Published in final edited form as:

J Proteome Res. 2019 February 01; 18(2): 759–764. doi:10.1021/acs.jproteome.8b00726.

Proxl (Protein Cross-linking Database): A public server, QC tools, and other major updates.

Michael Riffle^{1,2,*}, Daniel Jaschob¹, Alex Zelter¹, and Trisha N. Davis¹

¹Department of Biochemistry, University of Washington, Seattle, WA 98195, USA

²Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA

Abstract

Proxl is an open source web application for sharing, visualizing, and analyzing bottom-up protein cross-linking mass spectrometry data and results. Proxl's core features include comparing datasets, structural analysis, customizable and interactive data visualizations, access to all underlying mass spectrometry data, and quality control tools. All features of Proxl are designed to be independent of specific cross-linker chemistry or software analysis pipelines. Proxl's sharing tools allow users to share their data with the public or securely restrict access to trusted collaborators. Since being published in 2016, Proxl has continued to be expanded and improved through active development and collaboration with cross-linking researchers. Some of Proxl's new features include a centralized, public site for sharing data, greatly expanded quality control tools and visualizations, support for stable isotope labelled peptides, and general improvements that make Proxl easier to use, data easier to share and import, and data visualizations more customizable. Source code and more information may be found at <http://proxl-ms.org/>.

Keywords

cross-linking; data visualization; structure; proteomics; software; bioinformatics; database

INTRODUCTION

Protein cross-linking mass spectrometry (XL-MS or CL-MS) has been gaining momentum in recent years as a tool for characterizing protein interactions and protein complex architecture, as well as how these respond to different conditions or perturbations. At a high level, XL-MS experiments are similar to typical bottom up liquid chromatography tandem mass spectrometry (LC-MS/MS) experiments: proteins are digested into peptides, peptides separated on a LC column, and injected into and analysed by a mass spectrometer. For XL-MS analysis, proteins are subjected to a cross-linking reaction prior to digestion, which covalently attaches one residue of a protein to another (typically by a spacer arm of known length). Identification and analysis of cross-linked peptide pairs may then be used to infer spatial characteristics of interacting proteins in solution. For a more comprehensive review of XL-MS, see Chavez et al (2018)¹.

*Corresponding author: UW Box 357350, 1705 NE Pacific St., Seattle WA 98195-7350, Phone: 206-685-3740, MR: mriffle@uw.edu.

As XL-MS became more widely adopted, several tools emerged for visualizing identified cross-links in proteins²⁻⁵. The primary motivation behind the development and publication of Proxl⁶ was to build a single platform that provided dynamic interactive data visualization of results and access to all underlying mass spectrometry data. This needed to be done in a way that was independent of any specific cross-linking method or software analysis pipeline. When published, Proxl achieved these design goals and included robust sharing tools (public or private), multiple customizable and interactive data visualizations, and tools to compare results from multiple experiments (including experiments from different pipelines). However, potential users of Proxl were required to install their own copy of Proxl on their own server, which required expertise in setting up a web server, database, and system administration. Additionally, uploading data into Proxl required running a separate importer program by hand, which also required skill with databases and underlying operating systems.

Since publication, Proxl has been updated to address these limitations while adding many new capabilities. Proxl now supports uploading data directly via the web interface and a new publicly accessible server has been set up that allows anyone to create an account and immediately upload, analyze, and share data (public and private) with no need to install any software. Other updates include greatly expanded quality control (QC) visualizations, native support for stable isotope labelled peptides, a new interactive, dynamic Circos-style⁷ visualization of cross-links, support for more software analysis pipelines, other optimizations and user interface improvements. More information about the public server, comprehensive documentation, and source code may be found at <http://proxl-ms.org/>.

NEW FEATURES

Public Server and Web-based Data Upload

To make it as simple as possible to begin using Proxl, we have set up a public server at https://www.yeastrc.org/proxl_public/. This service is free to use and available to anyone. All features are enabled, there are no restrictions on how much data may be uploaded, the data may be publicly shared (e.g., for publication) or private (e.g., shared only between collaborators), and there is no expiration date on the data. In spite of not being published previously, 154 researchers have created accounts and have uploaded 378 runs to 172 projects. Raw search results have been uploaded to the public Proxl server from 23 distinct combinations of software pipelines and versions of those pipelines (9 distinct analysis programs). To gain access to the public Proxl server, visit the above URL, click the “Signup” tab, and fill out a simple form to create an account. Note that Proxl may still be installed and run privately by downloading it from our source code repository (<http://proxl-ms.org/>).

