Using SIM-XL to identify and annotate cross-linked peptides analyzed by mass spectrometry

Diogo B. Lima (✉ diogobor@gmail.com)

Laboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, Fiocruz, Paraná, Brazil

Fabio C. Gozzo (✉ fabio@iqm.unicamp.br)

Dalton Mass Spectrometry Laboratory, University of Campinas, São Paulo, Brazil

Paulo C. Carvalho (✉ paulo@pcarvalho.com)

Laboratory for Proteomics and Protein Engineering, Carlos Chagas Institute, Fiocruz, Paraná, Brazil

Tatiani B. de lima

Dalton Mass Spectrometry Laboratory, University of Campinas, São Paulo, Brazil

Tiago S. Balbuena

College of Agricultural and Veterinary Sciences, State University of São Paulo, Jaboticabal, São Paulo, Brazil

Ana Gisele C. Neves-Ferreira

Laboratory of Toxinology, Oswaldo Cruz Institute, Fiocruz, Rio de Janeiro, Brazil

Valmir C. Barbosa

Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

Method Article

Keywords: cross-linked peptide identification, cross-linked peptides, cross-linking, search engine, mass spectrometry, search tool, computational biology, structural proteomics

Posted Date: February 13th, 2015

DOI: <https://doi.org/10.1038/protex.2015.015>

License: (cc) (i) This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

State-of-the-art structural mass spectrometry-based proteomics is often accomplished by using cross-linkers to covalently bind two or more amino-acid groups. This strategy is complementary to classical structural biology and therefore broadens the toolset for analyzing protein and protein-protein complex structures. One of the greatest challenges in identifying cross-linked peptides in complex protein mixtures is computationally dealing with the large search space, which grows quadratically with each peptide included in the sequence database. The **Spectrum Identification Machine for Cross-Linked Peptides \ (SIM-XL)** software uses an algorithm that overcomes this limitation by capitalizing on experimental features that allow it to effectively address the massive combinatorial problem at hand, and presents the results in a user-friendly manner. Thus, SIM-XL is recommended for studies dealing with protein structure and protein-protein interaction in either simple or complex protein mixtures. SIM-XL also allows the sharing of results through PRIDE by exporting them in the mzIdentML format.

Introduction

One of the goals of systems biology is to determine how a system works, beginning at the molecular-level characterization of proteins and their interactions up to the level of cellular physiological pathways¹. In a broad sense, the determination of protein structures and protein-protein interactions has an immediate impact on many biological and biotechnological fields, including medicinal chemistry, immunology, and molecular medicine, to name a few. For example, the determination of a protein's three-dimensional structure allows us to better understand its biological function and consequently opens the doors to the design of new drugs or even provides new insights into how to completely engineer new proteins to fulfill specific biological functions². This qualifies the understanding of cell biology, at the atomic level, as a key to answering a number of important biological questions and thus providing a roadmap for numerous biotechnological applications such as the design of new drugs. The current gold-standard methods for determining a protein's structure are X-ray diffraction³ and nuclear magnetic resonance (NMR)⁴. However, the majority of proteins and their complexes are not amenable to these strategies, as they either do not crystallize, require large amounts of high-purity protein, or the system is just too large to be analyzed⁵. These facts make evident the need for developing novel structural approaches applicable to larger and more complex systems. Chemical cross-linking coupled to large-scale mass spectrometry (XL-MS) has become a key method to broaden the toolset for protein structural characterization and in determining protein-protein interactions. During sample preparation, synthetic cross-linkers are included; these covalently link to the side chains of amino acids in proteins and/or their complexes. The sample is then digested, ultimately allowing the cross-linked peptides to be identified by analyzing the mixture using tandem mass spectrometry. The covalently linked peptides carry a very unique piece of information, i.e., distance constraints, which allow us to further the key step of determining the protein-protein interactions^{6,7}. In this paper, we describe the key steps for analyzing 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 environment⁸, to analyze tandem mass spectrometry data derived from cross-linked peptides.

A video demonstrating SIM-XL v1.0 in action is available at "<http://patternlabforproteomics.org/sim-xl/video>".

Equipment

Hardware • A computer with at least a 4 GB RAM and 2 computing cores is recommended.

