JOURNAL OF PROTEOMICS 89 (2013) 179–182

Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot

Technical note

Pinpointing differentially expressed domains in complex protein mixtures with the cloud service of PatternLab for Proteomics

F.V. Leprevost[a], D.B. Lima[a], J. Crestani[b], Y. Perez-Riverol[c,d], N. Zanchin[a], V.C. Barbosa[e], P.C. Carvalho[a,*]

[a]Laboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, Fiocruz, Paraná, Brazil
[b]Laboratory for Regulation of Gene Expression in Microorganisms, Chemistry Institute, University of São Paulo, São Paulo, Brazil
[c]Department of Proteomics, Center for Genetic Engineering and Biotechnology, Ave 31 e/158 y 190, Cubanacán, Playa, Ciudad de la Habana, Cuba
[d]EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
[e]Systems Engineering and Computer Science Program, COPPE, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

ARTICLE INFO

Article history:
Received 20 April 2013
Accepted 13 June 2013
Available online 21 June 2013

Keywords:
Computational proteomics
Bioinformatics
Proteomics
Protein domains
Functional analysis

ABSTRACT

Mass-spectrometry-based shotgun proteomics has become a widespread technology for analyzing complex protein mixtures. Here we describe a new module integrated into PatternLab for Proteomics that allows the pinpointing of differentially expressed domains. This is accomplished by inferring functional domains through our cloud service, using HMMER3 and Pfam remotely, and then mapping the quantitation values into domains for downstream analysis. In all, spotting which functional domains are changing when comparing biological states serves as a complementary approach to facilitate the understanding of a system's biology. We exemplify the new module's use by reanalyzing a previously published MudPIT dataset of *Cryptococcus gattii* cultivated under iron-depleted and replete conditions. We show how the differential analysis of functional domains can facilitate the interpretation of proteomic data by providing further valuable insight.

1. Introduction

One of the goals of shotgun proteomics is to provide in-depth, holistic insights into cellular biology by first pinpointing differentially expressed proteins when comparing physiological states. Inferring exactly which proteins are in a mixture is an extremely challenging task, especially when analyzing data from higher-order organisms, in which case the number of peptides shared among proteins increases rapidly [1]. Typically, peptide spectrum matches (PSMs), the building blocks of computational shotgun proteomics, are mapped into protein groups that share identified peptides (Supplementary Fig. 1). In more complex scenarios, proteins share subsets of peptides, with different proteins giving rise to complex dependency graphs for treatment by protein-inference algorithms. The problem of deciding which proteins are truly in the mixture has been very well characterized by Nesvizhskii and collaborators [2,3], and computational solutions to tackle

* *Corresponding author at:* Laboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, Fiocruz, Rua Prof. Algacyr Munhoz Mader, 3775, ZIP: 81350-010, City: Curitiba, Paraná, Brazil. Tel.: +55 41 3316 3230; fax: +55 41 3316 3267.
E-mail address: paulo@pcarvalho.com (P.C. Carvalho).

1874-3919/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jprot.2013.06.013

it have been proposed [4,5]. It is now consensual in the proteomics community that a maximum-parsimony list of proteins is to be reported; i.e., one reports the smallest subset of proteins that explains all identified peptides. As a result, proteomic experiments cannot in general determine a mixture's correct protein contents. For example, a typical result for a *Homo sapiens* analysis might report somewhere near 2000 proteins according to the maximum-parsimony criterion but about 4000 when considering redundancies.

These limitations can, to some extent, obfuscate downstream functional analysis. In an attempt to help circumvent such difficulties, we introduce a data-analysis strategy stemming from the fact that proteins are composed of one or more regions that establish their biochemical functions. These "building blocks", known as functional domains, tend to be strongly conserved in nature. As such, inferring functional domains at a large scale and mapping peptide identifications into them, instead of into proteins, provides key benefits such as: a) simplifying (and eliminating redundancies in) the process of gaining functional insight into the biological system at hand; b) specifically addressing differentially expressed domains, and thus the key functional content of a given biological sample; c) providing direct access to the current functional state, thereby helping drive biochemical conclusions and culminating in an easier way to understand the relevant biochemical mechanisms.

PatternLab for Proteomics is a one-stop shop for proteomic data analysis, providing tools to handle data from mass spectra, conduct differential proteomics analyses, and more [6–8]. Within this environment, a cloud service has now been implemented that infers domains from the FASTA sequences of all proteins identified in the experiment at hand and maps peptide quantitation values into the corresponding functional domains. Recently, the growth of MS/MS data has motivated the proteomics community to seek cloud computing tools to enable small laboratories to analyze complex datasets [9,10]. Because inferring domains at a large scale is computationally intensive, resource demanding, and requires installing specialized software and databases, all this functionality is already integrated into PatternLab following a cloud-client model. In essence, FASTA sequences of identified proteins are transmitted to our cloud servers, which perform domain inference by executing HMMER3 [11] on demand. The latter, briefly, uses a hidden Markov model (HMM) approach to scan profiles against the Protein Family (Pfam) database. Detailed instructions on how to use PatternLab are available [7,8]. The new option for performing a differential proteomic domain analysis (DPDA), or simply "differential dominomics", is integrated into the Regrouper module of the PatternLab pipeline and can be used by simply choosing to map values into domains instead of proteins. A brief overview of the pipeline is presented in Fig. 1.

We note that our cloud service, termed FioCloud, is hosted at the Fiocruz foundation (http://fiocruz.br), more specifically at Fiocruz Paraná. Fiocruz is one of the world's largest governmental agencies devoted to public health and research.

Our method significantly simplifies the process of understanding an organism's biology; therefore, some sensitivity loss may occur as a side effect. For example, for a group of proteins that share a common domain it is possible that while some are up-regulated others are down-regulated. Whenever this happens, dispersion is generated in the quantitation values and the proteins in question, consequently, are missed by our differential domain approach. Moreover, our method is blind to high-quality PSMs that do not map into any functional domain. On the other hand, we advocate that a differential domain prediction strategy can pave the way to a more effective analysis of organisms with poorly annotated genomes or when performing homology-driven proteomics [17], since protein domains tend to be more conserved than full-length protein sequences and are easier to predict [18]. That is, the method described herein is to be regarded as complementary to others, such as those based on homologous sequences. Note also that our method is blind to proteoforms, so these should be accounted for in standard differential proteomics analyses. Moreover, while a complementary strategy for inferring functional biology is by

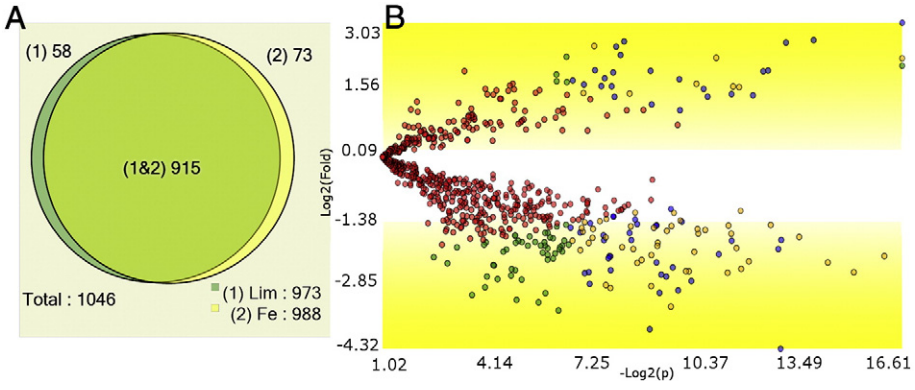

Fig. 1 – Workflow. Biological samples are analyzed by mass spectrometry. Mass spectra must be analyzed from a .sqt-compliant output [12], such as that generated by SEQUEST [13], ProLuCID [14], or the Spectrum Identification Machine (SIM) [15]. The PSMs are statistically filtered and organized using the Search Engine Processor (SEPro) [16]. The SEPro files are combined into PatternLab's index and sparse-matrix files for a single experiment using Regrouper, which offers the user the possibility of mapping quantitation values into protein functional domains. Differentially expressed domains are pinpointed using PatternLab's differential analyzer.

Fig. 2 – Differentially expressed domains. 58 and 73 domains were exclusively identified in the iron-depleted (Lim) and replete (Fe) conditions ($p < 0.05$), respectively, and 59 domains (blue dots) were found in both conditions but marked as differentially expressed ($q < 0.01$) by the TFold analysis.

mapping quantitation values into Gene Ontology (GO) terms (in fact, PatternLab includes specialized tools tailored to doing this [19]), usually an effective GO analysis can only be easily accomplished on well annotated organisms.

We exemplify the use of the new PatternLab functionality by reanalyzing the data of Crestani and collaborators, who have compared proteins from *Cryptococcus gattii* cultivated under iron-depleted and replete conditions [20]. The data were acquired using MudPIT [21] and searched with SEQUEST [13]. We then filtered for significant hits using the default parameters of SEPro [16] and used Regrouper for automatically inferring domains and mapping spectral counts into them by relying on HMMER3 and Pfam-A over the cloud, with Pfam accepting domains with HMMER E-Value < 10E-6 and i-EValue < 10E-3. Finally, PatternLab's statistical Venn Diagram [22] and TFold [23] modules were used for pinpointing domains exclusively identified in a single condition, as well as those differentially expressed (Fig. 2). Each domain out of a total of 1303 had at least one peptide mapped to it, and similarly none of 340 domains had any peptides mapped to it. Following domain identification, the information on which proteins were mapped to each domain is included in the domains' descriptions located in the index file. Downstream analysis in PatternLab can retrieve this information. We point out that the domain search results revealed interesting aspects that had remained unnoticed in the differential expression analysis. Examples are a FeoB_N domain, related to iron transportation and over-expressed in the iron-depleted condition, and three mitochondria-related domains (Mito_carr, FAD_binding_3, and Cyt-b5), over-expressed in the iron-replete condition. The complete results of all differentially expressed domains and all SEPro files discriminating our identified proteins and peptides are available at http://proteomics.fiocruz.br/pcarvalho/dominomics. A detailed protocol on how to use the new feature is available in Supplementary Part II.

2. Availability

All PatternLab modules are available for download at http://proteomics.fiocruz.br. All results of our *C. gatti* analysis are available at http://proteomics.fiocruz.br/dpda.

Financial support

The authors acknowledge CNPq, CAPES, FAPERJ, FAPESP, Fundação Araucária, and Fiocruz–PDTIS for the financial support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jprot.2013.06.013.

REFERENCES

[1] Perez-Riverol Y, Sanchez A, Ramos Y, Schmidt A, Muller M, Betancourt L, et al. In silico analysis of accurate proteomics, complemented by selective isolation of peptides. J Proteomics Sep 6 2011;74(10):2071–82.
[2] Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods Oct 2007;4(10):787–97.
[3] Nesvizhskii AI, Aebersold R. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. Drug Discov Today Feb 15 2004;9(4):173–81.
[4] Zhang B, Chambers MC, Tabb DL. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. J Proteome Res Sep 2007;6(9):3549–57.
[5] Ma ZQ, Dasari S, Chambers MC, Litton MD, Sobecki SM, Zimmerman LJ, et al. IDPicker 2.0: improved protein assembly with high discrimination peptide identification filtering. J Proteome Res Aug 2009;8(8):3872–81.
[6] Carvalho PC, Fischer JS, Chen EI, Yates III JR, Barbosa VC. PatternLab for proteomics: a tool for differential shotgun proteomics. BMC Bioinformatics 2008;9:316.
[7] Carvalho PC, Yates Jr I, Barbosa VC. Analyzing shotgun proteomic data with PatternLab for proteomics. Curr Protoc Bioinformatics Jun 2010:13.13.1–5 [Chapter 13:Unit-13.13].
[8] Carvalho PC, Fischer JS, Xu T, Yates III JR, Barbosa VC. PatternLab: from mass spectra to label-free differential shotgun proteomics. Curr Protoc Bioinformatics Dec 2012:13.19.1–13.19.18 [Chapter 13:Unit13.19].
[9] Muth T, Peters J, Blackburn J, Rapp E, Martens L. ProteoCloud: A full-featured open source proteomics cloud computing pipeline. J Proteomics Aug 2013;88:104–8.

[10] Trudgian DC, Mirzaei H. Cloud CPFP: a shotgun proteomics data analysis pipeline using cloud and high performance computing. J Proteome Res Dec 7 2012;11(12):6282–90.
[11] Eddy SR. A new generation of homology search tools based on probabilistic inference. Genome Inform Oct 2009;23(1):205–11.
[12] McDonald WH, Tabb DL, Sadygov RG, MacCoss MJ, Venable J, Graumann J, et al. MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. Rapid Commun Mass Spectrom 2004;18(18):2162–8.
[13] Eng JK, McCormack L, Yates A, Yates III JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom 1994;5:976–89.
[14] Xu T, Venable JD, Park SK, Cociorva D, Lu B, Liao L, et al. ProLuCID, a fast and sensitive tandem mass spectra-based protein identification program. Mol Cell Proteomics 2006;5(S174).
[15] Borges D, Perez-Riverol Y, Nogueira FC, Domont GB, Noda J, Leprevost FD, et al. Effectively addressing complex proteomic search spaces with peptide spectrum matching. Bioinformatics 2013;29(10):1343–4.
[16] Carvalho PC, Fischer JS, Xu T, Cociorva D, Balbuena TS, Valente RH, et al. Search engine processor: filtering and organizing peptide spectrum matches. Proteomics Apr 2012;12(7):944–9.
[17] Shevchenko A, Sunyaev S, Loboda A, Shevchenko A, Bork P, Ens W, et al. Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. Anal Chem May 1 2001;73(9):1917–26.
[18] Junqueira M, Carvalho PC. Tools and challenges for diversity-driven proteomics in Brazil. Proteomics Aug 2012;12(17):2601–6.
[19] Carvalho PC, Fischer JS, Chen EI, Domont GB, Carvalho MG, Degrave WM, et al. GO Explorer: a gene-ontology tool to aid in the interpretation of shotgun proteomics data. Proteome Sci 2009;7:6.
[20] Crestani J, Carvalho PC, Han X, Seixas A, Broetto L, Fischer JS, et al. Proteomic profiling of the influence of iron availability on *Cryptococcus gattii*. J Proteome Res Jan 1 2012;11(1):189–205.
[21] Washburn MP, Wolters D, Yates III JR. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol Mar 2001;19(3):242–7.
[22] Carvalho PC, Fischer JS, Perales J, Yates JR, Barbosa VC, Bareinboim E. Analyzing marginal cases in differential shotgun proteomics. Bioinformatics Jan 15 2011;27(2):275–6.
[23] Carvalho PC, Yates III JR, Barbosa VC. Improving the TFold test for differential shotgun proteomics. Bioinformatics Jun 15 2012;28(12):1652–4.