Where previously data could only be uploaded to Proxl by running an external program, data may now be uploaded directly through the web application. To upload data log into Proxl, visit (or create) a project, expand the “Upload Data” section, and click “Import Proxl XML File” to submit a Proxl XML file (and optionally an mzML or mzXML file). Proxl requires that the native output of a software analysis pipeline be converted to Proxl XML for import. To facilitate this, the authors of Proxl currently maintain Proxl XML converters for Kojak⁸ (used with Percolator⁹ or Trans-Proteomic Pipeline¹⁰), Crux (search-for-xlinks)¹¹, pLink

(1.x and 2.x)¹², StavroX¹³, MetaMorpheus¹⁴, and xQuest¹⁵. For more comprehensive documentation of uploading data and to download the Proxl XML converters, visit https://proxl-web-app.readthedocs.io/en/latest/using/upload_data.html.

Expanded QC Visualizations

The QC and run/search summary statistic visualizations in Proxl have been greatly expanded since its original publication. Proxl now includes 18 distinct plots designed to visually assess poor performance, misconfiguration, or biases in chromatography, mass spectrometer performance, sample preparation, and analysis software pipelines. A selection of the available QC plots are shown in Figure 1.

Current plots include summary statistics displaying peptide-spectrum match (PSM), peptide, and protein counts for different classes of peptides (cross-links, loop-links, or unlinked); digestion statistics, which examine fractions of peptides and PSMs with missed cleavages; scan file statistics, which interrogate abnormal MS1 and MS2 ion current with respect to retention time or m/z; PSM level statistics, which aim to show unexpected distributions of PSM counts with respect to quality scores, charge states, m/z, peptide length, and retention time; PSM error, which aims to show unexpected distributions of measured error for PSMs with respect to retention time or m/z; modification statistics, which show relative distributions of PSMs with post-translational modifications; and peptide level statistics, which shows the distribution of peptide lengths for each peptide class. Images of each plot or the underlying data in tab-delimited format may be downloaded by selecting the download icon and choosing the desired format.

The QC plots can help users identify problems with their mass spectrometry data from a wide array of sources. For example, errors in chromatography can be spotted viewing the MS1 ion current plots under “Scan File Statistics” or examining peptide length versus retention time under “Chromatography Statistics”. Mass spectrometer performance can be examined by viewing the ratio of scans that resulted in identifications, error distributions, error versus retention time, and error versus m/z of identified ions. Digestion or cross-linker inefficiencies can be interrogated by viewing the number of PSMs (both unlinked and cross-linked) and the number of peptides and PSMs with missed cleavages. The performance and character of the underlying software analysis pipeline can be examined by plotting any score against the number of PSMs and by plotting any score against any other score, such as Xcorr versus calculated q-value.

For a demonstration of the QC page, see https://www.yeastrc.org/proxl_public/go?SEuNw6AEGU. For a demonstration of the QC page comparing two runs, see https://www.yeastrc.org/proxl_public/go?YgmuQ8IAkk.

Stable Isotope Labelled Peptides and Intra-/Inter-Protein Distinction

If two cross-linked peptides identified in an XL-MS experiment map to the same protein, it is typically not possible to determine if that cross-link is within the same copy of a protein (intra-protein) or between two copies (inter-protein). However, in a mixture of stable isotope labelled proteins subjected to a cross-linking reaction (e.g. N14/N15), N14 peptides linked to N15 peptides indicate inter-protein cross-links.

In collaboration with biochemistry researchers, Proxl has been updated with native support for stable isotope labelled peptides. The code base, database, and Proxl XML format have all been updated to encode isotope label state for identified peptides. Using the updated user interface, researchers can now include labelled and unlabelled versions of proteins of interest in the visualization tools. If the user adds the labelled and unlabelled version of the same protein to the visualization, these are treated as separate proteins. Any links drawn between them must be inter-protein cross-links—since the labelled peptide in a cross-link will only be mapped to a labelled protein in Proxl. This allows users to quickly interrogate which links must be between two different copies (one heavy and one light) of the same protein in a protein complex. The built-in Lorikeet spectrum viewer¹⁶ has been updated to be isotope label aware to ensure peaks are annotated correctly for labelled peptide identifications.

Data visualization and UI Improvements

Circle Plots—A new visualization, here referred to as a “circle plot”, of cross-links within and between proteins was added to Proxl (Figure 2). This is an alternate representation of the same data present in the “bar plot” that may make broader patterns in cross-linking easier to visualize—particularly when viewing larger numbers of proteins or cross-links. The circle plot contains fewer intersecting links and cross-links do not intersect with the protein bars. The circle plot is based on the visualizations produced by the Circos visualization package, which is a flexible and powerful tool written in Perl that is widely used to illustrate connections between genomic features. However, the outputs of Circos are non-interactive static images. For Proxl, we re-implemented the aesthetics of the Circos output to produce dynamic, fully-interactive vector images—where users may choose which proteins are shown, choose how those protein sequences are annotated, click or mouse over proteins or links for more information and to view underlying mass spectrometry data, and more. The Proxl circle plot is implemented in Javascript and runs entirely within the web browser and requires no 3rd party extensions. The interactive image is rendered as scalable vector graphics (SVG), and may be downloaded either as vector art (for use in presentation slides or programs such as Adobe Illustrator) or as rasterized images (JPEG or PNG). For a demonstration of circle plots, visit https://www.yeastrc.org/proxl_public/go?EPJ8ci26UI. For comprehensive documentation of circle plots, visit <https://proxl-web-app.readthedocs.io/en/latest/using/image-circle.html>.

