JOURNAL OF PROTEOMICS 129 (2015) 51–55

Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot

Technical note

SIM-XL: A powerful and user-friendly tool for peptide cross-linking analysis☆

Diogo B. Lima[a,*], *Tatiani B. de Lima*[b], *Tiago S. Balbuena*[c], *Ana Gisele C. Neves-Ferreira*[d], *Valmir C. Barbosa*[e], *Fábio C. Gozzo*[b,*], *Paulo C. Carvalho*[a,*]

[a]*Laboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, Fiocruz, Paraná, Brazil*
[b]*Dalton Mass Spectrometry Laboratory, University of Campinas, São Paulo, Brazil*
[c]*College of Agricultural and Veterinary Sciences, State University of São Paulo, Jaboticabal, São Paulo, Brazil*
[d]*Laboratory of Toxinology, Oswaldo Cruz Institute, Fiocruz, Rio de Janeiro, Brazil*
[e]*Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil*

ARTICLE INFO

Article history:
Received 17 December 2014
Accepted 21 January 2015
Available online 29 January 2015

Keywords:
Cross-linked
Cross-linking
Protein–protein

ABSTRACT

Chemical cross-linking has emerged as a powerful approach for the structural characterization of proteins and protein complexes. However, the correct identification of covalently linked (cross-linked or XL) peptides analyzed by tandem mass spectrometry is still an open challenge. Here we present SIM-XL, a software tool that can analyze data generated through commonly used cross-linkers (e.g., BS3/DSS). Our software introduces a new paradigm for search-space reduction, which ultimately accounts for its increase in speed and sensitivity. Moreover, our search engine is the first to capitalize on reporter ions for selecting tandem mass spectra derived from cross-linked peptides. It also makes available a 2D interaction map and a spectrum-annotation tool unmatched by any of its kind. We show SIM-XL to be more sensitive and faster than a competing tool when analyzing a data set obtained from the human HSP90. The software is freely available for academic use at http://patternlabforproteomics.org/sim-xl. A video demonstrating the tool is available at http://patternlabforproteomics.org/sim-xl/video. SIM-XL is the first tool to support XL data in the mzIdentML format; all data are thus available from the ProteomeXchange consortium (identifier PXD001677).
This article is part of a Special Issue entitled: Computational Proteomics.

Recently, chemical cross-linking coupled to high-resolution mass spectrometry (XL-MS) emerged as a powerful strategy to broaden the toolset for protein structural characterization and for determining protein–protein interactions. In this approach, the side chains of amino acids in proteins and/or their complexes are covalently linked by reactions with cross-linkers. After enzymatic digestion of the cross-linked protein(s), cross-linked peptides can be identified by tandem mass spectrometry, generating spatial constraints between amino acid residues. In other words, the distance between two interacting partners (e.g., amino acids of the same protein or different ones or even lipids, RNA, DNA, and carbohydrates) can be inferred through the establishment of the covalent bond, therefore allowing for low-resolution characterization.

☆ This article is part of a Special Issue entitled: Computational Proteomics.

* *Corresponding authors.*
E-mail addresses: diogobor@gmail.com (D.B. Lima), fabio@iqm.unicamp.br (F.C. Gozzo), paulo@pcarvalho.com (P.C. Carvalho).

http://dx.doi.org/10.1016/j.jprot.2015.01.013
1874-3919/

Ultimately, these distance restraints enable a variety of structural information to be obtained, unraveling important information for understanding protein folding, complex topology and interaction regions [1,2].

