Bioinformatics, 2022, 1–2
https://doi.org/10.1093/bioinformatics/btac638
Advance Access Publication Date: 21 September 2022
Applications Note

Sequence analysis

Increasing confidence in proteomic spectral deconvolution through mass defect

Milan A. Clasen [1,*], Louise U. Kurt [1], Marlon D. M. Santos [1], Diogo B. Lima [2], Fan Liu[2], Fabio C. Gozzo[3], Valmir C. Barbosa [4] and Paulo C. Carvalho [1,*]

[1]Laboratory for Structural and Computational Proteomics, Carlos Chagas Institute, Fiocruz—Paraná 81310-020, Brazil, [2]Department of Structural Biology, Leibniz—Forschungsinstitut für Molekulare Pharmakologie (FMP), Berlin 13125, Germany, [3]Dalton Mass Spectrometry Laboratory, Unicamp, Campinas 13083-970, Brazil and [4]Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro, Rio de Janeiro 21941-972, Brazil

*To whom correspondence should be addressed.

Associate Editor: Olga Vitek

Received on May 2, 2022; revised on August 24, 2022; editorial decision on September 11, 2022; accepted on September 19, 2022

Abstract

Motivation: Confident deconvolution of proteomic spectra is critical for several applications such as *de novo* sequencing, cross-linking mass spectrometry and handling chimeric mass spectra.

Results: In general, all deconvolution algorithms may eventually report mass peaks that are not compatible with the chemical formula of any peptide. We show how to remove these artifacts by considering their mass defects. We introduce Y.A.D.A. 3.0, a fast deconvolution algorithm that can remove peaks with unacceptable mass defects. Our approach is effective for polypeptides with less than 10 kDa, and its essence can be easily incorporated into any deconvolution algorithm.

Availability and implementation: Y.A.D.A. 3.0 is freely available for academic use at http://patternlabforproteomics.org/yada3.

Contact: milanclasen@gmail.com or paulo@pcarvalho.com

Supplementary information: Supplementary information is available at *Bioinformatics* online.

1 Introduction

Mass spectra from proteomic experiments originate mainly from ionized polypeptides that appear in a spectrum as clusters of peaks whose relative intensities (*y*-axis) and *m*/*z* (*x*-axis) result from the isotopic distribution of their elements; these clusters are the so-called isotopolog envelopes. De-isotoping a mass spectrum aims to simplify it by summing the intensities of all ions in an envelope into a single (monoisotopic) peak. Likewise, de-charging is accomplished by creating surrogate spectra whose corresponding ions are all singly charged. Finally, deconvolution is the combined process of de-isotoping and de-charging a mass spectrum.

In proteomics, deconvoluting spectra have proven useful for simplifying various tasks. For example, years ago our group introduced Yet Another Deconvolution Algorithm (Y.A.D.A.) and showed that it could increase peptide identification by some 10% when applied to chimeric spectra (Carvalho *et al.*, 2009). At about the same time, the authors of MS-Deconv demonstrated that precise deconvolution was critical for searching complex tandem mass spectra of intact proteins (Liu *et al.*, 2010). Advances in deconvolution speed have also been reported; for example, FLASHDeconv was optimized for high speed on top-down data by using a log-transform to make the search for mass spectral patterns 'ultrafast' and with fewer artifacts (Jeong *et al.*, 2020). Nevertheless, despite how sophisticated modern deconvolution algorithms have become, there is always room for improvements when assigning monoisotopic ions, with no clear way to measure the proportion of misassignments or to benchmark algorithms.

The term 'mass defect', in chemistry, refers to the difference between an atom's exact mass and its mass number. For each of the 20 amino acids, the mass defect will be simply the sum of the decimal part of its atoms' exact masses. Here, we propose taking the peaks' mass defects into consideration and making them part of a deconvolution algorithm's core, aiming to single out those that are not compatible with the mass of any peptide. We demonstrate the effectiveness of our approach on a previously generated Q-Exactive™ Plus raw file acquired from human cancer (meningioma) cells (Silva *et al.*, 2020).

We also present a revamped, about 10 times faster version of Y.A.D.A. We refer to the original version as Y.A.D.A. 1.0, to the new one as Y.A.D.A. 3.0. In contrast to FLASHDeconv, Y.A.D.A. 3.0 is faster on low-mass datasets. Details are given as Supplementary Information.

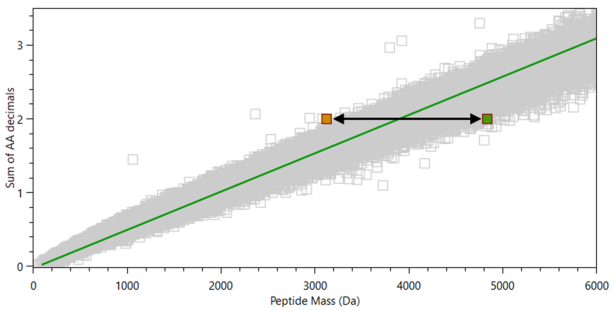

Fig. 1. Sum of a peptide's amino acid residue mass defects (decimals) versus its monoisotopic mass. The diagonal line is Model 1, relating each sum of amino acid decimals to a central monoisotopic mass. An arrow stretches to the left and right of this line for a fixed sum, each arrow three standard deviations long, representing the acceptable interval for mass given the sum. This standard deviation is given by Model 2

