**iPRG-2023 Proteome Informatics Research Group Study on Crosslinking MS Data Analysis**

Dear iPRG-2023 Study Participant,

Thank you for your interest in this year's ABRF Proteome Informatics Research Group (iPRG) study. This letter provides the instructions needed to access the data files, complete your analysis, and submit your results. This the first iPRG study in which guided tutorials will be made available for analysis of the first half of the data, facilitating participation by novices to crosslinking MS data analysis. Guidance will not be provided for analysis of the second half of the data. Experienced participants are encouraged to use their own analysis pipeline for the entirety of the data, to provide greater diversity in the results collected. This will permit better comparison of analytical techniques for analysis of crosslinking MS data. The deadline for submission of your results is January 15, 2024 to qualify for inclusion in the study report that will be presented at the next ABRF Annual Meeting in Minneapolis, April, 2024.

**Study Overview**

Crosslinking MS is performed to infer interaction sites between proteins and structural restraints from individual proteins in purified or enriched protein complexes, or the full complement of proteins in a cellular lysate. For most experiments, proteins are chemically crosslinked, enzymatically digested, and analyzed by data-dependent acquisition on a high-performance mass spectrometer. Protein structure restraints and interactions are inferred by the identification of two distinct peptide sequences crosslinked in a tandem mass spectrum derived from a single precursor ion. Although crosslinked peptide analysis is fundamentally the same as typical data-dependent MS peptide analysis, processing pipelines for crosslinked peptide data often require specialized algorithms with parameters and constraints that differ from the more familiar ones employed for traditional protein identification/characterization. Coupled with the fact that crosslinked peptides are often sparsely observed in samples, validation and visualization of the data and results to produce confident and coherent conclusions is not trivial.

The goal of this study is to gain information from the proteomics community about the different approaches for crosslinked peptide MS data analysis, showcasing the multitude of different pipelines and tools available, and providing guidance for improving the presentation of crosslinked MS results. In contrast to iPRG studies in the past, we are providing tutorials for analyzing the first half of the data in the study. These tutorials are designed such that any participant with basic knowledge of data-dependent acquisition MS data analysis can participate without any prior knowledge of crosslinked peptide mass spectrometry. With this approach, we hope to encourage participation by non-specialists and to better understand the common pitfalls users encounter when learning to perform crosslinked peptide analysis. Since we're not providing guidance for the analysis of the second half of the data, novice participants will be challenged to apply what they've learned and process a separate dataset on their own. Exploring pipelines in addition to the ones presented in the tutorials and obtained from the scientific literature and community are highly encouraged. Participants with prior experience in crosslinked peptide analysis are encouraged to forego the tutorials entirely and proceed with their preferred pipeline on the complete data set.

**Instructions**

Participants are asked to use one or more database search methods to identify crosslinked residues of protein complexes from data-dependent mass spectrometry data. Two datasets of different protein complexes are provided for which the participants must identify a) the proteins that are crosslinked to each other, b) the specific residues that are crosslinked, and c) the certainty or confidence (e.g., probabilities or q-values) of their conclusions. For both datasets, the proteins were crosslinked with DSSO (Thermo Fisher Scientific) and the same data acquisition protocol was used (see Dataset Details below).

Tutorials for three analysis pipelines are provided for analysis of the first half of the data. Participants are welcome to use one or more of these pipelines or any MS crosslinking software tools available to them. It is requested that participants use the same pipeline(s) for subsequent analysis of the second half of the data.

Participants will be able to report their findings in a variety of formats, ranging from plain text to R Markdown or Jupyter notebooks, as long as the results are clearly described and human-interpretable. In addition to tabulated data, results may be submitted as visual interaction maps. However, it is essential that all software tools used in the analysis be named, with all parameters provided.

