HAL
open science

Fast Dictionnary-Based Approach for Mass Spectrometry Data Analysis

Afef Cherni, Emilie Chouzenoux, Delsuc Marc-André

To cite this version:

Afef Cherni, Emilie Chouzenoux, Delsuc Marc-André. Fast Dictionnary-Based Approach for Mass Spectrometry Data Analysis. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2018), Apr 2018, Calgary, Canada. 10.1109/ICASSP.2018.8461720 . hal-01803419

HAL Id: hal-01803419

https://hal.science/hal-01803419

Submitted on 30 May 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

FAST DICTIONARY-BASED APPROACH FOR MASS SPECTROMETRY DATA ANALYSIS

Cherni Afef [1,3], *Chouzenoux Émilie*[1,2], *Delsuc Marc-André*[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] Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), INSERM U596, UMR CNRS 7104, Université de Strasbourg, Illkirch-Graffenstaden, France.

ABSTRACT

Mass spectrometry (MS) is a fundamental technology of analytical chemistry for measuring the structure of molecules, with many application fields such as clinical biomarker analysis or pharmacokinetics. In the context of proteomic analysis with MS, the superposition of the isotopic patterns of different proteins, in various charge-states produces MS spectra difficult to decipher. The complexity of the pattern models and the large size of the data again increase the difficulty of the analysis step. In this paper, we propose to formulate the problem of proteins characterization as the estimation of a positive-valued sparse signal thanks to a dictionary-based approach relying on the protein averagine concept. A proximal primal-dual splitting convex optimization method is considered to solve the resulting variational problem. Moreover, the large size of the dictionary matrix is circumvented by proposing a suitable block circulant approximation of it, allowing to limit the computational burden of the method. Numerical experiments on synthetic and real MS datasets illustrate the good performance of our approach.

***Index Terms*—** Mass spectrometry, Proteomic Analysis, Sparsity, Dictionary-based strategy, Primal-Dual algorithm.

1. INTRODUCTION

Mass Spectrometry (MS) is a powerful tool used for robust, accurate, and sensitive detection and quantification of molecules of interest. Thanks to its sensibility and selectivity, MS is widely used in proteomics such anti-doping, metabolomics, medicine or structural biology [1, 2]. In particular, it has applications in clinical research [3], personalized medicine [2], diagnosis process and tumours profiling [4] and pharmaceutical quality control [5]. In a MS experiment, the raw signal arising from the molecule ionization in an ion beam is measured as a function of time via Fourier Transform-based measures such as Ion Cyclotron Resonance (FT-ICR) and Orbitrap. A spectral analysis step is then performed, possibly involving a series of operations/algorithms [6, 7] to improve the quality of data, transforming the time-domain data into the frequency domain. The frequency spectrum is then converted to the so-called MS spectrum through a calibration function. This spectrum presents a set of positive-valued peaks distributed according to the charge state and the isotopic distribution of the studied molecule, generating several typical patterns in the signal. The goal is then to determine from this observed pattern distribution the most probable chemical composition of the sample, through the determination of the monoisotopic mass, charge state and abundance of each present molecule. Unfortunately, the superposition of the isotopic patterns in different charge-states can produce MS spectra difficult to decipher, and the complexity of the problem again increases with the number and the size of molecules. Proteins, the family of molecules that will be targeted in this work, present a particular challenge since they are rather large molecules with wide isotopic patterns, ionized by ElectroSpray Ionization (E.S.I) leading to a mixture of several charge states. Additionally, the usual high resolution and thus very large size of the measurements in this context (usually > 500k data point) make their analysis cumbersome. Peak-peaking and pattern recognition approaches are the most common methods currently used for MS spectrum analysis in the context of proteins samples [8, 9, 10]. Although these methods may be quite fast, they suffer from slow performance, instability and sensitivity to high noise level since they all require a preprocessing step to threshold the data. In particular, their performance can be highly degraded when a strong peak overlap masks the position of other peaks, which usually happens when several distinct proteins and/or several charge states are in presence [9].

