Journal of Proteomics 277 (2023) 104853

Contents lists available at ScienceDirect

Journal of Proteomics

journal homepage: www.elsevier.com/locate/jprot

Technical Note

DiagnoMass: A proteomics hub for pinpointing discriminative spectral clusters

Marlon D.M. Santos[a], Amanda C. Camillo-Andrade[a], Diogo B. Lima[b], Tatiana A.C.B. Souza[a], Juliana de S. da G. Fischer[a], Richard H. Valente[c], Fabio C. Gozzo[d], Valmir C. Barbosa[e], Carlos Batthyany[f], Julia Chamot-Rooke[g], Rosario Duran[f], Paulo C. Carvalho[a,*]

[a] Laboratory for Structural and Computational Proteomics, Carlos Chagas Institute, Fiocruz, Paraná, Brazil
[b] Department of Structural Biology, Leibniz - Forschungsinstitut für Molekulare Pharmakologie (FMP), Berlin, Germany
[c] Laboratory of Toxinology, Oswaldo Cruz Institute, Fiocruz – Rio de Janeiro, Brazil
[d] Universidade de Campinas, São Paulo, Brazil
[e] Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
[f] Analytical Biochemistry and Proteomics Unit, Institut Pasteur de Montevideo/IIBCE, Montevideo, Uruguay
[g] Institut Pasteur, University de Paris Cité, CNRS UAR 2024, Mass Spectrometry for Biology, F-75015 Paris

ARTICLE INFO

Keywords:
Protein identification
Biomarker
Proteomics

ABSTRACT

Motivation: There are several well-established paradigms for identifying and pinpointing discriminative peptides/proteins using shotgun proteomic data; examples are peptide-spectrum matching, *de novo* sequencing, open searches, and even hybrid approaches. Such an arsenal of complementary paradigms can provide deep data coverage, albeit some unidentified discriminative peptides remain.
Results: We present DiagnoMass, software tool that groups similar spectra into spectral clusters and then shortlists those clusters that are discriminative for biological conditions. DiagnoMass then communicates with proteomic tools to attempt the identification of such clusters. We demonstrate the effectiveness of DiagnoMass by analyzing proteomic data from *Escherichia coli, Salmonella,* and *Shigella*, listing many high-quality discriminative spectral clusters that had thus far remained unidentified by widely adopted proteomic tools. DiagnoMass can also classify proteomic profiles. We anticipate the use of DiagnoMass as a vital tool for pinpointing biomarkers.
Availability: DiagnoMass and related documentation, including a usage protocol, are available at http://www.diagnomass.com.

1. Introduction

One of the goals of proteomics is to identify and quantify as many peptides/proteins as possible to pinpoint those unique or differentially abundant to a biological condition. A variety of software tools rooted in different paradigms is available for analyzing proteomic tandem mass spectrometry data. PatternLab V [1], henceforth PLV, wraps Comet [2] to compare experimental spectra to those theoretically generated from a sequence database, following the classical and widely adopted peptide-spectrum matching approach, filtering results to achieve a 1% false-discovery rate (FDR) using SEPro [3]. Novor [4] employs *de novo* sequencing, a strategy that can infer sequences without a sequence database. FragPipe [5] can do an open search to facilitate inferring posttranslational modifications and mutations while performing spectral matching. All these paradigms have advantages and disadvantages and are complementary to one another. Their many qualities notwithstanding, discriminative spectra still elude them.

2. Results

We present DiagnoMass, a software tool that uses complete linkage hierarchical clustering (*hclust*) to cluster tandem mass spectra based on precursor tolerance, spectral angle, chromatography elution time, and charge state. In a previous study [6], we found *hclust* to be the most effective clustering approach for this task. Before clustering, DiagnoMass removes tandem mass spectra having more than 90% of their ion current on a single isotopic envelope. Our software saves the spectral clusters on an SQLite database, henceforth referred to as a knowledge

[*] Corresponding author.
E-mail address: carvalho.paulo@fiocruz.br (P.C. Carvalho).

https://doi.org/10.1016/j.jprot.2023.104853
Received 7 February 2023; Accepted 14 February 2023
Available online 16 February 2023

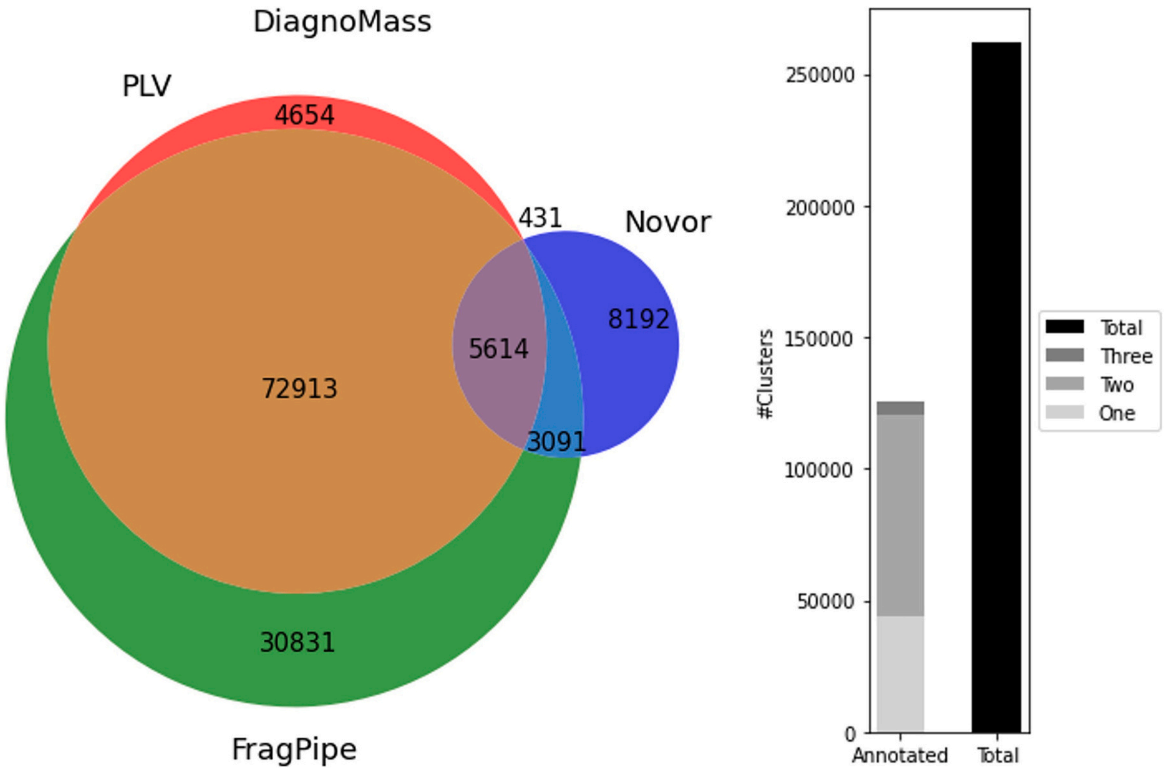

Fig. 1. A total of 261,664 spectral clusters were generated using default clustering parameters. Of these, 43,677, 76,435, and 5614 were annotated by 1, 2, and 3 tools, respectively.

base, to enable the seamless pinpointing of spectral clusters that are discriminative to one (or more) conditions as well as all further analyses. While DiagnoMass stems from our previous tool DiagnoProt [7], we note that it is a complete rewrite, including a redesigned graphical user interface, tools for viewing the data (Supplementary figs. 1 and 2), and significantly improved performance (almost two orders of magnitude faster for the data set used in this study). Most importantly, only DiagnoMass can interface with widely adopted proteomic identification tools, provide several data visualization strategies, and classify unknown proteomic profiles. Such interfacing enables our software to work as a proteomic hub, and annotate spectral clusters. As such, it makes available an unbiased approach for comparing biological conditions, waiving identification by a search engine as a first step. This proves convenient if we consider that, in general, the most interesting biological alterations are not found in reference sequences of public databases, being therefore overlooked by current standard procedures. In fact, according to [8], 75% of the spectra deposited in PRIDE are not identified.

DiagnoMass is moreover capable of spotlighting high-quality mass spectra that were missed by widely adopted tools based on complementary spectral identification paradigms. In this regard, DiagnoMass allows users to resort to specific methods (or even manual interpretation) to focus on discriminative and biologically relevant spectra. DiagnoMass offers several data analysis options - such as PCA, t-SNE (Supplementary fig. 1), and heat maps - to analyze the data set and help interpret biological patterns.

We demonstrate the effectiveness of DiagnoMass in shortlisting discriminative clusters not identified by widely adopted proteomic tools, focusing on a shotgun proteomic data set generated from *Escherichia coli, Salmonella,* and *Shigella* [9]. Our motivation for choosing these bacteria was that most MALDI-profiling diagnostic solutions experience difficulty discriminating among them [10]. Briefly, our data set contains three biological replicates of each bacterium, each having three technical replicates, adding up to 27 raw files acquired on a Q-Exactive Plus. Our data set was clustered with default parameters and automatically annotated while respecting the following minimum quality scores: 90 or higher for Novor, an XCorr greater than 2.0 for PLV, and a Peptide Prophet Probability of 0.95 or higher for FragPipe. The overlap of identifications provided by these proteomic tools is summarized in Fig. 1. DiagnoProt took ~3115 min to process the data, while DiagnoMass was 84 times faster (37 min) for knowledge base generation, which requires several computations other than spectral clustering.

DiagnoMass shortlisted 2751, 14,149, and 1519 spectral clusters exclusively associated with *E. coli*, *Salmonella*, and *Shigella,* respectively; each cluster was detected in all biological replicates (Supplementary fig. 2). Of these clusters, 1022 (*E. coli*), 6660 (*Salmonella*), and 603 (*Shigella*) were identified by at least one proteomic tool. Examples of high-quality discriminant spectra (according to our visual assessment) that had not yet been identified by some tool are shown in supplementary figs. 3, 4, and 5 for each biological condition.