Although identifying unmodified peptides (i.e., linear peptides) by mass spectrometry is a rather solved problem in proteomics, reliably identifying pairs of covalently linked peptides (i.e., interpeptide / type 2 cross-links and intrapeptide / type 1 cross-links) is still a bottleneck. To date, there are only a few reference engines for XL search, most prominently Crux [3], CrossWork [4], StavroX[5], and pLink[6]. Differently than in classical proteomics, where proteins are identified from a large redundancy of peptides, generally allowing a 1% false-discovery rate (FDR), we argue that XL-MS studies should avoid false-positives at all costs, as the structural information brought by each cross-link is not only fundamental but also unique. We therefore advocate that the FDR control of classical proteomics is not sufficient in the context of cross-linking: besides having each identification associated with a stringent family-wise error rate estimate or empirically derived score, a personal assessment must be carried out. This is because a single wrong XL identification is enough to create a conflicting protein model or to incorrectly suggest an interaction between proteins. Yet the problem of identifying XL peptides is far more challenging than those of conventional proteomics, as the search space for cross-linked peptides grows quadratically with the number of peptides, which naturally decreases sensitivity and selectivity in the classical search engine approach [7]. Moreover, the population of XL spectra in an LC/MS/MS run is minute when compared to those originating from linear peptides and from type 0 cross-links (i.e., peptides containing dead-end modifications). Consequently, XL identification tools need to be very sensitive and selective to provide means for the user to easily interpret and verify each identification. In our hands, existing tools presented false-positives among their top-scoring hits and were computationally costly (data not shown). Additionally, they provided limited or no resources at all for viewing, editing, and manually validating XL peptide identifications, which is a fundamental and time-consuming step in any experiment addressing XL-MS.

Here we present the Spectrum Identification Machine for Cross-Linked Peptides (SIM-XL), a fast and sensitive XL search engine that is part of the PatternLab for proteomics environment [8]. The SIM-XL software was programmed in C# with .NET Framework 4.5. The software requires a computer with Windows 7 or later, and at least 4 GB of RAM. To perform a search, the user begins by specifying the tandem mass spectrum file, the FASTA sequence database file, as well as parameters related to sample preparation (e.g., modifications) and mass spectrometry features (e.g., ppm). A detailed description of each SIM-XL parameter is available in its integrated manual, which is accessible through the Help menu, Read Me option. The current version is compatible with the Mascot Generic Format [9], MS2 [10], and mzML[11] and can work directly with Thermo .RAW files if the freely available MSFileReader is locally installed.

Among its novelties, we point out three: (I) SIM-XL builds on a new paradigm for search-space reduction. As previously mentioned, the larger the search space (i.e., the set of possibilities of theoretical peptides or, in this case, combinations of peptides originating from a database matching the experimental precursor mass), the lower the sensitivity of the search engine [12]. To address the quadratic growth arising from cross-linked peptide candidates, our search engine employs a dynamic database reduction heuristic to eliminate possibilities by considering only combinations that contain at least one linear peptide identified with a dead-end modification. (II) SIM-XL search engine takes advantage of reporter ions [13], i.e., fingerprints of mass spectral peaks found almost exclusively in tandem mass spectra derived from cross-linked peptides. By searching only tandem mass spectra with these reporter ions, the chance of false-positive identifications is decreased and the search speed increases considerably. We note that feature I and II are optional and therefore can be switched on or off. (III) Our search engine provides a user-friendly Graphical User Interface (GUI) that allows the user to assess each identification interactively through a spectrum viewer and annotation tool.

As described, SIM-XL can reduce the search space when working in dynamic database reduction mode. To do this, it begins by wrapping the Comet [14] search engine to perform a preliminary search aiming to identify peptide spectrum matches (PSMs) of linear peptides with a user-configurable XCorr cutoff (default value: 1.5). A secondary database is then dynamically generated with all possible pairs containing a linear peptide with a dead-end and a peptide from the identified proteins having a reactive site on the sequence. When this mode is not activated, by contrast, all pairs of peptides containing a reactive site in the sequence database are considered. SIM-XL makes use of reporter ions by only considering tandem mass spectra that contain reporter ions from cross-linked peptides [13]. The idea of using reporter ions for different strategies has been previously reported [7,15–17]. This significantly decreases the number of spectra to be searched and thus improves on both selectivity and processing time, especially on large data sets. As previously reported, these so-called reporter fragment ions are specific to Lys-Lys cross-linked or dead-end modified peptides and consist of a rearranged lysine side chain and the spacer arm of the linker. SIM-XL can work with any set of diagnostic ions that are specified in its GUI or XML parameter file. Yet we note that to take advantage of reporter-ion filtering, MS/MS acquisition should start at least at *m/z* of the lowest mass reporter ion (in the case of DSS/BS3, m/z 220). Although this is usually not a problem for TOF instruments, special attention should be paid when acquiring data using Orbitrap analyzers, as the *m/z* range is more restricted. For cases such as these, mass spectra from the same precursor can optionally be acquired in different ranges of *m/z* and our tool will automatically generate consensus (merged) spectra.