Region Selections and Exclusions—Researchers are often interested in visualizing the cross-links between specific regions of proteins or excluding a particularly “promiscuous” region of a protein that ubiquitously cross-links an entire protein complex. This is particularly important if the visualization includes many proteins and/or many cross-links. Both the linear protein bar plot and circle plots were updated to allow selecting specific segments of individual proteins to highlight or exclude. These features may be accessed by clicking “Manage Protein Selections” or “Manage Link Exclusions” on either the linear protein bar or circle plot views. Additionally, regions may be selected by holding control and clicking and dragging over regions of interest in proteins on the protein bar view. For an example of region selections, please visit https://www.yeastrc.org/proxl_public/go/?olC47bO6M (protein bar view) or https://www.yeastrc.org/proxl_public/go?IE2FYYQKgY (circle plot view).

Custom Feature Annotations—Another common need for researchers is to annotate protein sequences with regions of interest, known domain motifs, or some other type of descriptive text. On both the protein bar plot and circle plot, project owners may click the “Custom Annotation Manager” to define custom annotations for regions of proteins in the experiment. (Note this option only appears for project owners.) They are able to designate the start and stop position, text to be shown, and the color to use to indicate that segment. After saving these, users may choose the “Custom Regions” option in the “Show Feature Annotations” pull-down menu to view the custom region annotations on the proteins.

USAGE AND SCALABILITY

In the two years since publication, data has continued to be collected and inserted into the authors’ installation of the Proxl web application and database. To address questions of scalability, the authors note that this installation contains the results from 1,102 separate XL-MS runs, identifying 2,482,776 distinct peptide ions (unlinked, loop-linked, and cross-linked peptides), from 24,777,398 PSMs. This Proxl installation and MySQL database run on a single Ubuntu 16.04.4 Linux installation using a 4 TB NVMe solid state hard disk and 256 GB of RAM. To date the Proxl installation has not experienced any issues with performance degradation or stability.

CONCLUSION

Proxl is a web application for sharing, visualizing and analyzing XL-MS data. Proxl was originally published with the motivation of providing a single application to dynamically and interactively visualize XL-MS data that was independent of any cross-linking technology or software analysis pipeline. It provided access to all underlying mass spectrometry data and provided robust public and private sharing tools. Since publication, using Proxl has been simplified via establishment of a public Proxl server at https://www.yeastrc.org/proxl_public/, which allows anyone to begin using Proxl without installing software. Uploading data has been streamlined, QC tools have been expanded, stable isotope labelled peptides have been supported, and new visualizations and customization options have been added. More information, including access to source code, can be found at <http://proxl-ms.org/>.

ACKNOWLEDGMENTS

This work is supported by grants P41 GM103533 from the National Institute of General Medical Studies from the National Institutes of Health; and the University of Washington Proteomics Resource (UWPR95794).

ABBREVIATIONS:

CL-MS	Protein Cross-linking Mass Spectrometry
PSM	peptide-spectrum match
QC	quality control
XL-MS	Protein Cross-linking Mass Spectrometry

