Matching experimental and accurately-simulated Orbitrap mass spectra improves feature extraction and annotation in the analysis of small and large molecules

Konstantin O. Nagornov,[1] Sergey Girel,[2] Anton N. Kozhinov,[1] Serge Rudaz,[2] and Yury O. Tsybin[1]

[1]Spectroswiss, EPFL Innovation Park, 1015 Lausanne, Switzerland; [2]Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, 1211 Geneva, Switzerland.

Introduction

FTMS instruments, such as Orbitraps, often generate mass spectral datasets of very high complexity that require advanced data processing and analysis software tools.

Rigorous feature extraction from FTMS data determines the quality and confidence of the subsequent data analysis and biologically-relevant conclusions. Furthermore, data processing procedures must be computationally efficient and suitable for automation.

We suggest that consideration of the specific nature of the FTMS data, namely the underlying time-domain transient processing, will enhance the feature extraction performance. Previously, we reported on developing the FTMS Simulator – a software tool to accurately simulate FTMS data via modeling the time-domain transients.

Here, we report on developing and evaluating **feature extraction** workflows for metabolomics, complex mixtures and protein analysis based on the **FTMS Simulator**.

Workflow

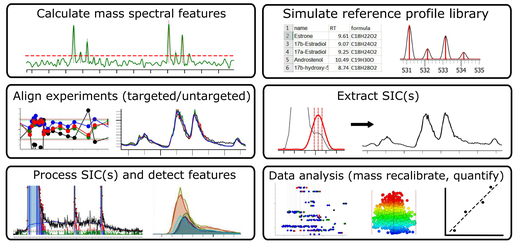


Figure 1. The MS only workflow for data annotation and quantification based on feature extraction from LC/MS or GC/MS experiments via transient mediated simulations:

1. Calculate mass spectral features: mass spectra baseline correction, data dependent noise thresholding, three-point interpolation peak picking.
2. Simulate reference profile library: automatic instrument selection, resolution settings and FT processing parameters; simulation of isotopic envelope profile and centroids via transient calculation for compounds of interest in targeted database.
3. Align experiments: targeted - user selected mass values from mass spectra or reference compounds; untargeted - parameter-dependent feature extraction, without annotation;
4. Extract SIC(s): group compounds of interest (single compounds, charge state grouping, compound classes); isotopic group selection (A, A+1, etc.); mass tolerance based SIC extraction.
5. Feature detection: baseline correction, smoothing and noise thresholding of SIC(s) optionally, detect feature (SIC peak) based on its prominence and matching experimental and simulated data; filter detected features: minimum points per peak, number of adducts, etc.
6. Data analysis: re-calibration of mass spectra; mass accuracy plots; dynamic global results matrix overview; hits map, quantification: absolute, semi, standard addition.

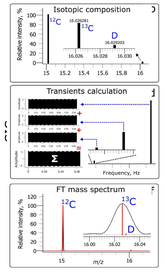


Figure 2. The workflow of isotopic envelope simulation via time domain signal (transient) calculation. Transients are calculated for each isotope separately and averaged together.

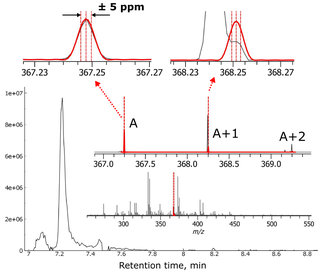


Figure 3. Example of selected ion current (SIC) calculation in LC/MS experiment. SIC was calculated for three isotopologues of the pre-calculated isotopic envelope of a target compound. In fact, each data point of SIC is the summation of experimental intensities of isotopologues found within mass tolerance window of ± 5 ppm.

Protein analysis (average mass)

Analysis of proteins of any size (10 kDa - 1 MDa) at low resolution acquisition settings, where isotopic envelopes are unresolved and only average mass can be deduced, may be improved using the combination of a standard untargeted deconvolution (UniDec, etc.) with targeted deconvolution method via simulations, Figures 4-6.

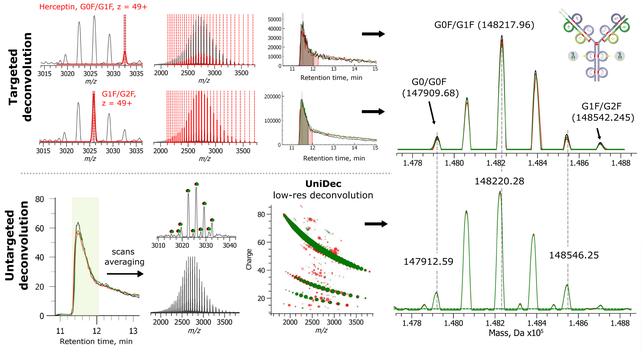


Figure 4. Implementation of a combined deconvolution method: targeted - developed in this work and a standard (untargeted) deconvolution method. **Targeted** (top panels)**:** broadband charge state profile patterns (red) were simulated for a number of expected proteoforms of Herceptin and Orbitrap™ Q Exactive HF™ instrument. SICs were calculated for the simulated reference patterns and features were detected according to the "Feature Detection" algorithm (see the "Workflow" section). **Untargeted** (bottom panels)**:** standard low resolution deconvolution was performed using the UniDec algorithm (DOI: 10.1021/acs.analchem.5b00140), integrated into the software tool. (Right panel) the deconvolved spectra as output of both algorithms were analyzed together.