SIM-XL uses multi-threading to take advantage of multiple hardware cores and therefore significantly increase the search speed. These speedups become evident especially in those cases in which (i) no dead-end modifications are specified, so the software has to work on the full search space; (ii) MS/MS acquisition does not contain reporter ions, so no reporter-ion spectrum filtering is possible; or (iii) the number of proteins in the database is large, so the search space is huge, even using the two filtering options described previously.

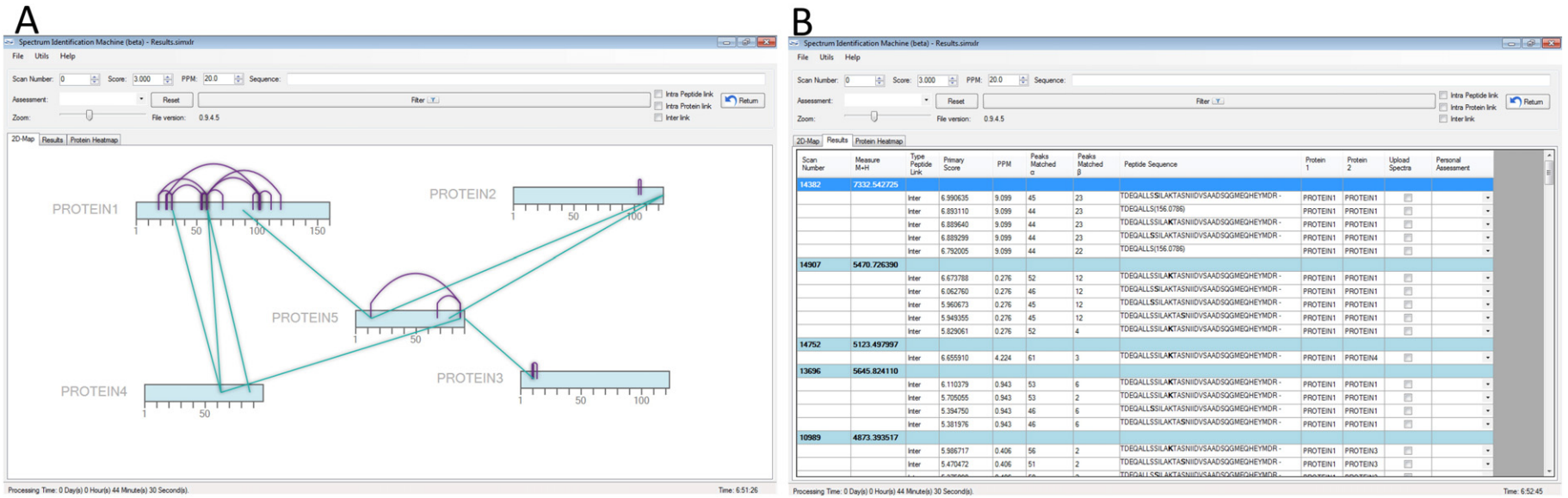

Fig. 1 – Panel A shows the 2D interaction map for a data set from a protein complex consisting of five proteins (data undisclosed). By clicking on the Results tab, the dynamic report is displayed. Mass spectra can be accessed by either clicking on the 2D-map link or on any dynamic report result. Dynamic cutoff scores can be applied and combined with personal assessments (i.e., excellent, good, medium, fair, or poor).