In this paper, we propose a new dictionary-based approach to solve efficiently and automatically this problem. We start by introducing in Section 2 the chemical problem statement, the measurement model and our dictionary-based strategy. In Section 3, we formulate the selection of the dictionary elements as the resolution of a convex non-smooth optimization problem, and describe a primal-dual proximal algorithm to solve it efficiently. In Section 4, we discuss the practical implementation of our method, and propose a block-circulant approximation of the dictionary matrix to reduce the computational cost of the processing. Finally, an illustration of the good performance of our method for recovering synthetic and real data in the context of Mass Spectrometry is presented in Section 5.

2. PROBLEM STATEMENT

2.1. Some reminders on chemistry

An atom is the basic unit of matter and the smallest defining structure of elements, defined in the periodic table with a symbol and a nucleon number. 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. Each unique molecule presents thus a specific mass in Daltons, depending on the sum of the masses of each of its constituting isotopes and on its charge state z.

This work was supported by the CNRS MASTODONS project under grant 2016TABASCO and by the Agence Nationale pour la Recherche (ANR, France) under grant 2010FT-ICR2D and grant Défi de tous savoirs 2014, ONE-SHOT-FT-ICR-MS-2D).

When a large number of samples with various charge states is considered, for instance when measuring a MS spectrum, a distribution of peaks, named multi-charged isotopic pattern, is observed on the mass over charge (i.e. (m/z)) axis, following the composition of the elementary distributions of all atoms [11, 12] and their charge distribution [13]. For a fixed charge state $z > 0$, the smallest (m/z) position of the peaks is associated with the most abundant isotope mass and allows to determine the monoisotopic mass at charge z. This quantity is independent on the relative isotopic abundances, and helps in a non-ambiguous determination of the molecule. However, for large molecules, the probability of having a single charge state and no isotopes is extremely low so that the peak intensity at the monoisotopic masses can be vanishingly small and their direct detection impossible.

The purpose of this work is to provide an automatic tool to characterize monoisotopic mass and charge state quantities from the measured MS data, in the context of proteomic analysis, i.e. when the chemical sample to be studied is made of several proteins. We recall that a protein is a large molecule having the generic formula $\mathrm{C}_{N_\mathrm{C}}\mathrm{H}_{N_\mathrm{H}}\mathrm{O}_{N_\mathrm{O}}\mathrm{N}_{N_\mathrm{N}}\mathrm{S}_{N_\mathrm{S}}$ where $(N_C, N_H, N_O, N_N, N_S)$ are respectively the number of Carbon C, Hydrogen H, Oxygen O, Nitrogen N and Sulfur S. The most probable isotopes of the latter atoms, along with their mass at the neutral state and their associated abundances, are given in Table 1. Note that for $z > 0$, as it is the case in MS data, the mass values are all shifted proportionally to z and to the mass of the hydrogen adduct ions [13].

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

Table 1: Isotopic mass and natural abundance of atoms found in proteins [14, 15].

2.2. Observation model

Let us consider a given chemical sample, composed of P different proteins 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 $p \in \{1, \ldots, P\}$. The acquired MS spectrum y can be modeled as the weighted sum of each individual isotopic pattern $y = \sum_{p=1}^{P} a_p D(m_p^{\text{iso}}, z_p) + n$ where n models the acquisition noise and possible errors arising from the spectral analysis preprocessing step. The measurements are taken on a discrete grid of (m/z) values with size M, so that the observation model finally reads:

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

with $\mathbf{y} \in \mathbb{R}^M$, $\mathbf{d}(m_p^{\text{iso}}, z_p) \in [0, +\infty[^M$ and $\mathbf{n} \in \mathbb{R}^M$. The aim of the MS spectrum analysis is then to reconstruct the set of coefficients $(a_p, m_p^{\text{iso}}, z_p)_{1 \leq p \leq P}$ from $\mathbf{y}$. The lack of knowledge of the complicated and nonlinear function $\mathbf{d}(m, z)$ and the large value of M present the main locks of this problem.