DiagnoMass also provides new functionalities for classifying unknown proteomic profiles according to one of the biological conditions cataloged in its knowledge base. Such a feature can be used to obtain diagnostics based on mass spectrometry data, e.g., classifying bacteria. Classification occurs by scoring each spectrum relatively to the knowledge base. The scoring function we use returns a value $s(b)$ for each spectrum and each of the b clusters in which it is found, each cluster for a different biological condition. As explained in the online supplementary material, $s(b)$ is a fraction of S, the spectrum's cosine score relative to the consensus spectrum of the cluster. It decreases from $s(1) = (1 - \epsilon)S$ to $s(B) = \epsilon S$, where B is the total number of biological conditions and $\epsilon < 0.5$ is a positive stringency parameter (Supplementary fig. 6). The sum of $s(b)$ regarding all spectra in all clusters for a given biological condition provides its final classification score. DiagnoMass correctly classified all 9 proteomic profiles in the knowledge base using the leave-one-out cross-validation.

Funding

This work received partial financial support from CNPq, Inova Fiocruz, Fiocruz PEP, Fundação Araucária, CAPES, and a FAPERJ BBP grant.

Declaration of Competing Interest

None declared.

Data availability

ProteomeXchange identifier PXD035961.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jprot.2023.104853.

References

[1] M.D.M. Santos, D.B. Lima, J.S.G. Fischer, M.A. Clasen, L.U. Kurt, A.C. Camillo-Andrade, L.C. Monteiro, P.F. de Aquino, A.G.C. Neves-Ferreira, R.H. Valente, M.R. O. Trugilho, G.V.F. Brunoro, T.A.C.B. Souza, R.M. Santos, M. Batista, F.C. Gozzo, R. Durán, J.R. Yates, V.C. Barbosa, P.C. Carvalho, Simple, efficient and thorough shotgun proteomic analysis with PatternLab V, Nat. Protoc. (2022), https://doi.org/10.1038/s41596-022-00690-x.

[2] J.K. Eng, M.R. Hoopmann, T.A. Jahan, J.D. Egertson, W.S. Noble, M.J. MacCoss, A deeper look into Comet–implementation and features, J. Am. Soc. Mass Spectrom. 26 (2015) 1865–1874, https://doi.org/10.1007/s13361-015-1179-x.

[3] P.C. Carvalho, J.S.G. Fischer, T. Xu, D. Cociorva, T.S. Balbuena, R.H. Valente, J. Perales, J.R. Yates 3rd, V.C. Barbosa, Search engine processor: filtering and organizing peptide spectrum matches, Proteomics. 12 (2012) 944–949, https://doi.org/10.1002/pmic.201100529.

[4] B. Ma, Novor: real-time peptide de novo sequencing software, J. Am. Soc. Mass Spectrom. 26 (2015) 1885–1894, https://doi.org/10.1007/s13361-015-1204-0.

[5] A.T. Kong, F.V. Leprevost, D.M. Avtonomov, D. Mellacheruvu, A.I. Nesvizhskii, MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics, Nat. Methods 14 (2017) 513–520, https://doi.org/10.1038/nmeth.4256.

[6] A.R.F. Silva, D.B. Lima, L.U. Kurt, M. Dupré, J. Chamot-Rooke, M.D.M. Santos, C. A. Nicolau, R.H. Valente, V.C. Barbosa, P.C. Carvalho, Leveraging the Partition Selection Bias to Achieve a High-Quality Clustering of Mass Spectra [Manuscript submitted for publication], 2021.

[7] A.R.F. Silva, D.B. Lima, A. Leyva, R. Duran, C. Batthyany, P.F. Aquino, J.C. Leal, J. E. Rodriguez, G.B. Domont, M.D.M. Santos, J. Chamot-Rooke, V.C. Barbosa, P. C. Carvalho, DiagnoProt: a tool for discovery of new molecules by mass spectrometry, Bioinformatics. 33 (2017) 1883–1885, https://doi.org/10.1093/bioinformatics/btx093.

[8] J. Griss, Y. Perez-Riverol, S. Lewis, D.L. Tabb, J.A. Dianes, N. Del-Toro, M. Rurik, M.W. Walzer, O. Kohlbacher, H. Hermjakob, R. Wang, J.A. Vizcaíno, Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets, Nat. Methods 13 (2016) 651–656, https://doi.org/10.1038/nmeth.3902.

[9] M. Dupré, M. Duchateau, C. Malosse, D. Borges-Lima, V. Calvaresi, I. Podglajen, D. Clermont, M. Rey, J. Chamot-Rooke, Optimization of a top-down proteomics platform for closely related pathogenic bacterial discrimination, J. Proteome Res. 20 (2021) 202–211, https://doi.org/10.1021/acs.jproteome.0c00351.

[10] D. Borges Lima, M. Dupré, M.D. Mariano Santos, P.C. Carvalho, J. Chamot-Rooke, DiagnoTop: a computational pipeline for discriminating bacterial pathogens without database search, J. Am. Soc. Mass Spectrom. 32 (2021) 1295–1299, https://doi.org/10.1021/jasms.1c00014.