Once the search engine finishes, SIM-XL presents its results in three interconnected modes: the 2D interaction map and the dynamic report (Fig. 2) and a "Heat map" of a pairwise comparison (not shown). The former provides a graphical representation of all XL links among the protein(s) and the latter allows the user to sort identifications by several criteria (e.g., primary score and ppm) and to make an assessment for each spectrum. This assessment can be saved in SIM-XL's dynamic report to help in keeping track of which spectra were already evaluated and approved to be considered, say, when determining a protein's structure or inferring protein–protein interactions. Fig. 1 shows a screenshot of SIM-XL's main GUI exemplifying its 2D interaction map and the dynamic report.

We note than in the dynamic report, buttons are made available to allow for the upload of high-quality annotated XL-MS spectra to the online database we are developing to support the creation of even more effective machine learning approaches for identifying cross-linked peptide species originating from different cross-linkers, mass spectrometers, etc. When a spectrum is uploaded, only information pertaining to that single spectrum, including which peptides were cross-linked, is sent to our server. As the number of XL mass spectra is generally low for an experiment, we advocate that libraries such as in this initiative can become fundamental for the development of future, more sensitive tools.

SIM-XL's spectrum viewer and 2D interaction map are its high points, constituting unique features that greatly simplify the assessment of identification candidates, each of which can be easily visualized by double-clicking on the identification provided in the dynamic search engine report or in the graphical representation in the 2D map. The spectrum viewer allows the user to view the annotated ions and to zoom in on a region of interest in the mass spectrum. Its importance to a SIM-XL user resides in that it allows for the manual validation of all assignments given by the software and, importantly, for the easy verification of other assignment possibilities for the same mass spectrum, thus supporting unbiased judgments, independent of SIM-XL's scoring heuristic through an immediate comparison assisted by SIM-XL's theoretical spectrum predictor. We strongly encourage viewing further details and functionalities in our online supplementary video available at http://patternlabforproteomics.org/sim-xl/video.

We demonstrate the effectiveness of SIM-XL by analyzing a data set aiming to aid in establishing a structural model for the Human HSP90; the data set was generated as previously described [13]. Briefly, disuccinimidyl suberate (DSS) cross-linker was dissolved in dimethylformamide (DMF, Thermo Scientific) at a stock concentration of 27.1 mM. DSS was added to the human C-terminal of HSP90 at a 1:50 (protein: DSS) ratio and incubated with the sample for 2 h at room temperature. Cross-linking reaction was quenched with ammonium bicarbonate 100 mM. Reduction and alkylation of cysteine residues were performed using dithiothreitol and iodacetamide during 30 min at 60 °C and at room temperature, respectively. The sample was digested with trypsin (Promega) at 1:50 for 16 h at 37 °C. The peptides were fractionated using an Oasis HLB cartridge (Waters Corp.) and eluted with different concentration of acetonitrile and analyses were performed using a Thermo Q-Exactive mass spectrometer equipped with a nano-electrospray source coupled to a nano EasyLC (Thermo, San Jose – CA).

The search engines used were Crux v. 2.0 and SIM-XL 1.0. All searches were performed using carbamidomethylation of cysteine as fixed modification; for SIM-XL, the variable modifications were a dead-end DSS of 156.0786 Da and a DSS cross-linker mass modification of 138.0681 Da; the remaining parameters were defaults. The precursor and fragment ion-mass tolerances were of 20 ppm. The sequence database comprised the sequence of HSP90 plus those from five decoy sequences. Benchmarking was performed on a MacPro with Intel Xeon X5670 processors.

The searching times were 1 h 4 min 10 s, 1 h 5 min, 1 min 49 s, and 37 s, respectively, for Crux, SIM-XL in normal mode (i.e., with features I and II off), SIM-XL with dynamic database reduction activated, and SIM-XL with both dynamic database reduction and the use of reporter ions activated. Plots of the cumulative number of non-decoy hits among the 50 top-scoring spectra for these searches are found in Fig. 2.