Software • Windows 7 (32 or 64 bits) or later. • For reading Thermo RAW files, the "MSFileReader":<http://sjsupport.thermofinnigan.com/public/detail.asp?id=703> must be installed. • The .NET framework 4.5, which will be automatically updated by SIM-XL if necessary. • The SIM-XL software is available at "<http://patternlabforproteomics.org/sim-xl/>".

Data files

• SIM-XL v1.0 is compatible with data files in the formats "mzML 1.0":http://www.psdev.info/mzml_1_0_0, MS2⁹ and Mascot Generic Format (MGF)¹⁰ and can work directly with Thermo RAW files if the freely available MSFileReader is locally installed. • SIM-XL exports results to its own format (i.e., *.simxlr) and in the "mzidentML 1.2 draft":<http://www.psdev.info/mzidentml> format that is currently under development by the "HUPO Proteomics Standards Initiative":<http://www.psdev.info/> to support the identification of cross-linked peptides. We note this is the first search engine that enables complete submissions of XL-MS data to PRIDE¹¹, being therefore is compatible with the "PRIDE Inspector software":<http://www.nature.com/nbt/journal/v30/n2/full/nbt.2112.html>.

Procedure

1. Software installation: Download SIM-XL by clicking on the _Download_ button at "<http://patternlabforproteomics.org/sim-xl/>".

2. Workflow

The following workflow demonstrates how to perform a search using the SIM-XL search engine.

2.1. Execute the _Spectrum Identification Machine for Cross-linked Peptides_ (Figure 1) See figure in Figures section.

Figure 1: Graphical User Interface for the main window of SIM-XL.

2.2. Specify the protein sequence database file by clicking on the _Browse_ button. The file extension has to be *.FASTA, *.T-R, or *.T. We refer the reader to Basic Protocol 1: Preparing a sequence database to be searched by ProLucid or the academic Sequest⁸ for more on how to generate target-decoy (T-R) databases compatible with PatternLab.

2.3. Specify a directory containing XL-MS data in any of the following formats: "mzML 1.1.0":http://www.psdev.info/mzml_1_0_0, MS2, MGF, or Thermo RAW. If _Recursive Search_ is checked, all subdirectories will be searched.

2.4. Select one of the pre-registered cross-linkers in the drop-down list. By default, there are five cross-linkers registered: • Disuccinimidyl Suberate (DSS); • Disuccinimidyl Glutarate (DSG); • Disuccinimidyl Glutarate (DSSeb); • Disulfide; and • Zero-length.

Optionally including a new cross-linker in the XL library

2.4.1. To register a new cross-linker, click on the _Edit_ button beside the _Cross-linker_ field and the _XL library_ window will pop up (Figure 2). See figure in Figures section.

Figure 2: XL library. A cross-linker can be inserted or removed in this window.

2.4.2. Fill out the fields _XL Name_, _XL Mass Shift_ (reaction XL mass) in Daltons, _Reaction Sites_, and optionally, the _Modification Mass Shift_ and _Reporter Ions_ masses.

Parameter descriptions

2.4.2.1. _XL Name_ is a user-defined identifier for the cross-linker.

2.4.2.2. _XL

Mass Shift_ is the net mass of the cross-linker that will be added to the peptide masses. 2.4.2.3.