2.3. Proposed dictionary-based strategy

The mass distribution function $D(m, z)$, at a given (m, z) is actually easy to evaluate from the molecular formula using a recursive program [11]. We thus propose to adopt a dictionary-based approach for solving the estimation problem, under the assumption that we know approximately the range of mass and charge for the P proteins present in the sample. To do so, let us define a search grid, with size $T := MZ$ which defines M possible values of isotopic masses and Z possible values for the charges. From this grid, we build the dictionary $\mathbf{D} \in \mathbb{R}^{M \times T}$ so that each column $i \in \{1, \ldots, T\}$ of $\mathbf{D}$ is $\mathbf{d}(m_i, z_i)$ where (m_i, z_i) is the couple charge-mass in the i-th position of the grid. Then, the problem is reformulated as:

$$\mathbf{y} = \mathbf{D}\bar{\mathbf{x}} + \mathbf{n}' \tag{2}$$

where $\bar{\mathbf{x}}$ is a sparse vector with positive entries, for which the P non-zeros coefficients allow to determine the isotopic mass and charge state of each protein, along with their abundance. Moreover, $\mathbf{n}' = \mathbf{n} + \mathbf{e}$ models the acquisition noise and possible errors arising from the spectral analysis and discretization steps ($\mathbf{e} \to 0$ with high accuracy).

3. OPTIMIZATION STRATEGY

3.1. Variational formulation

Because of the presence of noise in $\mathbf{y}$ and the ill-conditioning of the dictionary matrix $\mathbf{D}$, direct inversion is not suitable to find an estimate of $\bar{\mathbf{x}}$. We propose instead to employ a penalized approach that defines the estimate $\widehat{\mathbf{x}} \in \mathbb{R}^T$ as a solution of the constrained minimization problem:

$$\underset{\mathbf{x} \in \mathbb{R}^T}{\text{minimize}} \;\; \Phi(\mathbf{x}) \quad \text{subject to} \quad \|\mathbf{D}\mathbf{x} - \mathbf{y}\| \leq \eta \tag{3}$$

where $\Phi : \mathbb{R}^T \mapsto (-\infty, +\infty]$ is a proper, lower semicontinuous, convex regularization function used to enforce positivity and sparsity on the solution, and $\eta > 0$ is a parameter that depends on the noise characteristics. When P is unknown, as it is usually the case in practical MS experiments, a simple choice for Φ is to consider, for every $\mathbf{x} \in \mathbb{R}^T$, $\Phi(\mathbf{x}) = \sum_{i=1}^{T} \max(0, x_i)$. More sophisticated penalties can also be used, involving for instance block-sparsity regularizers [16], or entropy-like priors [17].

3.2. Primal-dual optimization strategy

To resolve Problem (3), we propose to use the proximal Primal-Dual Splitting algorithm from [18] which is an efficient algorithm for convex optimization. In particular, this approach allows to treat efficiently the non-necessarily differentiable function Φ, and does not require any inversion step on the linear operator $\mathbf{D}$ [19].

Hereabove, $\text{prox}_{\tau\Phi}(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^T$, $\tau > 0$, states for the proximity operator of function $\tau\Phi$ at $\mathbf{x}$ [20] which is defined as the unique minimizer of $\tau\Phi + 1/2\|\cdot - \mathbf{x}\|^2$ [21]. Moreover, the projection operator $\text{proj}_{\|\cdot - \mathbf{y}\| \leq \eta}$ is defined, for every $(\mathbf{y}, \mathbf{v}) \in (\mathbb{R}^N)^2$, as:

$$\text{proj}_{\|\cdot - \mathbf{y}\| \leq \eta}(\mathbf{v}) = \mathbf{v} + (\mathbf{v} - \mathbf{y}) \min\left(\frac{\eta}{\|\mathbf{v} - \mathbf{y}\|}, 1\right) - \mathbf{y}. \tag{4}$$

The convergence of the iterates $(\mathbf{x}^{(k)})_{k \in \mathbb{N}}$ to a solution of Problem (3) is ensured, according to [18, 22].

