findMF - analysis of qTOF DIA - LC MS/MS images

Witold E Wolski[1,2*] and Ruedi Aebersold[1]

[1]Institute for Molecular System Biology, ETH Zurich

[2]Systems Biology IT (SyBIT), SystemsX.ch Zurich

ABSTRACT

Motivation: Liquid chromatography-mass spectrometry measurements record ion counts of molecules which are annotated with two dimensions: hydrophobicity and mass over charge ratio (m/z). *findMF* (find Molecular Features) software evaluates these data with image processing methods to recognize and remove noise, and to detect isotope features with high sensitivity and accuracy.

Availability and Implementation: *findMF* is freely available as C++ source code from https://github.com/findmf under a BSD style license. Building on Linux is regularly tested on https://travis-ci.org/findMF/findMFHCS. All *findMF* dependencies are freely available under either BSD or Apache style licenses.

1 INTRODUCTION

Liquid chromatography (LC) connected on-line to mass spectrometry (MS) constitutes a system that separates molecules with high resolution, based on two orthogonal properties: hydrophobicity represented by the chromatographic retention time (RT) domain and mass over charge (m/z) ratio. LC-MS/MS operated in data independent acquisition (DIA) mode cycles through the precursor-ion m/z range in segments of specified width providing an additional MS dimension for ion separation [1]. DIA-LC-MS/MS allows recording millions of molecular features (MF) in a single experiment.

Due to the two dimensional (2D) nature of LC-MS separation, image processing methods are well suited to scrutinize raw data files [2]. *findMF* software was developed to support various levels of data processing, specifically: i) resampling of profile mode spectra and LC-MS runs to spectra or images; ii) removal of electronic and chemical noise from spectra and images; iii) detection of peaks in spectra as well as of isotope features (2D peaks) in images.

findMF is geared toward quadrupole - time of flight (qTOF) instruments operated in data dependent acquisition (DDA) and DIA mode, but could be extended to Fourier transform instruments.

findMF consists of two separate entities, *findMFBase* and *findMFHCS*, which are built and tested separately. *findMFBase* contains C++ functions, the interfaces of which are inspired by the corresponding GNU-R functions (i.e. filter, interpolate, diff, rnorm, dnorm, median, mean, sd, kurtosis and skewness, as well as their weighted versions). This simplifies porting of peak detection prototypes implemented in S (GNU-R) language to C++ while at the same time assures collection independence - known from C++ STL algorithms - - by employing the iterator pattern. In addition, *findMF* includes algorithms to resample spectra or to detect peak apices. To simplify the use of *findMFBase* by other projects, we implemented it as a header only library, and the only non-standard dependencies are the boost-C++ libraries (www.boost.org).

findMFHCS implements command line applications for qTOF MS data processing: the **qtofpeakpicker** detects peaks in single spectrum; **conv2Image** generates, pre-processes images and stores them as TIFF; **filterSingleMap** denoises and removes background from LCMS maps; **findMF** detects isotope features and stores them in a configuration-free relational database (*sqlite*), which greatly simplifies data exchange and downstream data analysis. *findMFHCS* depends on proteowizard [4] to access raw profile data; *findMFBase* to resample the spectra; VIGRA to process and analyze images; and Qt to provide visualization and connectivity with *sqlite*.

Overall, *findMF* can be used to convert raw MS data into either dense (images) or structured data representations such as peak-list or isotope features, facilitating image processing and targeted or exploratory analysis of LC-MS/MS DDA or DIA data.

2 METHODS

The feature detection implemented in **findMF** tool (Figure 1 top) consists of:

2.1 Data import and image generation

The applications provided by *findMFHCS* employ the file format agnostic proteowizard API, which supports binary vendor formats. LC-MS/MS DIA datasets typically consist of 200k spectra sampled with 500k data points per spectrum, which represents a 1000-fold increase in the number of data points recorded, as compared with LCMS experiments performed a few years ago, rendering text-based raw data formats impractical.

The width of peaks in TOF instruments changes with mass (m) and is approximated by $FWHM = m/R$, where $FWHM$ is the full width at half maximum peak height and R, the instrument resolution. As the distance between sampling points changes slower ($\sim\sqrt{m}$), peaks at higher mass are sampled with more data points than those at lower mass. Appreciably, a constant number of sampling points per peak would simplify spectra smoothing and peak detection. Therefore, we resample the raw data using a mass-dependent bin width. Given the value of m and R, together with the number of points with which the peaks are to be sampled, we can determine new sampling points for each peak. To determine, for the result spectrum, the number of bins among which the input intensity values are to be allotted, the bin width in the input spectrum needs to be specified. This input bin width at mass, m can be determined by $\Delta m = \sqrt{m}/a$, where a can be inferred from the input data.

Modern MS instruments have a practical resolution of ~20k. If a peak apex is sampled with 3 to 5 points and there are ~3000 spectra per map, the generated image can be rather large (1.5 GByte).