Reaction Sites are all the combinations of amino acids that react with the cross-linker. If the cross-linker reacts with the N-terminus, the keyword _N-TERM_ should be included. For example, the entry for the DSS cross-linker should be _KK KS SS KN-TERM SN-TERM_. Similarly, C-terminal reactivity can be specified using the keyword _C-TERM_. Note that reaction sites must be separated by a single space. Also, when N-TERM is specified SIM-XL performs an unspecific digestion on the first 20 amino acids of all database entries, starting with Methionine; this is done to address the possibility of signal peptides, frequently present in public databases. The optional _Modification Mass Shift_ field defines an artificial modification caused by the cross-linker. For example, DSS/BS3 can react with a single lysine residue, generating the so-called dead-end modification. If this field is filled out and the _Dynamic DB Reduction_ box \ (see **item 2.10.5**) is checked, SIM-XL will use Comet¹² to identify the modified peptides and use this information to only consider theoretical cross-linked peptides that had at least one chain identified with the modification. Consequently, this reduces the search space and is thus indicated when searching complex samples. If the _Modification Mass Shift_ field is left blank, SIM-XL will always consider all possible combinations of cross-linked peptides. The optional _Reporter Ions_ field defines the _m/z_ of fragments that are specific to cross-linked peptides¹³. If these _m/z_ values are given, SIM-XL will only search MS/MS spectra containing at least one of the corresponding fragments, thus speeding up the search; otherwise, SIM-XL will search all spectra. 2.4.3. To finish the new cross-linker inclusion, click on the _Update_ button and then on the _OK_ button. This will make the new cross-linker available within the library and usable for searching. 2.4.4. In order to delete a cross-linker entry from the library, the whole line must be selected \ (by clicking on the row's header cell) and then the DEL key must be pressed, followed by a click on the _Update_ button and another on the _OK_ button. 2.5 Select an enzyme from the drop-down enzyme list. By default, Trypsin and Lys-C are registered. ****Optionally including a new enzyme in the Enzyme library**** 2.5.1. To include a new enzyme, click on the _Edit_ button beside the _Enzyme_ field; the enzyme library window will pop-up \ (**Figure 3**). In an empty line, complete the corresponding fields with the enzyme's name and a regular expression encoding the enzymatic cleavage. [See figure in Figures section](#). **Figure 3: Enzyme library.** An enzyme can be inserted or removed in this window. A regular expression is required to specify the cleavage sites of a new enzyme. For example, the regular expressions for Trypsin and Lys-C are '[KR](?![P])' and '[K]', respectively. For more on building regular expressions we refer the reader to "<http://www.regular-expressions.info/>":<http://www.regular-expressions.info/>. ****** 2.5.2. In case one wishes to remove an enzyme from the library, the whole line must be selected \ (by clicking on the row's header cell) and then the DEL key must be pressed, followed by a click on the _Update_ button and another on the _OK_ button. 2.6. Choose _Enzyme Specificity_ from the drop-down list; the options are: _Semi-Specific_ or _Fully Specific_. The latter refers to peptides originating from a complete digestion \ (i.e., with enzyme cleavage sites at both the C- and the N-terminus). _Semi-specific_ means that the constraint of having a cleavage site at one end is lifted. For example, in the sequence _R.APBCK.A_, where "." denotes the occurrence of cleavage, selecting _Semi-Specific_ will make SIM-XL consider _A_, _AP_, _APB_, _APBC_, _K_, _CK_, _BCK_, _PBCK_, and _APBCK_. Otherwise \ (i.e., if _Fully Specific_ is selected), the search space is limited to _APBCK_. 2.7. Specify the _Precursor_ and _Fragment ppm_ tolerances. 2.8. Check the _Deconvoluted MS/MS_ option if the spectra in the data files

are deconvoluted (i.e., decharged and de-isotoped). We refer the reader to YADA as a tool for deconvoluting mass spectra¹⁴.

Considering modifications 2.9. To include a modification from the `_Modification library_`, select it from the drop-down list and then click on the `_OK_` button found in the `_Modifications_` group box (Figure 4). See figure in Figures section. Figure 4: Select a pre-defined modification or add a new one. To edit a pre-defined modification, click on the `_Edit_` button.

Optionally including new modifications in the Modification library 2.9.1. To include a new modification or edit an existing one, click on the `_Edit_` button. 2.9.1.1. A new window will open. To include a new modification, fill out the fields `_Name_`, `monoisotopic _Mass Shift_`, and `_Amino acid(s)_` to can carry the modification. 2.9.1.2. To delete a modification from the library, select the whole line by clicking on the row's header cell, then press the DEL key, click on the `_Update_` button, and then on the `_OK_` button. 2.9.2. The user should indicate whether the modification is a variable one and whether it applies to the C-term and/or the N-term by checking the corresponding boxes. For example, if not all Methionines in the sample are expected to be oxidized, the modification should be checked as `_variable_`; however, for modifications that are expected in all occurrences of the amino acid, such as, say, carbamidomethylation of cysteine, the `_variable_` option should remain unchecked. 2.9.3. To remove a modification, select the desired one, then click on the `_Remove Modification_` button and confirm the exclusion. 2.10. The `_XL Advanced Parameters_` tab allows access to various parameters that are not usually required to be changed for XL-MS analyses. In this tab, the parameters are divided into three groups: `_SIM-XL Parameters_`, `_Dynamic DB Reduction_` (which uses Comet¹² for performing a preliminary search), and `_Common Parameters_` (indicating that the parameters belong to both the SIM-XL and the Comet search engines).