Algorithm 1 Primal-Dual Splitting Algorithm

Initialization

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

$0 < \sigma < \|\mathbf{D}\|^2/\tau$, $\rho \in (0,2)$, $\tau > 0$

Minimization

For $k = 0, 1, \ldots$

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

4. PRACTICAL IMPLEMENTATION

We now discuss the practical implementation of our approach in the applicative context of MS spectrum analysis. In particular, the question of the computation and storage of matrix $\mathbf{D}$ is raised, and an approximated strategy is proposed for dealing with very large dimension.

4.1. Dictionary construction

As already mentioned, for any given (m, z), it is possible to estimate precisely the isotopic distribution $D(m, z)$ of an average protein having a monoisotopic mass m and charge state z using the so-called averagine model [23]. Given a range of masses $[m_{\min}, m_{\max}]$ and charges $[z_{\min}, z_{\max}]$, we define a regular grid:

$$(\forall i \in \{1, \ldots, T\}) \quad m_i = m_{\min} + (j-1)m_{\max}, \tag{5}$$
$$z_i = z_{\min} + (\ell - 1)z_{\max}, \tag{6}$$

with the convention $i = \ell M + j$, $j \in \{1, \ldots, M\}$ and $\ell \in \{1, \ldots, Z\}$. Then, the i-th column of $\mathbf{D}$ is taken as $\mathbf{d}(m_i, z_i)$ which corresponds to a sampled version of $D(m, z)$ on the mass grid with size M. Here, we propose to perform this sampling in the Fourier domain, and we normalize the result, so as to preserve the sum of squared amplitudes from $D(m, z)$ to $\mathbf{d}(m, z)$.

4.2. Circulant approximation

Mass spectrometry aims at providing a very high mass accuracy, so that the value of M can be very large. Even for small Z, the large number of columns of $\mathbf{D}$ presents a computational challenge as large memory resources may be needed to store this matrix. In order to avoid such memory issues, we propose an approximation $\overline{\mathbf{D}}$ of $\mathbf{D}$ whose structure will allow the use of Fourier transform operations for computing the products of $\mathbf{D}$ and $\mathbf{D}^\top$ with vectors. Our approximation relies on the important facts that (i) the isotopic patterns for similar mass values mainly differs by a simple shift of peaks positions, (ii) these patterns are sparse with non-zero elements located in a limited range of indexes near the monoisotopic mass value. Therefore, an alternative to the storage of isotopic patterns for each mass/charge couple that was proposed in Section 4.1, is to decompose the mass axis into windows onto which the isotopic pattern is assumed to be constant up to a circular shift. Let us introduce the notation $\mathbf{D} = [\mathbf{D}_1 |\ldots| \mathbf{D}_\ell |\ldots| \mathbf{D}_Z]$ where, for every $\ell \in \{1, \ldots, Z\}$, $\mathbf{D}_\ell \in \mathbb{R}^{M\times M}$ maps for the dictionary associated to charge $z_{\min} + (\ell - 1)z_{\max}$. Let $L \leq M$ the chosen window width and $\overline{\mathbf{d}}_{s,\ell}$ the average isotopic pattern for a mass within the range $[(s-1)L + 1, sL]$, and a fixed charge state $z_{\min} + (\ell - 1)z_{\max}$. We propose to approximate each $\mathbf{D}_\ell$ by the following block diagonal (BDiag) matrix made of M/L blocks assumed to be circulant (Circ) matrices with first line $\overline{\mathbf{d}}_{s,\ell}$, $s \in \{1, \ldots, M/L\}$:

$$\overline{\mathbf{D}}_\ell = \mathrm{BDiag}\left(\left[\mathrm{Circ}\left(\overline{\mathbf{d}}_{s,\ell}\right)\right]_{1\leq s\leq M/L}\right). \tag{7}$$

As a consequence, for every charge value, the products $\overline{\mathbf{D}}_\ell$ and $\overline{\mathbf{D}}_\ell^\top$ with vectors can be easily computed using Fourier operations. Under this approximation, Algorithm 1 can still be used to estimate the mass and charge positions, where $\mathbf{D}$ has now been replaced by $\overline{\mathbf{D}} = \left[\overline{\mathbf{D}}_1 |\ldots| \overline{\mathbf{D}}_\ell |\ldots| \overline{\mathbf{D}}_Z\right]$, and the norm of $\overline{\mathbf{D}}$ is computed using power iteration.