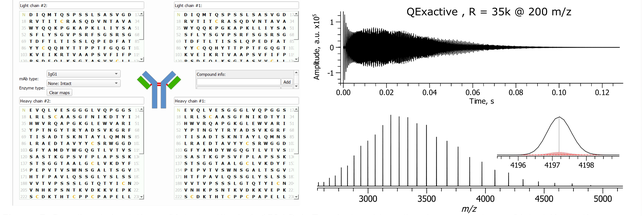


Figure 5. Broadband simulations of intact antibody (~150 kDa). Transient and broadband mass spectrum of intact Herceptin were simulated for Orbitrap™ Q Exactive™ instrument settings for charge states of z = 30+ - 60+.

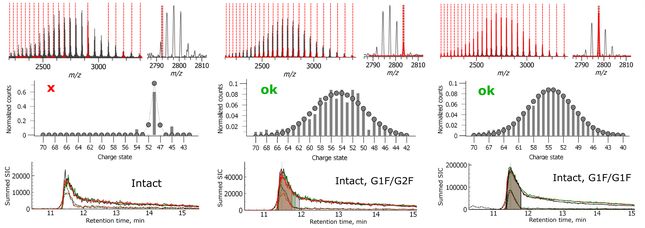


Figure 6. Filtering of features, detected using the targeted deconvolution method. Top panel shows the overlay of experimental and simulated data for different Herceptin proteoforms acquired with Orbitrap™ Q Exactive HF™ instrument. Middle panel shows the extracted charge state distributions used as an automatic validation parameter for feature detection. Bottom panel shows SIC and detected features in it for the proteoforms of interest. The SIC feature is detected if charge state distribution of the corresponding proteoform satisfies a number of parameters: a minimum number of charge states, a deviation from a normal distribution, etc.

Protein analysis (isotopic mass)

Protein analysis performed at high resolution, sufficient to provide isotopically resolved and mono isotopic mass can be detected, may be also improved using the combination of a standard untargeted deconvolution (Xtract, Thrash, Hardklor, etc.) and targeted deconvolution method via simulations, Figures 7-9.

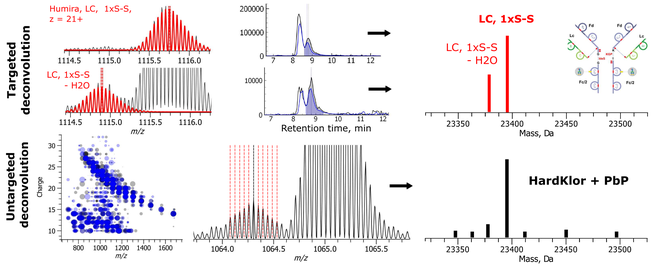


Figure 7. Implementation of a combined targeted deconvolution method, developed in this work, with a standard (untargeted) deconvolution method. **Targeted** (top panels)**:** multiple charge state profile isotopically resolved patterns (red) were simulated for number of expected subunits of IdeS digested Humira and Orbitrap™ Q Exactive HF™ instrument. SICs were calculated for the simulated reference patterns and features were detected according to the "Feature Detection" algorithm (see the "Workflow" section). **Untargeted** (bottom panels)**:** standard high resolution deconvolution was performed using the HardKlor algorithm, integrated into the software tool. (Right panel) the deconvolved spectra as output of both algorithms were analyzed together.

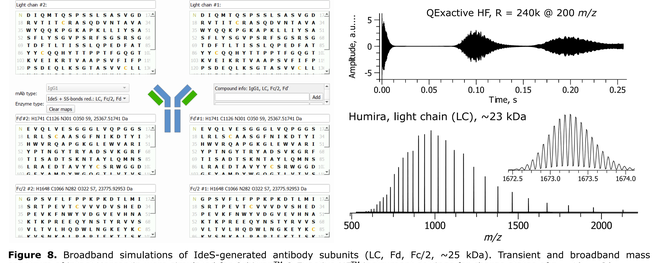


Figure 8. Broadband simulations of IdeS-generated antibody subunits (LC, Fd, Fc/2, ~25 kDa). Transient and broadband mass spectrum of intact Herceptin were simulated for Orbitrap™ Q Exactive HF™ instrument settings for charge states of z = 10+ - 30+.