Our data set consisted of 1,788 tandem mass spectra, of which 973 contained at least one XL reporter ion. Among the top-50 mass spectra reported by SIM-XL with both dynamic database reduction and reporter ion modes turned off, three XL originated from an HSP90-decoy peptide pair, two of which presented reporter ions; in this case, an HSP90 peptide could actually be there, but having its counterpart wrongly attributed. As for the remaining non-decoy identifications, all but four did not have at least one reporter-ion peak. We also note that

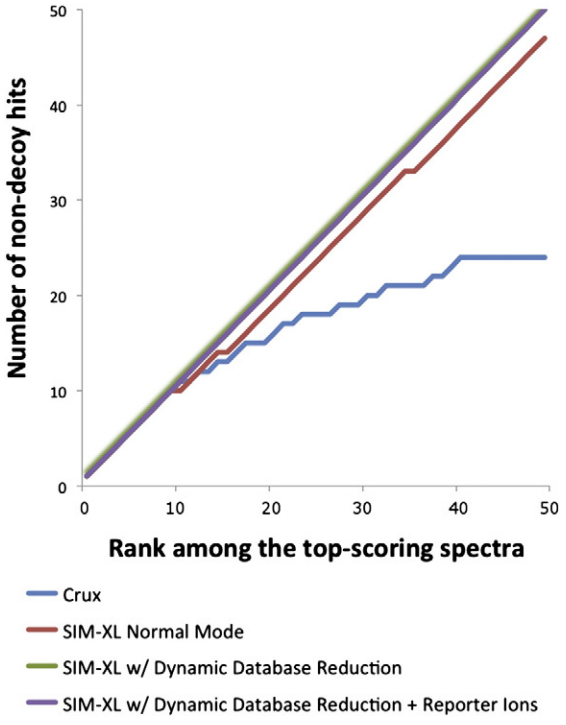

Fig. 2 – Plots of the cumulative number of non-decoy hits for SIM-XL operating in different modes and for Crux, considering in all cases the 50 top-scoring mass spectra. The "SIM-XL w/ Dynamic Database Reduction" and "SIM-XL w/ Dynamic Database Reduction + Reporter Ions" lines coincide.

among all non-decoy identifications appearing in normal mode, at least one of the cross-linked peptides had their linear peptide version identified with a dead-end. Two of the three HSP90-decoy duets did not have dead-end counterparts.

We recommend using SIM-XL with both the dynamic database reduction and the reporter ion modes activated. These can drastically decrease the chances of a false-positive and significantly increase search speed; for the task at hand, this resulted in reducing the search time from 1 h 5 min to 37 s. We believe these two features underscore SIM-XL as a promising tool for addressing next generation challenges such as *in vivo* cross-linking [19].

Finally, we note that SIM-XL is the first XL tool capable of exporting results in the forthcoming mzIdentML 1.2 format [18], established by the Proteomics Standards Initiative (PSI). This has enabled us to perform the first complete submission of an XL data set to the ProteomeXchange consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository [20] (data set identifier PXD001677, DOI 10.6019/PXD001677). Consequently, all our data are readily available to the scientific community. The remaining files, which include the search results, parameter files, and the sequence database, are available at the project's website (http://patternlabforproteomics.org/sim-xl).

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

The authors thank FAPESP, FAPERJ, CAPES, Universal CNPq, Microsoft Research—Microsoft Azure Research Award, Programa Estratégico de Apoio à Pesquisa em Saúde (PAPES), and Fundação Oswaldo Cruz for financial support. We thank the PRIDE Team for working together with us to enable SIM-XL to support the next version of mzIdentML. The authors declare no competing financial interest.