5. EXPERIMENTAL RESULTS

5.1. Synthetic data

In order to evaluate the performance of our method, we simulate two synthetic sparse signals A and B. Signal A models a mono-charge MS spectrum with $M = 3000$, $Z = 1$, $z_{\min} = z_{\max} = 1$ and $P = 10$ proteins. Signal B represents a multi-charge MS spectrum with $M = 5000$, containing $Z = 3$ charge values with $z_{\min} = 1$, $z_{\max} = 3$, and $P = 50$ proteins. In both cases, $m_{\min} = 1000$ Daltons, and $m_{\max} = 1100$ Daltons. Noisy data are then created using the linear model (2), where the noise is assumed to be zero-mean Gaussian, i.i.d, with known standard deviation $\sigma > 0$. We solve (3) with the penalty function Φ described in Section 3.1, and $\eta = \theta\sigma\sqrt{M}$ where $\theta > 0$ is a weight closed to 1.

Tab. 2 presents the results obtained for several noise levels. The quality of the results is evaluated in terms of signal to noise ratio (SNR) in dB defined as $10\log_{10}\left(\|\overline{\mathbf{x}}\|^2/\|\widehat{\mathbf{x}} - \overline{\mathbf{x}}\|^2\right)$, and in terms of the number of recovered peaks $\widehat{P}$. As one can observe, the exact dictionary approach and its block-circulant approximation both allow to recover the required number of peaks with good SNR values even for high noise level. The reconstruction quality can also be confirmed by a visual inspection of Fig. 1 which displays the input MS spectrum, the exact signal and the result of our method (using $\mathbf{D}$ or $\overline{\mathbf{D}}$) in the case of the multi-charged dataset B, and $\sigma = 10^{-2}$. Note that the signals are represented here along the abscissa axis (m/z), which is a standard representation in the context of MS. We also provide the computation time when running our method on a Macintosh Mac Pro Intel Xeon with a total of 8 cores, equipped with 12 GB of memory and running MacOsX 10.7.5. Algorithm 1 is used with $\rho = 1.99$, $\tau = \|\mathbf{D}\|^{-1}$, $\sigma = 0.9\tau$, and it is stopped whenever the relative error between two consecutive iterates is lower than 10^{-8}. In our tests, a maximum of 1000 iterations was sufficient to reach such precision. As expected, when using the block-circulant approximation of the dictionary matrix, the computation time depends on the value of L, the fastest reconstruction being obtained with a large L. Moreover, even if the reconstruction quality is slightly deteriorated, the SNR values remain stable as soon as L is sufficiently large as it can be observed on Fig. 2. According to our practical observations, the value of L mainly impacts on the peak intensity values, and has little effect on the estimation quality of peak positions, which is of main interest in the MS application. We also provide in Tab. 2 the memory required for the storage of the dictionary elements when using the exact matrix $\mathbf{D}$. As expected, the memory requirement in that case may be quite high, especially in the multi-charged case (dataset B). In contrast, the block-circulant approximation we proposed allows to avoid any dictionary storage, since the products with $\overline{\mathbf{D}}$ and its adjoint are performed using Fourier operators where the first lines of each circulant term are computed on the fly.

		Exact dictionary approach				Block-circulant approximation											
						$L = 2$			$L = 10$			$L = 50$			$L = 100$		
	σ	SNR	Time	$\widehat{P}$	Memory	SNR	Time	$\widehat{P}$	SNR	Time	$\widehat{P}$	SNR	Time	$\widehat{P}$	SNR	Time	$\widehat{P}$
Dataset A, P=10	1	15.07	10.18	10	72	15.05	15.16	10	11.31	3.31	10	13.88	0.95	10	11.37	0.61	10
	0.1	35.09	10.40	10		34.98	15.22	10	31.95	3.32	10	32.60	0.27	10	32.28	0.24	10
	0.01	38.38	10.46	10		34.33	15.14	10	34.44	3.35	10	34.92	0.94	10	36.91	0.61	10
Dataset B, P=50	1	16.18	303.33	50	600	15.57	127.85	50	13.99	27.20	50	14.01	6.98	50	15.59	3.52	50
	0.1	35.73	206.84	50		35.43	44.48	50	33.95	8.19	50	33.53	3.09	50	28.26	2.06	50
	0.01	39.56	377.80	50		38.38	290.56	50	34.40	58.92	50	35.44	11.66	50	28.99	16.67	50