SIM-XL Parameters 2.10.1. `_Number of Isotopic Possibilities_`: The precursor mass stored in raw data files may not correspond to the monoisotopic peak. This option allows the software to find the correct monoisotopic peak, which is required to identify the molecule but at the cost of opening up the search space. So, for example, for a peptide with a monoisotopic mass of 4000 Da, the most intense peak in the isotopic envelope is M+2 (that is, 4002 Da), which will most likely be selected as the precursor mass. If the number of possibilities is set to three, SIM-XL will search this MS/MS spectrum considering the precursor masses 4002, 4001, and 4000, plus or minus the given ppm tolerance. In this example, the correct monoisotopic precursor mass is 4000 and thus can be correctly identified by SIM-XL. If a high number of isotopic possibilities is set, the search space will increase accordingly and impact SIM-XL's sensitivity negatively. 2.10.2. `_Minimum AA Residues per chain_`: Minimum number of amino acids a peptide should have to be considered a candidate for cross-linking. 2.10.3. `_Maximum results to report_`: Number of top-scoring XL candidates to be reported for each queried spectrum. 2.10.4. `_Intra-link charge_`: Maximum charge of precursor ions to be searched in an intra-molecular cross-link candidate. All candidates are also considered for the inter-molecular searches.

Dynamic DB Reduction parameters 2.10.5. `_Enable_`: If enabled, SIM-XL will run Comet¹² to perform a preliminary search to identify peptides containing the modification specified in the cross-linker definition as modification mass shift (see item 2.4.2). These identifications are used to generate a theoretical cross-linked peptides database in which all entries contain at least one chain previously identified with the modification. 2.10.6. `_XCorr Threshold_`: Minimum Comet XCorr value for identifying peptides containing the modification mass shift specified in item 2.4.2. 2.10.7. `_Minimum number of`

peptides_: Minimum number of peptides required to include a protein to be later used during the generation of the theoretical cross-linked peptide sequence database. ****Common Parameters**** 2.10.8. _Maximum missed cleavages_: Maximum number of missed cleavages allowed during the theoretical digestion of the sequence database. 2.10.9. _Minimum and Maximum MH_: Minimum and maximum masses of singly-charged peptide ions to be searched. 2.10.10. _Peaks Matched cutoff_: Minimum number of matching fragments between the theoretical and experimental spectra. Identifications not satisfying this constraint will be discarded. 2.10.11 _Merge High Resolution Spectra_: Enabling this option will let the search engine merge two or more high-resolution spectra that likely belong to the same precursor. The motivation is that several MS/MS spectra may be acquired from the same precursor during its elution peak; the merged spectrum will have a better signal-to-noise ratio than the individual spectra. 2.10.11.1 _Chromatogram Tolerance \ (seconds)_: Maximum time difference, in seconds, between tandem mass spectra having the same precursor mass to be merged. Note: The _Log_ field reports notes on the progress of the search. [See figure in Figures section.](#) ****Figure 5: XL Advanced Parameters tab.**** 2.11 Once all parameters have been set, we strongly recommend saving them for future searches. This is accomplished by selecting Save SIM-XL Params from the File menu or pressing CTRL + S \ (****Figure 6****). 2.12. To load a previously saved search parameter file, select Load SIM-XL Params from the File menu or press ALT + L \ (****Figure 6****). [See figure in Figures section.](#) ****Figure 6: Save or Load SIM-XL params.**** 2.13. To begin searching, click on the _GO_ button in the _XL Search_ tab.

****3. Exploring the results**** Note: At this point we recommend saving the results by selecting Save results from the File menu or pressing CTRL + S \ (****Figure 7****). [See figure in Figures section.](#) ****Figure 7: Results Browser's File menu.** Here the user can access many features, such as loading or saving the search results, exporting the 2D-Map to an image or PDF file, exporting the Protein Heat map to an image or Excel® \ (XLS) file, and printing or exporting the results to a spreadsheet with all protein interactions and their hit details.