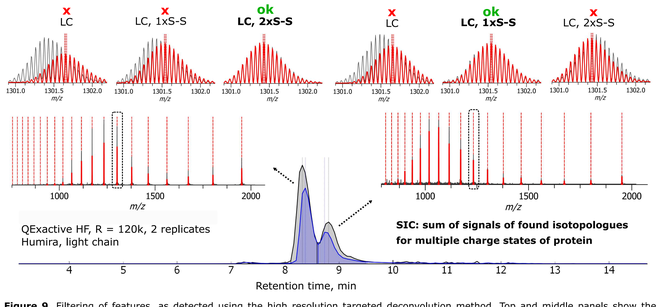
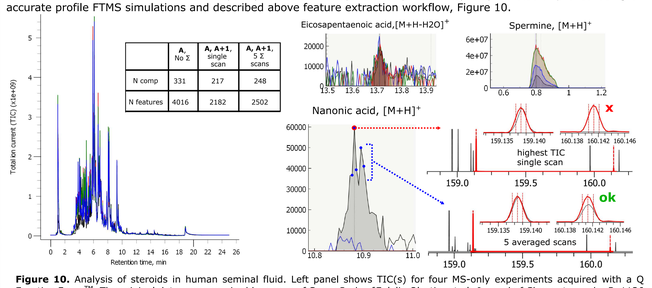


Figure 9. Filtering of features, as detected using the high resolution targeted deconvolution method. Top and middle panels show the overlay of (black) experimental and (red) simulated data for different proteoforms of Humira light chain acquired with Orbitrap™ Q Exactive HF™. Bottom panel shows SIC and detected features in it for the proteoforms of interest. The SIC feature is detected, if experimental and simulated data are matched within the user-defined similarity score.

Small molecules

The confidence for small molecules analysis (metabolomics, lipidomics, proteomics) may be also improved using an accurate profile FTMS simulations and described above feature extraction workflow, Figure 10.



	A, No Σ	A, A+1, single scan	A, A+1, 5 Σ scans
N comp	331	217	248
N features	4016	2182	2502

Figure 10. Analysis of steroids in human seminal fluid. Left panel shows TIC(s) for four MS-only experiments acquired with a Q Exactive Focus™. The original data were acquired by group of Serge Rudaz [Eulalia Olestia, et.al. *Journal of Chromatography B*, 1136 (2020) 121929]. Here we performed data annotation using 789 steroids database and mass tolerance of ± 5 ppm. Top right panels show the detected features for the high abundance and low abundance compounds different by 4 orders of magnitude. The identification confidence is improved using two isotopologues instead of only the monoisotopic peak. The total number of identified compounds was increased via spectral averaging averaging.

Compound classes

The analysis of complex mixtures may be improved by decreasing the complexity of mass spectra using LC or GC sample separation. Similarly to the individual compound analysis, compound classes can be qualified and quantified via simulations and further grouping of the corresponding multiple compounds, Figure 11.

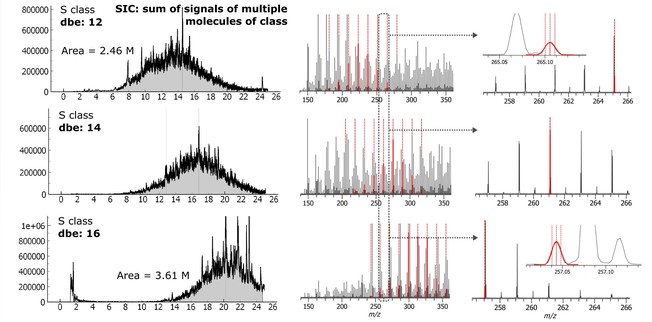


Figure 11. LC-MS analysis of a complex mixture on a 7 T LTQ FT Ultra (Thermo Scientific). Left panel shows a selected ion chromatogram (SIC) calculated for monoisotopic peaks of multiple compounds of the S class ($C_nH_{n+2}S_1$) and different DBE values (12, 14, and 16). Middle and right panels show distribution of the grouped isotopic envelopes over a broadband mass range at different DBE values. Experimental data are shown in black and simulated data are shown in red.

Conclusions

- Accurate simulation of FTMS (profile) mass spectra for small and, especially, large molecules open the doors toward accurate, rapid, and targeted deconvolution based on feature extraction;
- Features are confirmed by similarity scoring between the isotopic envelopes of the experimental and simulated data. The isotopic envelopes can be matched individually or grouped. The groups can contain charge state distributions, proteoform clusters (e.g., glycosylation profiles of mAbs), or classes of compounds, e.g., N-class in petroleomics-type measurements;
- FTMS specific software tool was developed for targeted feature extraction from LC/GC-MS data for molecules of any size. The tool is supported with a graphical user interface, demonstrated excellent analytical specificity, high sensitivity, and quantitative precision when applied to metabolomics, complex mixture and protein analysis.

We acknowledge financial support from European Horizon 2020 research and innovation program under grant agreement No 829157 (TopSpec).

Poster presented at the 70-th ASMS Conference, Minneapolis, MN, USA, June 5-9, 2022