Table 2: SNR (in dB), computation time (in s), memory storage (in Mb) for matrix $\mathbf{D}$ and number $\widehat{P}$ of detected peaks for the restored signals A and B for various values of noise level. Block-circulant approximation $\overline{\mathbf{D}}$ is tested for four L values.

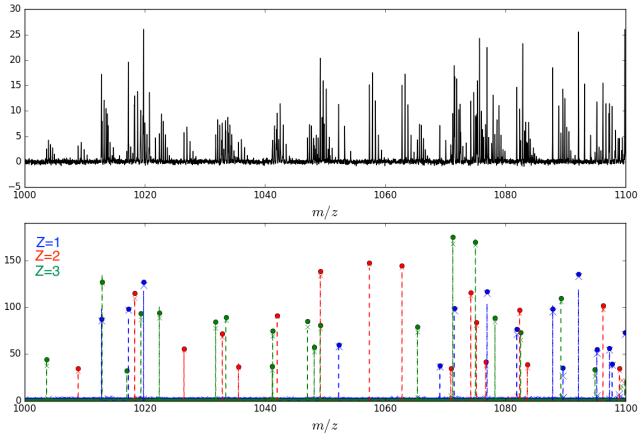

Fig. 1: Reconstruction results of the signal B with $\sigma = 10^{-2}$: (top) input data $\mathbf{y}$, (bottom) exact spectrum (dots), restored spectrum with exact dictionary (dashed line), and with its block-circulant approximation for $L = 10$ (asterisks).

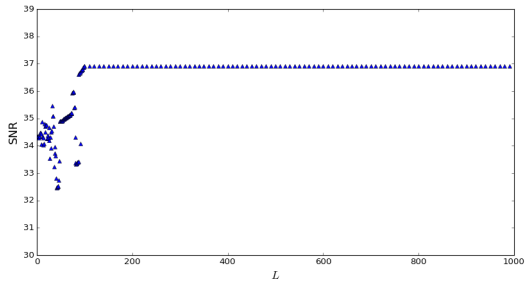

Fig. 2: Reconstruction results of the signal A with $\sigma = 10^{-2}$: SNR of the restored spectrum using $\overline{\mathbf{D}}$ for various values of L.

5.2. Real data

We also perform numerical tests on a real MS dataset measured on a Brucker Solarix 15 T, FT-ICR instrument with an ESI source. The considered sample was constituted of $3\,\mu$M of the peptide EVEALEKKVAALESKVQALEKKVEALEHG-NH2 ($C_{140}H_{240}N_{38}O_{45}$) in its trimer form within 50 mM of NH_4OAc, acquired in native conditions. The input data is of size $M = 8130981$ with $m_{\min} = 153.57$ Daltons, $m_{\max} = 4999.96$ Daltons, and $Z = 5$ with $z_{\min} = 1$ and $z_{\max} = 5$. Parameter η is set in a similar manner than for the synthetic data, the noise level σ being estimated from an empty frame of the measured signal.