****3.1. 2D-Map**** 3.1.1. The 2D-Map is an interactive map showing all the cross-links identified with a score above the cutoff value given in the _Score_ field. In this map \ (****Figure 8****), each protein is represented as a rectangle of size directly proportional to its sequence length, with residue numbers marking the ticks at the bottom. The protein's ID is shown outside the rectangle, on the left. Each intra-protein cross-link is represented as a red arc. Likewise, inter-protein cross-links are given in blue straight lines. The position of each linker corresponds to the amino-acid number. The user can customize the view by left-clicking on the rectangles and dragging them around, as well as zoom in or out using the zoom bar. By letting the mouse pointer hover over the cross-linker representation, a window will pop up showing linker details such as the linking amino acids and their positions. 3.1.2. By right-clicking on a cross-linker representation, a pop-up window will be displayed showing all the identified cross-links, with corresponding scan numbers, scores, and charge states. The user can then left-click on a desired identification to access the spectrum with the _Spectrum Viewer_ \ (see ****item 4****). 3.1.3. The 2D-Map can be exported as a PNG, TIFF, or JPG image, or as a PDF file: On the File menu, select 2D-Map, then Save Image \ (ALT + I), or select Export 2D-Map \ (ALT + R) \ (****Figure 7****). [See figure in Figures section.](#) ****Figure 8: Protein-protein interaction map \ (2D-Map).**** ****3.2. Loading results**** 3.2.1. SIM-XL can load results in the SIM-XL \ (=**simxlr**= or "mzIdentIML 1.2 draft ":<http://www.psidev.info/mzidentiml> file

formats. This can be accomplished in several ways, the easiest one being to double-click on a SIM-XL results file. If the `_Result Browser_` window is open, select Load Results from the File menu or press ALT + S, as seen in **Figure 7**. Otherwise, if the main window is open, select Load SIM-XL Results from the File menu (or press ALT + R), as seen in **Figure 6**. We note that mzIdentML results can only be loaded within SIM-XL by accessing the Load Results option, which will open another windows where the user can specify both required files, the `_Result file_` (=`**mzIdentML`)= and the `_Data file_` (e.g, =`**mzML`, =`**MGF`, =`**MS2`, or =`**RAW`). [See figure in Figures section.](#) **Figure 9: Input file window.** SIM-XL accepts the mzIdentML format, in addition to its own format (simxlr).

3.3. Dynamic Result Report

3.3.1. A dynamic report is made available by clicking on the `_Dynamic Result Report_` tab (**Figure 10**). The user can sort/search the results according to user-specified criteria. By double-clicking on an entry, the `_Spectrum Viewer_` will open, enabling access to the spectrum for the identification in that entry (see **item 4**). [See figure in Figures section.](#) **Figure 10: Dynamic Result Report.**

3.3.1.1. Filtering results

3.3.1.1.1. `_ScanNumber_`:

In case this field is not empty, only spectra whose scan numbers match that of this field will be displayed.

3.3.1.1.2. `_Score_`:

Only results containing identification scores greater than or equal to this value will be displayed.

3.3.1.1.3. `_ppm_`:

Only results containing a ppm less than or equal to this value will be displayed.

3.3.1.1.4. `_Sequence_`:

Only results from peptides containing the sequence input to this field will be displayed. The user can further specify whether only results from intra-link or inter-link peptides/proteins are to be displayed.

3.3.1.1.5. `_Assessment_`:

Only results containing a `_Personal Assessment_` equal to this value will be displayed. All criteria specified in these fields will be reflected in all tabs (2D-Map, Dynamic Result Report, and Protein Heat map). **Fields in the Dynamic Result Report**

3.3.2. The `_Peptide Sequence_` field shows the search result's identified peptides. The amino acids interacting with the cross-linker are shown in a bold typeface. Double-clicking on this field makes the `_Spectrum Viewer_` (see **item 4**) pop up, enabling the user to assess the spectrum that resulted in the respective identification.

3.3.3. The `_Protein 1_` and `_Protein 2_` columns display the protein(s) that contain the identified peptide sequence(s). The remaining proteins having conserved regions containing the peptide(s) can be assessed by double-clicking on one of these columns (**Figure 11**). [See figure in Figures section.](#) **Figure 11: Proteins containing identified peptides window.** The header displays the identified peptides: TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR and LAVLSSSLTHWKK. Below, the protein that contains these sequence(s) (PROTEIN1) is listed. 3.3.4. The `_Upload Spectra_` column is part of a global effort for improving cross-linker identification scoring functions. By checking beside the desired spectra and then selecting Send Spectra to Server from the Utils menu (or pressing ALT + S), the user will donate the identifications and spectra for further research on the topic.3.3.5. The `_Personal Assessment_` column allows the user to input a personal assessment on the quality of each identification. This is accomplished by selecting from the drop-down list one of the five choices ranging from Excellent to Poor.3.3.6. At this point, we recommend saving the results once again so that the personal assessments can be included. This is done by selecting Save Results from the File menu or pressing CTRL + S.3.4. Protein Heat map3.4.1. The Protein Heat map (**Figure 12**) displays regions where inter-protein cross-linkers were identified. To generate such a map, two proteins must be selected by using the horizontal-axis and the vertical-axis drop-down lists. The heat map can be limited to desired amino acids

