging*", Journal of mathematical imaging and vision, vol. 40, no. 1, pp. 120-145, 2011.

THIS WORK WAS SUPPORTED BY THE CNRS MASTODONS PROJECT UNDER GRANT 2016TABASCO, BY THE AGENCE NATIONALE POUR LA RECHERCHE, GRANT ANR2014 ONE SHOT 2D FT ICR