Fig. 3(top) displays a zoom on the input signal $\mathbf{y}$ for m/z within $[1890, 1920]$ Daltons. We also show on Fig. 3(bottom) the recovered

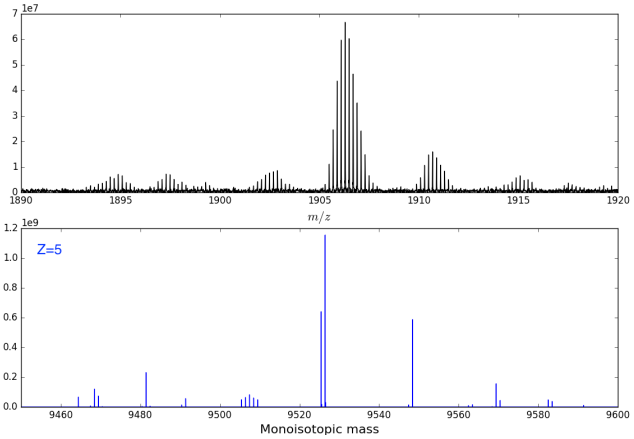

Fig. 3: Analysis of the real FT-ICR-MS spectrum of a peptide in trimer form: (top) zoom on the acquired data; bottom) recovered spectrum at $z = 5$ using block-circulant approximated dictionary with $L = 10$.

signal along the monoisotopic mass axis, for the same mass range, and $z = 5$ charge state. It is worth mentioning that, due to the very large size of the dataset, such results were only made possible by the block-circulant approximation we proposed, since the storage of a dictionary would have required GigaBytes of memory. Here, the best visual results were obtained using $L = 10$, with a processing time of 108 minutes. A major peak can be distinguished at $m = 9526.439$ Daltons which fits well with the theoretical monoisotopic mass of the studied peptide equals to $m = 9526.337$ Daltons. A second peak, shifted by -1 Dalton, is also observed, due to unavoidable grid ambiguity. Finally, we observe a third important peak distant with $+21.959$ Daltons of the main peak, which allows to identify the sodium adduct (with theoretical relative position of $+21.982$ Daltons), thus validating the faithfulness of our approach.

6. CONCLUSION

This work presents a new dictionary-based approach based on the averagine function to solve the isotopic pattern analysis problem arising in Mass Spectrometry of proteins. We propose a sparsity-aware variational strategy to determine the dictionary elements, associated with a primal-dual splitting minimization strategy. To counteract any computation burden, we propose a suitable block-circulant approximation of the dictionary. Our experimental results illustrate the efficiency of our method to solve the MS problem. Future work will address the extension of the approach to the processing of multi-dimensional MS spectra [24, 25].

7. REFERENCES

[1] L.-M Heaney, D.J. Jones, and T. Suzuki, "Mass spectrometry in medicine: a technology for the future?," *Future Science OA*, vol. 3, no. 3, June 2017.

[2] I.-E. Sodal, "The medical mass spectrometer," *Biomedical Instrumentation & Technology*, vol. 23, no. 6, pp. 469–476, 1989.

[3] O. Gaillot, N. Blondiaux, C. Loiez, F. Wallet, N. Lemaitre, S. Herwegh, and R. Courcol, "Cost-effectiveness of switch to matrix-assisted laser desorption ionization-time of flight mass spectrometry for routine bacterial identification," *Journal of Clinical Microbiology*, vol. 49, no. 12, pp. 4412–4412, 2011.

[4] S. A. Schwartz, R. J. Weil, M. D. Johnson, S. A. Toms, and R. M. Caprioli, "Protein profiling in brain tumors using mass spectrometry," *Clinical Cancer Research*, vol. 10, no. 3, pp. 981–987, 2004.

[5] F. J. Belas and I. A. Blair, "Mass spectrometry in pharmaceutical analysis," *Journal of liposome research*, vol. 11, no. 4, pp. 309–342, 2001.

[6] A.-G. Ferrige, M.-J. Seddon, B.-N. Green, S.-A. Jarvis, J. Skilling, and J. Staunton, "Disentangling electrospray spectra with maximum entropy," *Rapid Communications in Mass Spectrometry*, vol. 6, no. 11, pp. 707–711, 1992.

[7] A. Mohammad-Djafari, J.-F. Giovannelli, G. Demoment, and J. Idier, "Regularization, maximum entropy and probabilistic methods in mass spectrometry data processing problems," *International Journal of Mass Spectrometry*, vol. 215, no. 1, pp. 175 – 193, 2002, Detectors and the Measurement of Mass Spectra.