by selecting them in the `_Reaction Site_` field. [See figure in Figures section.](#) **Figure 12: A Protein Heat map showing the interaction regions defined by cross-linkers. The red scale is associated with the number of identified XL spectra. By clicking on a cell, all identifications supporting that interaction will be displayed.**

3.4.2. The Protein Heat map can be exported as an image or a spreadsheet file containing the information about the interactions. To export the map as an image, select Save Image, then Protein Heat map, from the File menu \ (or press CTRL + I); to export it as a spreadsheet, select Spreadsheet from the Export Data menu \ (or press ALT + T), or Hit Details \ (i.e., information of the identifications contained in each cell) \ (ALT + H) \ (b>Figure 7).

3.5. Utils menu

3.5.1. _Report Fusion_: This option allows merging several SIM-XL results into a single report. To accomplish this, select Report Fusion from the Utils menu \ (or press ALT + F) and select all files to be joined.

3.5.2. _Custom Report Results_: This option allows the addition or removal of columns in the Dynamic Result Report. For this, select Custom Report Results from the Utils menu \ (or press ALT + C). Following that, a new window containing all columns that can be included or removed will be displayed. After checking beside the features of interest, click on the `_OK_` button. b>4.

Spectrum Viewer

4.1. The _Spectrum Viewer_ \ (b>Figure 13) displays an annotated XL mass spectrum. The `_Spectrum View_` tab allows the user to browse the spectrum, zoom in and out, and easily view which peak was attributed to the corresponding fragment. To zoom in, click and drag the mouse over the desired `_m/z_` range \ (b>Figure 14). To zoom out, double left-click. [See figure in Figures section.](#) **Figure 13: XL Spectrum Viewer. The _Spectrum View_ tab allows the user to browse the spectrum, zoom in and out, as well as easily view which peaks were attributed to which series. A ppm deviation plot is available above the annotated mass spectrum.** [See figure in Figures section.](#) **Figure 14: Zoom-in on a specific _m/z_ range of the XL mass spectrum.**

4.2. A ppm deviation plot is available above the annotated mass spectrum \ (b>Figure 13). This plot displays the deviation between the theoretical and experimental peaks, in ppm, along the `_m/z_` range. The continuous blue line represents the `_Random Sample Consensus_` \ (RANSAC), which is a linear regression of matched peaks. The blue dotted lines represent three standard deviations from the regression line. To save this plot, right-click on the image and choose between `_Copy to clipboard_` or `_Save_`.

4.3. The _Peptide Annotation_ tab \ (b>Figure 15) shows a fragmentation diagram of the cross-linked peptide. The plot can be saved by right-clicking on the image and choosing between `_Copy to clipboard_` or `_Save_`. [See figure in Figures section.](#) **Figures 15: The _Peptide Annotation_ tab displays the fragmentation diagram of the cross-linked peptides.**

4.4. The _Spectrum Prediction Parameters_ tab \ (b>Figure 16) provides a table showing all theoretical fragments and their assignments when matched. Matching criteria are shown in the panel on the left. Changes in these parameters, followed by pressing the `_Plot_` button, will update the assignments. These features allow the user to verify, for example, the effects of changing the cross-linker position or even, say, to evaluate the impact on the score of oxidizing a methionine or even changing the sequence of the matched peptide(s).