2.2 Denoising and background removal

Electronic and chemical noise resulting from ion count statistics is reduced by Gaussian smoothing of the 2D image. Chemical noise and detector ringing are removed by subtracting the background in the RT and *m* dimensions using morphological filters. The choice of algorithms is primarily guided by performance considerations. Chemical noise removal in the RT domain increases signal specificity and sensitivity as previously demonstrated [2]. The de-noised features are well-defined and their location and intensities can be reliably determined (Figure 1 Bottom).

2.3 Feature detection and storage

Features are detected by local maximum searches and extended by seeded region growing to a minimum threshold or until a ridge (see

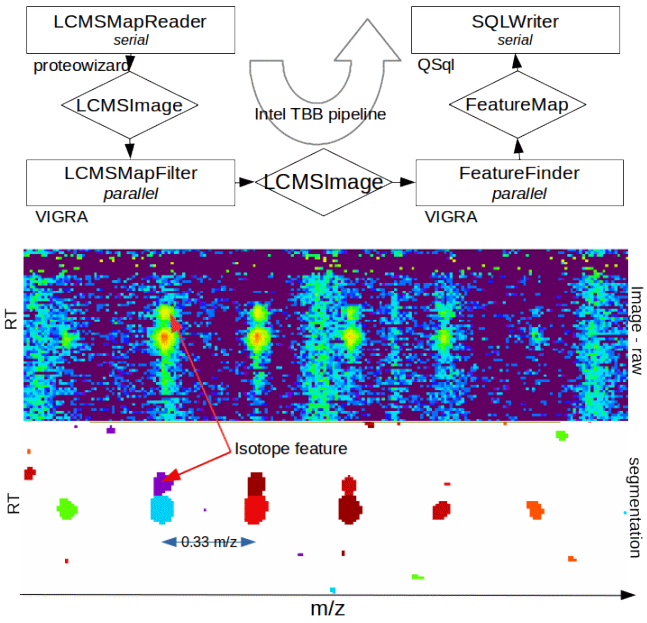

Figure 1 Top: Feature detection workflow implemented in findMF tool (squares - algorithms, rhombs - data structures), connected using the Intel TBB pipeline framework. Bottom: feature detection visualization using findMF **filterQtView** *tool: top - resampled input image intensities are color coded (terrain colors), bottom panel - detected segments (colors are used to distinguish segments).*

Figure 1, Bottom). Subsequently, feature statistics (i.e. peak location, area, volume and measures of peak shape) are computed using VIGRA or *findMFBase* functions to enable i.e. features classification. To provide visualization, the projections of the features in *m* and RT dimensions are stored. While HDF5 is the format of choice for storing raw MS or centroid data [6], *sqlite* is well suited for storing structured data such as isotope features.

3 RESULTS AND CONCLUSION

Processing of a LC-MS map takes 7−15 minutes, depending on instrument resolution and sample complexity. A single DIA LC-MS/MS measurement produces 32, 64 or more images. On a 4-core desktop computer, we can reduce the processing of an experiment with 32 maps to 2-3h using the Intel TBB pipeline framework. Feature detection is a variation of lossy data reduction, and the *sqlite* feature file is approximately 2-4 times smaller than the ABSciex input wiff file.

Sensitivity of the feature detection was examined by comparing with results of manually annotated DIA-LC-MS/MS experiments [5]. Depending on the complexity of the input matrix (water, yeast or human origin), the recall rates of spiked-in peptides were 71%, 93%, and 94% respectively (aggregate over 30 datasets and 10 dilution series). Interestingly, with increasing matrix complexity the recall rates improved.

We also examined how fold changes of dilution series can be reproduced. For complex samples (human and yeast background), the fold change estimation was more accurate but less precise compared with manual peak annotation.

Image processing techniques constitute computationally feasible approaches to process 2D spectrometric data. *findMF* intends to facilitate prototyping of MS applications based on image processing with algorithms and tools to convert qTOF MS data into images. Our image processing based feature detection demonstrated high sensitivity and quantification capacity when tested on various datasets. *sqlite* is well suited to store millions of isotope features in a single file, exhibits superior write and read performance to text-based alternatives and we recommend it for storing and sharing structured scientific data.

Software projects with few dependencies and a well-defined problem domain facilitate code reuse. The ongoing project aims at providing the feature *sqlite* API (used within *findMFHCS*) to a separate C++ project.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Dr Christian Panse (FGCZ Zurich), Dr Maciej Lalowski (University of Helsinki) and Ching Chiek Koh and Dr Tiannan Guo (IMSB Zurich) for useful discussions on the project.

REFERENCES

[1] Ludovic Gillet, et.al. (2012) Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. Mol Cell Proteomics,

[2] Kyriacos Leptos, et.al. (2006) Mapquant: open-source software for large-scale protein quantification. Proteomics

[3] Salvatore Cappadona, et.al. (2008) Wavelet-based method for noise characterization and rejection in high-performance liquid chromatography coupled to mass spectrometry. Anal Chem

[4] Darren Kessner et.al. (2008) Proteowizard: open source software for rapid proteomics tools development. Bioinformatics

[5] Hannes Roest, et.al. (2014) Openswath: Automated, targeted analysis of data-independent acquisition (dia) ms-data. Nature biotechnology

[6] Mathias Wilhelm, et al. (2012) mz5: space and time efficient storage of mass spectrometry data sets. Mol Cell Proteomics