[8] J. Samuelsson, D. Dalevi, F. Levander, and T. Rognvaldsson, "Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting," *Bioinformatics*, vol. 20, no. 18, pp. 3628–3635, 2004.

[9] B. Y. Renard, M. Kirchner, H. Steen, J. A. J. Steen, and F. A. Hamprecht, "Nitpick: peak identification for mass spectrometry data," *BMC Bioinformatics*, vol. 9, no. 1, pp. 355, 2008.

[10] D. P. A. Kilgour, S. Hughes, S. L. Kilgour, C. L. Mackay, M. Palmblad, B. Q. Tran, Y. A. Goo, R. K. Ernst, D. J. Clarke, and D. R. Goodlett, "Autopiquer - a robust and reliable peak detection algorithm for mass spectrometry," *Journal of The American Society for Mass Spectrometry*, vol. 28, no. 2, pp. 253–262, Feb. 2017.

[11] J.-A. Yergey, "A general approach to calculating isotopic distributions for mass spectrometry," *International Journal of Mass Spectrometry and Ion Physics*, vol. 52, no. 2-3, pp. 337–349, 1983.

[12] P. Kaur and P.-B O'Connor, "Use of statistical methods for estimation of total number of charges in a mass spectrometry experiment," *Analytical Chemistry*, vol. 76, no. 10, pp. 2756–2762, 2004.

[13] M. Mann, C. K. Meng, and J. B. Fenn, "Interpreting mass spectra of multiply charged ions," *Analytical Chemistry*, vol. 61, no. 15, pp. 1702–1708, 1989.

[14] G. Audi and A.-H. Wapstra, "The 1993 atomic mass evaluation:(i) atomic mass table," *Nuclear Physics A*, vol. 565, no. 1, pp. 1–65, 1993.

[15] K.-J.-R Rosman and P.-D.-P. Taylor, "Table of isotopic masses and natural abundances," *Pure and Applied Chemistry*, vol. 71, pp. 1593–1607, 1999.

[16] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Structured sparsity through convex optimization," *Statistical Science*, vol. 27, no. 4, pp. 450–468, 2012.

[17] A. Cherni, E. Chouzenoux, and Delsuc M.-A., "Proximity for a class of hybrid sparsity + entropy prior. Application to DOSY NMR signal reconstruction," in *Proceedings of the 8th International Symposium on Signal, Image, Video and Communications (ISIVC 2016)*, Tunis, Tunisia, 21-23 Nov. 2016, pp. x–x+6.

[18] A. Chambolle and T. 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.

[19] N. Komodakis and J.-C. Pesquet, "Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 31–54, Oct. 2015.

[20] H.H. Bauschke and P.L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*, CMS Books in Mathematics. Springer, New York, NY, 2011.

[21] J.-J. Moreau, "Proximité et dualité dans un espace hilbertien," *Bulletin de la Société mathématique de France*, vol. 93, pp. 273–299, 1965.

[22] L. Condat, "A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms," *Journal of Optimization Theory and Applications*, vol. 158, no. 2, pp. 460–479, Aug. 2013.

[23] M.-W. Senko, J.-P Speir, and F.-W. McLafferty, "Collisional activation of large multiply charged ions using Fourier transform mass spectrometry," *Analytical Chemistry*, vol. 66, no. 18, pp. 2801–2808, 1994.

[24] M.-A. Agthoven, Delsuc M.-A., G. Bodenhausen, and C. Rolando, "Towards analytically useful two-dimensional fourier transform ion cyclotron resonance mass spectrometry," *Analytical and Bioanalytical Chemistry*, vol. 405, pp. 51–61, Oct. 2013.

[25] F. Floris, M.-V. Agthoven, L. Chiron, A.-J. Soulby, C. A. Wootton, Y.-P.-Y. Lam, M.-P. Barrow, M.-A. Delsuc, and P.-B. O'Connor, "2D FT-ICR MS of Calmodulin: A Top-Down and Bottom-Up Approach.," *Journal of the American Society for Mass Spectrometry*, vol. 27, no. 9, pp. 1531–1538, Sep. 2016.