4.4.1. _Peptide Sequence 1 and 2_: These are the sequences of `_α_` and `_β_` chains. For intra-links, fill out the `_Peptide Sequence 1_` field only. Any modification mass shift must be enclosed in parentheses after the modified amino acid. For example: oxidation of methionine should be input as 'M(15.9949)'. **4.4.2. _Position XL 1 and 2_:** These are the positions of the cross-linked residues in both chains. For intra-links, both fields correspond to the cross-linking positions in the `_α_`

chain. 4.4.3. **_Deconvoluted MS/MS_**: Check this option if the spectrum is deconvoluted (see **item 2.8**). 4.4.4. **_XL Mass_**: The cross-linker mass shift (see **item 2.4.2.2**). 4.4.5. **_ppm_**: The tolerance of each spectrum peak match. 4.4.6. **_Precursor Charge and Precursor Mass_**: The charge and mass of the precursor ion. 4.4.7. **_Ion Series_**: Check the series to be considered by the **_Spectrum Viewer_**. For inter-links, both **_α_** and **_β_** series should be checked. 4.4.8. Click on the **_Plot_** button to update the **_Spectrum Viewer_**. 4.4.9. The **_Load Example_** button loads an example spectrum. [See figure in Figures section](#). **Figure 16: Spectrum Prediction Parameters tab**. The user can adjust parameters to check assignments. 4.5. The **_Spectrum Plotting Parameters_** tab (**Figure 17**) allows the user to enter an individual experimental mass-spectral peak list. The user can also specify the XL reporter ions. [See figure in Figures section](#). **Figure 17: Spectrum Plotting Parameters tab**. The user can add individual mass-spectral peak lists to visualize the spectral assignments. 4.6. To save the annotated XL mass spectrum, select **Save Spectrum** from the **File** menu or press **CTRL + S**. An image can be saved by selecting **Export Image**, then **Spectrum Image**, from the **File** menu or pressing **CTRL + I**. To load the annotated XL mass spectrum, select **Load Spectrum** from the **File** menu or press **ALT + L**. 4.7. The user can customize the XL mass spectrum annotation by selecting **Custom Spectrum View** from the **Utils** menu or pressing **CTRL + L**. A new window will open with several options, as shown in **Figure 18**. [See figure in Figures section](#). **Figure 18: Custom Spectrum View**. Spectrum annotations can be customized by checking the option in this menu. 4.8. By selecting **Send Spectrum to Server** from the **Utils** menu (or pressing **ALT + S**), SIM-XL will upload the annotated spectrum to a server and thus contribute to a global effort aiming towards creating more sophisticated cross-linker scoring functions through machine learning.

References

- [1] A. Ma'ayan, A. D. Rouillard, N. R. Clark, Z. Wang, Q. Duan, and Y. Kou, "Lean Big Data integration in systems biology and systems pharmacology," *Trends Pharmacol. Sci.*, Aug. 2014.
- [2] C. Jost and A. Plückthun, "Engineered proteins with desired specificity: DARPins, other alternative scaffolds and bispecific IgGs," *Curr. Opin. Struct. Biol.*, vol. 27C, pp. 102–112, Jul. 2014.
- [3] E. F. Garman, "Developments in x-ray crystallographic structure determination of biological macromolecules," *Science*, vol. 343, no. 6175, pp. 1102–1108, Mar. 2014.
- [4] R. Hänsel, L. M. Luh, I. Corbeski, L. Trantirek, and V. Dötsch, "In-Cell NMR and EPR Spectroscopy of Biomacromolecules," *Angew. Chem. Int. Ed Engl.*, Jul. 2014.
- [5] M. Strong, M. R. Sawaya, S. Wang, M. Phillips, D. Cascio, and D. Eisenberg, "Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from *Mycobacterium tuberculosis*," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 21, pp. 8060–8065, May 2006.
- [6] G. W. Preston, S. E. Radford, A. E. Ashcroft, and A. J. Wilson, "Covalent cross-linking within supramolecular peptide structures," *Anal. Chem.*, vol. 84, no. 15, pp. 6790–6797, Aug. 2012.
- [7] E. D. Merkley, J. R. Cort, and J. N. Adkins, "Cross-linking and mass spectrometry methodologies to facilitate structural biology: finding a path through the maze," *J. Struct. Funct. Genomics*, vol. 14, no. 3, pp. 77–90, Sep. 2013.
- [8] P. C. Carvalho, J. S. G. Fischer, T. Xu, J. R. Yates 3rd, and V. C. Barbosa, "PatternLab: from mass spectra to label-free differential shotgun proteomics," *Curr. Protoc. Bioinforma.* Ed. Board Andreas Baxevanis AI, vol. Chapter 13, p. Unit13.19, Dec. 2012.
- [9] W. H. McDonald, D. L. Tabb, R. G. Sadygov, M. J. MacCoss, J.