REFERENCES

[1] Preston GW, Radford SE, Ashcroft AE, Wilson AJ. Covalent cross-linking within supramolecular peptide structures. Anal Chem Aug. 2012;84(15):6790–7.
[2] Merkley ED, Cort JR, Adkins JN. Cross-linking and mass spectrometry methodologies to facilitate structural biology: finding a path through the maze. J Struct Funct Genomics Sep. 2013;14(3):77–90.
[3] McIlwain S, Draghicescu P, Singh P, Goodlett DR, Noble WS. Detecting cross-linked peptides by searching against a database of cross-linked peptide pairs. J Proteome Res May 2010;9(5):2488–95.
[4] Rasmussen MI, Refsgaard JC, Peng L, Houen G, Højrup P. CrossWork: software-assisted identification of cross-linked peptides. J Proteomics Sep. 2011;74(10):1871–83.
[5] Götze M, Pettelkau J, Schaks S, Bosse K, Ihling CH, Krauth F, et al. StavroX—a software for analyzing crosslinked products in protein interaction studies. J Am Soc Mass Spectrom Jan. 2012;23(1):76–87.
[6] Yang B, Wu Y-J, Zhu M, Fan S-B, Lin J, Zhang K, et al. Identification of cross-linked peptides from complex samples. Nat Methods Jul. 2012;9(9):904–6.
[7] Borges D, Perez-Riverol Y, Nogueira FCS, Domont GB, Noda J, da Veiga Leprevost F, et al. Effectively addressing complex proteomic search spaces with peptide spectrum matching. Bioinforma Oxf Engl May 2013;29(10):1343–4.
[8] Carvalho PC, Fischer JSG, Xu T, Yates III JR, Barbosa VC. PatternLab: from mass spectra to label-free differential shotgun proteomics". In: Andreas Baxevanis Al Board, editor. Curr. Protoc. Bioinforma., vol. Chapter 13; Dec. 2012. p. Unit13.19.
[9] Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis Dec. 1999;20(18):3551–67.
[10] 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 RCM 2004;18(18):2162–8.
[11] Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, et al. mzML—a community standard for mass spectrometry data. Mol Cell Proteomics Jan. 2011;10(1) [R110.000133–R110.000133].
[12] Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J Proteomics Oct. 2010;73(11):2092–123.
[13] Iglesias AH, Santos LFA, Gozzo FC. Identification of cross-linked peptides by high-resolution precursor ion scan. Anal Chem Feb. 2010;82(3):909–16.
[14] Eng JK, Jahan TA, Hoopmann MR. Comet: an open-source MS/MS sequence database search tool. Proteomics Jan. 2013;13(1):22–4.
[15] Perez-Riverol Y, Sánchez A, Noda J, Borges D, Carvalho PC, Wang R, et al. HI-bone: a scoring system for identifying phenylisothiocyanate-derivatized peptides based on precursor mass and high intensity fragment ions. Anal Chem Apr. 2013;85(7):3515–20.
[16] Tang X, Munske GR, Siems WF, Bruce JE. Mass spectrometry identifiable cross-linking strategy for studying protein–protein interactions. Anal Chem Jan. 2005;77(1):311–8.
[17] Perez-Riverol Y, Sánchez A, Ramos Y, Schmidt A, Müller M, Betancourt L, et al. In silico analysis of accurate proteomics, complemented by selective isolation of peptides. J Proteomics Sep. 2011;74(10):2071–82.
[18] Jones AR, Eisenacher M, Mayer G, Kohlbacher O, Siepen J, Hubbard SJ, et al. The mzIdentML data standard for mass spectrometry-based proteomics results. Mol Cell Proteomics Jul. 2012;11(7) [M111.014381–M111.014381].
[19] Weisbrod CR, Chavez JD, Eng JK, Yang L, Zheng C, Bruce JE. In vivo protein interaction network identified with a novel real-time cross-linked peptide identification strategy. J Proteome Res Apr. 2013;12(4):1569–79.
[20] Vizcaíno JA, Côté RG, Csordas A, Dianes JA, Fabregat A, Foster JM, et al. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res Jan. 2013;41(Database issue):D1063–9.