**Deliverables**

Submit the following no later than January 15, 2024:

1. A list of the software tools (and versions) used in the analytical pipeline *and their parameters.* A “Methods Description Template” and “Software Description Template” are provided for guidance. Participants should provide a brief written description of their pipeline that should resemble a methods section for publication. Additionally, any human-readable configuration files for any of the tools used should be submitted. Please specify which tools were used for each of these aspects of the crosslinked peptide analysis:

1. Crosslinked peptide identification.
2. Statistical validation of crosslinked peptides (FDR estimation, probability assignment, etc.) at a threshold of your choosing. Please provide a brief statement of why the threshold was chosen.
3. Visualization of crosslinked residues (if used).

2. The sets of identified and validated crosslinked residues for each stage, using the “XL Results Template” provided.

3. A list of crosslinked peptide spectrum matches (CSMs) and peptide spectrum matches (PSMs) and the probabilities or error estimates assigned to them in a suitable open data format (mzIdentML and PepXML preferred).

4. A brief written summary that identifies which proteins are crosslinked to each other in each of the two datasets, along with any comments about the biological significance.

**Study Package**

1. One copy of these instructions.
2. Two datasets of three Thermo Lumos raw files and mzML conversions.
3. One FASTA protein sequence file.
4. Three self-guided tutorials on MS crosslinked data analysis.
5. A set of submission templates for results.

The Study Package is available on https://osf.io/ycvw7

**Dataset Details**

DSSO cross-linked complexes were separated by SDS-PAGE (12% Bis-Tris) and stained with Coomassie blue. The bands of interest were manually excised, individually reduced in situ with TCEP and alkylated in the dark with iodoacetamide prior to treatment with trypsin (Promega, sequencing grade). Each digest was analyzed by capillary HPLC-electrospray ionization tandem mass spectrometry on a Thermo Scientific Orbitrap Fusion Lumos mass spectrometer. On-line HPLC separation was accomplished with an RSLC NANO HPLC system (Thermo Scientific/Dionex): column, PicoFrit (New Objective; 75 μm i.d.) packed to 15 cm with C18 adsorbent (Vydac; 218MS 5 μm, 300 Å). Precursor ions were acquired in the orbitrap in centroid mode at 120,000 resolution (m/z 200); data-dependent higher-energy collisional dissociation (HCD) spectra were acquired in the orbitrap using 30% normalized collision energy and 15,000 resolution. Other MS scan parameters included: mass window for precursor ion selection, 0.7; charge states, 2 – 5; dynamic exclusion, 15 sec (± 10 ppm); intensity to trigger MS2, 50,000.

For convenience, each raw data file was converted to mzML format using ProteoWizard (version 3.0.23264) and the following parameters: “--zlib --filter “peakPicking true 1-“--filter “zeroSamples removeExtra”. These parameters store the spectra in centroid format and zlib compressed (lossless) to preserve disk space. Participants are welcome to convert the raw data to a different format or with different parameters to suit the needs of their analytical pipeline.

**Submission of Results**

Participants are welcome to submit up to three sets of results for the study. Each submission will be assigned an identifier and private results submission portal. We ask that contact information be provided in case the study committee has follow-up questions. Although results submission is not anonymous, all results will be reported anonymously.

You may update a submission via the submission portal at any time until the submission deadline. Results will be final on January 15, 2024.

Important note to vendors and commercial laboratories: ABRF imposes strict guidelines on the use of study results for marketing purposes. These guidelines are described in https://abrf.memberclicks.net/assets/docs/research\_group\_study\_participation\_guidelines\_2010.pdf.

Questions? Please send questions to iPRG.XL.2023@gmail.com. We thank you for your support of the ABRF and look forward to receiving your results for the study.

Sincerely,

The ABRF Proteome Informatics Research Group (iPRG):

Michael Hoopmann (Co-Chair) - Institute for Systems Biology, Seattle, WA

Pratik Jagtap (Co-Chair) - University of Minnesota, Minneapolis, MN

Viktoria Dorfer - University of Applied Sciences, Upper Austria (FH OOE), Hagenberg, Austria

Melanie Föll - University Medical Center Freiburg, Germany

Magnus Palmblad - Leiden University Medical Center, The Netherlands

Yasset Perez-Riverol - European Bioinformatics Institute, Hinxton, UK

Susan T. Weintraub (EB Liaison) - University of Texas Health Science Center at San Antonio