Venable, J. Graumann, J. R. Johnson, D. Cociorva, and J. R. Yates 3rd, "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*, vol. 18, no. 18, pp. 2162–2168, 2004. \[10] D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell, "Probability-based protein identification by searching sequence databases using mass spectrometry data," *Electrophoresis*, vol. 20, no. 18, pp. 3551–3567, Dec. 1999. \[11] J. A. Vizcaíno, R. G. Côté, A. Csordas, J. A. Dianes, A. Fabregat, J. M. Foster, J. Griss, E. Alpi, M. Birim, J. Contell, G. O'Kelly, A. Schoenegger, D. Ovelheiro, Y. Pérez-Riverol, F. Reisinger, D. Ríos, R. Wang, and H. Hermjakob, "The PRoteomics IDentifications (PRIDE) database and associated tools: status in 2013," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D1063–1069, Jan. 2013. \[12] J. K. Eng, T. A. Jahan, and M. R. Hoopmann, "Comet: an open-source MS/MS sequence database search tool," *Proteomics*, vol. 13, no. 1, pp. 22–24, Jan. 2013. \[13] A. H. Iglesias, L. F. A. Santos, and F. C. Gozzo, "Identification of cross-linked peptides by high-resolution precursor ion scan," *Anal. Chem.*, vol. 82, no. 3, pp. 909–916, Feb. 2010. \[14] P. C. Carvalho, T. Xu, X. Han, D. Cociorva, V. C. Barbosa, and J. R. Yates, "YADA: a tool for taking the most out of high-resolution spectra," *Bioinformatics*, vol. 25, no. 20, pp. 2734–2736, Oct. 2009.

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. They also thank the PRIDE Team for working together with them to enable SIM-XL to support the next version of mzIdentML.

Figures

Figure 1

Figure 5 SIM-XL Advanced Parameters

Figure 2

Figure 6 Save or Load SIM-XL parameters

Figure 3

Figure 8 Protein-protein interactions Map (2D-Map)

Figure 4

Figure 9 Input file window SIM-XL accepts the mzIdentML format, in addition to its own format (simxlr).

Figure 5

Figure 10 Dynamic Result Report

Figure 6

Figure 11 Protein interactions window Proteins containing identified peptides window. The header displays the identified peptides: TDEQALLSSILAKTASNIIDVSAADSQGMEQHEYMDR and LAVLSSSLTHWKK. Below, the protein that contains these sequence(s) (PROTEIN1) is listed.

Figure 7

Menu File of results Here the user can access many features, such as loading or saving the search results, exporting the 2D-Map to an image or PDF file, exporting the Protein Heat map to an image or Excel© (XLS) file, and printing or exporting the results to a spreadsheet with all protein interactions and their hit details.

Figure 8

Figure 12 Protein interactions map A Protein Heat map showing the interaction regions defined by crosslinkers. The red scale is associated with the number of identified XL spectra. By clicking on a cell, all identifications supporting that interaction will be displayed.

Figure 11

Figure 15 Peptide Annotation The Peptide Annotation tab displays the fragmentation diagram of the cross-linked peptides.

Figure 12

Figure 16 Spectrum Prediction Parameters tab The user can adjust parameters to check assignments.

Figure 13

Figure 17 Spectrum Plotting Parameters tab The user can add individual mass-spectral peak lists to visualize the spectral assignments.

Figure 14

Figure 18 Custom Spectrum View Spectrum annotations can be customized by checking the option in this menu.

Figure 15

Figure 1 SIM-XL Main GUI In this window the user will perform the search filling out all fields according to each analysis.

Figure 16

Figure 2 Cross-linker (XL) library A cross-linker can be inserted or removed in this window.

Figure 17

Figure 3 Enzyme library An enzyme can be inserted or removed in this window. A regular expression is required to specify the cleavage sites of a new enzyme. For example, the regular expressions for Trypsin and Lys-C are '[KR](?!P)' and '[K]', respectively. For more on building regular expressions we refer the reader to <http://www.regular-expressions.info/>.

Figure 18

Figure 4 Add/Remove Modifications Select a pre-defined modification or add a new one. To edit a pre-defined modification, click on the Edit button.