REFERENCES

1. Chavez JD; Bruce JE, Chemical cross-linking with mass spectrometry: a tool for systems structural biology. *Curr Opin Chem Biol* 2018, 48, 8–18. [PubMed: 30172868]
2. Kosinski J; von Appen A; Ori A; Karius K; Muller CW; Beck M, Xlink Analyzer: software for analysis and visualization of cross-linking data in the context of three-dimensional structures. *J Struct Biol* 2015, 189 (3), 177–83. [PubMed: 25661704]
3. Combe CW; Fischer L; Rappsilber J, xiNET: cross-link network maps with residue resolution. *Mol Cell Proteomics* 2015, 14 (4), 1137–47. [PubMed: 25648531]
4. Grimm M; Zimniak T; Kahraman A; Herzog F, xVis: a web server for the schematic visualization and interpretation of crosslink-derived spatial restraints. *Nucleic Acids Res* 2015, 43 (W1), W362–9. [PubMed: 25956653]
5. Zheng C; Weisbrod CR; Chavez JD; Eng JK; Sharma V; Wu X; Bruce JE, XLink-DB: database and software tools for storing and visualizing protein interaction topology data. *J Proteome Res* 2013, 12 (4), 1989–95. [PubMed: 23413830]
6. Riffle M; Jaschob D; Zelter A; Davis TN, ProXL (Protein Cross-Linking Database): A Platform for Analysis, Visualization, and Sharing of Protein Cross-Linking Mass Spectrometry Data. *J Proteome Res* 2016, 15 (8), 2863–70. [PubMed: 27302480]
7. Krzywinski M; Schein J; Birol I; Connors J; Gascoyne R; Horsman D; Jones SJ; Marra MA, Circos: an information aesthetic for comparative genomics. *Genome Res* 2009, 19 (9), 1639–45. [PubMed: 19541911]
8. Hoopmann MR; Zelter A; Johnson RS; Riffle M; MacCoss MJ; Davis TN; Moritz RL, Kojak: efficient analysis of chemically cross-linked protein complexes. *J Proteome Res* 2015, 14 (5), 2190–8. [PubMed: 25812159]
9. Kall L; Canterbury JD; Weston J; Noble WS; MacCoss MJ, Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* 2007, 4 (11), 923–5. [PubMed: 17952086]
10. Deutsch EW; Mendoza L; Shteynberg D; Farrah T; Lam H; Tasman N; Sun Z; Nilsson E; Pratt B; Prazan B; Eng JK; Martin DB; Nesvizhskii AI; Aebersold R, A guided tour of the Trans-Proteomic Pipeline. *Proteomics* 2010, 10 (6), 1150–9. [PubMed: 20101611]
11. 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* 2010, 9 (5), 2488–95. [PubMed: 20349954]
12. Yang B; Wu YJ; Zhu M; Fan SB; Lin J; Zhang K; Li S; Chi H; Li YX; Chen HF; Luo SK; Ding YH; Wang LH; Hao Z; Xiu LY; Chen S; Ye K; He SM; Dong MQ, Identification of cross-linked peptides from complex samples. *Nat Methods* 2012, 9 (9), 904–6. [PubMed: 22772728]
13. Gotze M; Pettelkau J; Schaks S; Bosse K; Ihling CH; Krauth F; Fritzsche R; Kuhn U; Sinz A, StavroX—a software for analyzing crosslinked products in protein interaction studies. *J Am Soc Mass Spectrom* 2012, 23 (1), 76–87. [PubMed: 22038510]
14. Lu L; Millikin RJ; Solntsev SK; Rolfs Z; Scalf M; Shortreed MR; Smith LM, Identification of MS-Cleavable and Noncleavable Chemically Cross-Linked Peptides with MetaMorpheus. *J Proteome Res* 2018, 17 (7), 2370–2376. [PubMed: 29793340]
15. Rinner O; Seebacher J; Walzthoenl T; Mueller LN; Beck M; Schmidt A; Mueller M; Aebersold R, Identification of cross-linked peptides from large sequence databases. *Nat Methods* 2008, 5 (4), 315–8. [PubMed: 18327264]
16. Sharma V; Eng JK; MacCoss MJ; Riffle M, A mass spectrometry proteomics data management platform. *Mol Cell Proteomics* 2012, 11 (9), 824–31. [PubMed: 22611296]

Figure 1. Selected examples of QC visualizations in Proxl. The top panel, labelled “Summary Statistics,” shows bar charts comparing counts for PSMs, peptides, and proteins for each class of peptide (cross-link, loop-links, and unlinked). The next panel, labelled “Digestion Statistics,” compares the number of PSMs and peptides with missed cleavages for each class of peptide. “Scan File Statistics” shows information related to total MS1 and MS2 ion current and how these change as functions of retention time or m/z of parent ions. “PSM Error Estimates” shows the distribution of estimated PSM error (for PSMs meeting current cut offs), and how the distribution of error changes as a function of retention time or m/z of parent ions.

Figure 2.

Example of the utility of circle plots generated by Proxl. Both panels A and B depict the same set of proteins and crosslinks and the same display settings. Each image was directly downloaded from Proxl and was unaltered. (A) The protein bar view from Proxl. Each protein bar is the linear representation of a distinct protein sequence. The colored lines represent cross-links between proteins, and the arcs above and below each protein depict intra-protein cross-links and loop-links respectively. The shaded regions on the protein bars depict sequence coverage for each protein. (B) The circle plot view from Proxl. Each curved colored bar represents a distinct protein. The thinner outer bar of the same color represent the sequence coverage for that protein. The arcs within and between the arcs represent cross-links, loop-links, and intra-protein cross-links. Organizing as a circle aids in interpreting

highly connected regions of proteins, patterns that may not be as accessible in the protein bar view because of large numbers of highly intersecting lines.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript