BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 10 2013, pages 1343–1344
doi:10.1093/bioinformatics/btt106

Sequence analysis

Advance Access publication February 27, 2013

Effectively addressing complex proteomic search spaces with peptide spectrum matching

Diogo Borges[1,2,†], Yasset Perez-Riverol[3,4,†], Fabio C. S. Nogueira[5], Gilberto B. Domont[5], Jesus Noda[3], Felipe da Veiga Leprevost[2], Vladimir Besada[3], Felipe M. G. França[1], Valmir C. Barbosa[1], Aniel Sánchez[3] and Paulo C. Carvalho[2,*]

[1]Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro, 21941-972 Rio de Janeiro, Brazil, [2]Laboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, 81350-010 Fiocruz, Paraná, Brazil, [3]Department of Proteomics, Center for Genetic Engineering and Biotechnology, P.O. Box 6162, Cubanacán, Playa, Ciudad de la Habana, Cuba, [4]Proteomic Services, EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK and [5]Proteomics Unit, Institute of Chemistry, Federal University of Rio de Janeiro, 21941-909 Rio de Janeiro, Brazil

Associate Editor: Martin Bishop

ABSTRACT

Summary: Protein identification by mass spectrometry is commonly accomplished using a peptide sequence matching search algorithm, whose sensitivity varies inversely with the size of the sequence database and the number of post-translational modifications considered. We present the Spectrum Identification Machine, a peptide sequence matching tool that capitalizes on the high-intensity b1-fragment ion of tandem mass spectra of peptides coupled in solution with phenylisotiocyanate to confidently sequence the first amino acid and ultimately reduce the search space. We demonstrate that in complex search spaces, a gain of some 120% in sensitivity can be achieved.

Availability: All data generated and the software are freely available for academic use at http://proteomics.fiocruz.br/software/sim.

Contact: paulo@pcarvalho.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 22, 2012; revised on February 20, 2013; accepted on February 22, 2013

1 INTRODUCTION

One of the goals of shotgun proteomics is to perform large-scale identification and quantitation of thousands of proteins within complex protein mixtures (e.g. biological fluids or whole-cell lysates). The strategy comprises protein digestion, followed by peptide chromatographic separation online with tandem mass spectrometry (MS2) (Washburn *et al.*, 2001). The MS2 data are then generally identified using a peptide sequence matching (PSM) tool; examples are SEQUEST (Eng *et al.*, 1994) and, most recently, Andromeda (Cox *et al.*, 2011). Briefly, given a peptide's precursor ion mass and MS2, these algorithms pull out, from a peptide-sequence database, peptide sequences whose theoretical mass lies within a given tolerance from the experimental precursor mass. Following that, theoretical spectra are generated for all peptide candidates so that some similarity metric, be it empirical or statistical, can be used to select the most likely candidate. Finally, a list of identifications satisfying some false discovery rate (FDR) is obtained by using a statistical filtering tool such as SEPro (Carvalho *et al.*, 2012).

The sensitivity of a PSM tool varies inversely with the size of the sequence database and the number of post-translational modifications considered (Yen *et al.*, 2006). Consequently, studies addressing complex search spaces are challenging when seen from a computational perspective. Examples are analysing snake venoms for identifying naturally occurring peptides (Tashima *et al.*, 2012) or performing a meta-proteomic study of a microorganism biota (Muth *et al.*, 2012). The former requires not trypsinizing the samples and thus lifts the constraints of a PSM search engine to only tryptic peptides, which results in an exponential growth of the search space; the latter entails the concatenation of hundreds of sequence databases of different organisms. Nevertheless, the rewards at stake could be discovering a naturally occurring peptide with pharmaceutical properties or the in-depth comprehension of a system's biology.

Recently, Sanchez *et al.* (2010) and Perez-Riverol *et al.* (2011) demonstrated the possibility to identify peptides using the N-terminal residue and accurate precursor mass; for this, they coupled peptides in solution with phenylisotiocyanate (PITC). During the activation in the collision cell, these phenylthiocarbamoyl-derivatized peptides dissociate to specifically yield an intense b1 fragment. This unlocks the possibility to confidently determine the N-terminal residue in a single mass spectrum. The authors then demonstrated a peptide identification tool that considered only the b1 fragment ion mass and the high mass accuracy of the precursor and used it to identify peptides in an *Escherichia coli* tryptic digest. The shortcomings of this method are in the inability to discriminate between peptides with close masses and same first residue. As the remaining MS2 information is not taken into account, the method is blind to peptides not found in the database but also coinciding in mass and first residue, and thus prone to such false positives.

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

More on these limitations is found in a discussion in the Supplementary File. In brief, this strategy becomes inapplicable to studies addressing complex search spaces, where these 'coincidences' become increasingly frequent. Notwithstanding this, the authors demonstrated a way to potentially improve current PSM algorithms.

2 METHODS

To overcome these limitations, we present the Spectrum Identification Machine (SIM). SIM capitalizes on PITC-coupled peptides to reduce the search space by filtering peptide candidates to only those satisfying the precursor mass and the first amino acid obtained from the high-intensity b1 fragment. The reduced search space is then queried by comparing theoretically generated spectra to experimental ones with a similarity metric that is the dot product between the normalized experimental and theoretical spectra, multiplied by the number of matched peaks. This enables the selection of the highest-scoring candidate sequence. Some other scores, such as DeltaCN from SEQUEST, are also computed; in fact, the output of SIM is a .SQT file (i.e. it has the SEQUEST output format), which makes every tool that works with SEQUEST automatically compatible with SIM.

We benchmarked SIM, with results filtered by SEPro to achieve a 1% FDR (protein level), on a previously published yeast lysate MudPIT dataset (Barboza *et al.*, 2011) against the widely adopted Andromeda. We note that this is a non-PITC-labelled dataset; therefore, this benchmarking was carried out to verify whether SIM would perform acceptably. Search parameters and results are available at the SIM website. In our hands, Andromeda (v. 1.3.0.5) identified 53 997 MS/MS and SIM (v. 0.905) 73 639 MS/MS, both constrained by the same FDR of 1% at the protein level. This result demonstrates that SIM does have an effective algorithm for PSM and has allowed us to focus our efforts on showing the benefits of activating what we term the PITC logic.

We verify the efficiency of the PITC logic on a PITC-labelled *E.coli* extract that was trypsinized and analysed with a 1-h reversed-phase chromatography gradient on an Orbitrap Velos acquiring MS2 in higher energy collisional dissociation (HCD) mode. To verify how the increase in database complexity affected the results, we generated three peptide databases, one comprehending only fully tryptic peptides (one missed-cleavage accepted and no post-translational modifications (PTMs)), the second having a semi-tryptic specificity and the third with no enzymatic specificity. This generated search spaces comprising 566 070, 11 217 794 and 63 102 231 peptides, respectively. Results were filtered with SEPro to converge to a list of 1% FDR.

3 RESULTS

Search results with and without the PITC logic are presented in Figure 1. An example of a PITC peptide tandem mass spectrum is found in Supplementary Figure S1. Further comparisons to the method from Sanchez *et al.*, (2010) are provided in the Supplementary File.

4 DISCUSSION AND CONCLUSIONS

We have searched an *E.coli* tryptic digest labelled with PITC using SIM. We performed a proof of concept by testing the efficiency of our new PITC logic under increasing complexities, i.e. from tryptic to semi-tryptic to fully tryptic, and obtained an increase in sensitivity of some 120% in a large search space. As such, the SIM-PITC approach is recommended when addressing proteomic studies with complex search spaces. SIM has a graphical user interface to provide a user-friendly experience, is multiplatform and can be executed in cluster environments. SIM is integrated into PatternLab for proteomics (Carvalho *et al.*, 2008, 2010), which makes available an arsenal of tools for quantitative and differential proteomics.

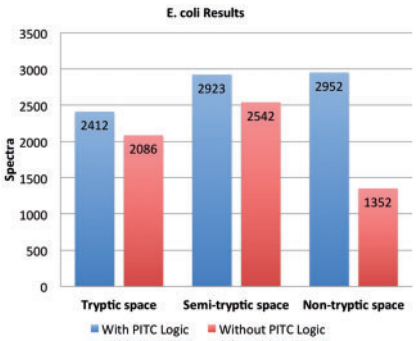

Fig. 1. Number of identified spectra with and without activating SIM's PITC logic

ACKNOWLEDGEMENTS

D.B. and Y.P.-R. have contributed equally to this work. The authors thank Dr Fabrico Marchini and Michel Batista for technical discussions.

Funding: Fundação de Amparo a Pesquisa do Rio de Janeiro (FAPERJ), Programa de Desenolvimento Tecnológico de Insumos para a Saúde (PDTIS), and Conselho Nacional de Pesquisa (CNPq).

Conflict of Interest: none declared.

REFERENCES

Barboza,R. *et al.* (2011) Can the false-discovery rate be misleading? *Proteomics*, **11**, 4105–4108.

Carvalho,P.C. *et al.* (2008) PatternLab for proteomics: a tool for differential shotgun proteomics. *BMC Bioinformatics*, **9**, 316.

Carvalho,P.C. *et al.* (2010) Analyzing shotgun proteomic data with PatternLab for proteomics. *Curr. Protoc. Bioinformatics*, Chapter 13, Unit 1–15.

Carvalho,P.C. *et al.* (2012) Search engine processor: filtering and organizing peptide spectrum matches. *Proteomics*, **12**, 944–949.

Cox,J. *et al.* (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome. Res.*, **10**, 1794–1805.

Eng,J.K. *et al.* (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom*, **5**, 976–989.

Muth,T. *et al.* (2012) Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. *Mol. Biosyst.*, **9**, 578–585.

Perez-Riverol,Y. *et al.* (2011) In silico analysis of accurate proteomics, complemented by selective isolation of peptides. *J. Proteomics*, **74**, 2071–2082.

Sanchez,A. *et al.* (2010) Evaluation of phenylthiocarbamoyl-derivatized peptides by electrospray ionization mass spectrometry: selective isolation and analysis of modified multiply charged peptides for liquid chromatography-tandem mass spectrometry experiments. *Anal. Chem.*, **82**, 8492–8501.

Tashima,A.K. *et al.* (2012) Peptidomics of three Bothrops snake venoms: insights into the molecular diversification of proteomes and peptidomes. *Mol. Cell Proteomics*, **11**, 1245–1262.

Washburn,M.P. *et al.* (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.*, **19**, 242–247.

Yen,C.Y. *et al.* (2006) Improving sensitivity in shotgun proteomics using a peptide-centric database with reduced complexity: protease cleavage and SCX elution rules from data mining of MS/MS spectra. *Anal. Chem.*, **78**, 1071–1084.