2 Materials and methods

Y.A.D.A. 3.0 can generate a model for acceptable polypeptide masses as follows:

1. For all peptides in a theoretical proteome digest, plot the sum of each peptide's amino acids' mass defects as a function of the nominal mass of the peptide (Fig. 1). Our default model considers all *Homo sapiens* sequences from UniProt and all cysteines as having undergone carbamidomethylation.
2. Use simple linear regression to obtain two equations. The first one (Model 1) takes all peptides in the theoretical digest into account and marks a central nominal mass for a given sum of amino acids' mass defects; this is the sloped line traversing all the plots in Figure 1. The second equation (Model 2), which is not shown explicitly in Figure 1, aims to model the nominal-mass standard deviation for a given sum of amino acids' mass defects; computing Model 2 considers the standard deviations of all peptides lying inside several horizontal strips in the plot, given by equally spaced values along the ordinate axis plus or minus a tolerance. Supplementary Information is available.
3. Use these two equations to estimate whether the monoisotopic mass of a given polypeptide is acceptable. Our default acceptance interval considers three standard deviations (Model 2) to each side of the central mark (Model 1).

Our software can use this (built-in) model as well as custom models that its model builder can generate seamlessly, to remove incompatible monoisotopic candidates with masses lower than 10 kDa. Like its previous version, it can also reprocess tandem mass spectra to spot those that have multiple precursors within the isolation window and thus enable the identification of multiplexed spectra in software environments such as PatternLab V (PLV) (Santos *et al.*, 2022).

We performed the following benchmark tests:

1. We searched our meningioma datafile with PLV according to its bioinformatics protocol (Santos *et al.*, 2022) and accepting up to 1% false-discovery rate and then its reprocessed tandem mass spectra using Y.A.D.A. 3.0.
2. We used Y.A.D.A. 1.0, FLASHDeconv and Y.A.D.A. 3.0, with default parameters, to deconvolute all MS1 spectra from our meningioma dataset, count the number of de-isotoped and de-charged peaks, and count how many did not comply with our mass defect model. These can be automatically removed, and moreover, we argue that they provide an indirect measure of deconvolution quality.

Table 1. Benchmark results of deconvolution algorithms

Software	Monoisotopic signals	Time (s)	Mass defect[a]
Y.A.D.A. 1.0	2 816 882	673	350 415
FLASHDeconv	2 307 360	465	72 167
Y.A.D.A. 3.0	2 499 281	70	60 038

[a]Number of mass spectral peaks not complying with our mass defect model.

3 Results

The search results with PLV on the Thermo file yielded 8791 identified spectra (6951 peptides) and the Y.A.D.A. 3.0 reprocessed file resulted in 9656 identified spectra (7344 peptides). An example of a multiplexed spectra that was identified is available (Supplementary Fig. S1). This 9.8% increase in spectral identifications was expected and is in accordance with our group's previous publication (Carvalho *et al.*, 2009). Y.A.D.A. 3.0 flagged 13 of the 9656 identifications as not complying with the mass defect model. However, given that our model considers three standard deviations, these 13 identifications may be taken as exceptions and be accepted. Table 1 shows a benchmarking of our software against its previous version (Y.A.D.A 1.0) and FLASHDeconv.

Some naturally occurring peptides will violate our model; for example, those with missed cleavages and rich in I, L, K or V are more prone to have higher mass defects. The same can be said for large peptides or intact proteins, both of which have a wide range of possible mass defects. Post-translational modifications may also influence the mass defect. Our software allows for generating customized models by providing a list of peptides for training; post-translational modifications mass differences can be indicated beside each amino acid within brackets (e.g. M[15.9949]).

In summary, we presented a mass defect metric that can remove peaks with unexpected masses built in a fast deconvolution algorithm. Further details are available on the project's website.

Funding

Fiocruz, Fundação Araucária, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho nacional de desenvolvimento científico e tecnológico (CNPq), Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) [2014/17264-3 and 2014/50867-3], the National Institute of Science and Technology in Bioanalitics, and FAPERJ provided financial aid.

Conflict of Interest: none declared.

References

Carvalho,P.C. *et al.* (2009) YADA: a tool for taking the most out of high-resolution spectra. *Bioinformatics*, **25**, 2734–2736.
Jeong,K. *et al.* (2020) FLASHDeconv: ultrafast, High-Quality feature deconvolution for top-down proteomics. *Cell Syst.*, **10**, 213–218.e6.
Liu,X. *et al.* (2010) Deconvolution and database search of complex tandem mass spectra of intact proteins. *Mol. Cell. Proteomics*, **9**, 2772–2782.
Santos,M.D.M. *et al.* (2022) Simple, efficient and thorough shotgun proteomic analysis with PatternLab V. *Nat. Protoc.*, **17**, 1553–1578.
Silva,J.M. *et al.* (2020) Proteomics pinpoints alterations in grade I meningiomas of male versus female patients. *Sci. Rep.*, **10**, 10335.