Published in final edited form as:

J Proteome Res. 2011 March 4; 10(3): 923–931. doi:10.1021/pr100848a.

Xlink-Identifier: An Automated Data Analysis Platform for Confident Identifications of Chemically Cross-linked Peptides using Tandem Mass Spectrometry

Xiuxia Du^{1,*}, Saiful M. Chowdhury⁴, Nathan P. Manes³, Si Wu², M. Uljana Mayer², Joshua N. Adkins², Gordon A. Anderson², and Richard D. Smith^{2,*}

¹Department of Bioinformatics & Genomics, University of North Carolina at Charlotte, Charlotte, NC 28023, USA

²Pacific Northwest National Laboratory, Richland, WA 99352, USA

³National Institute of Allergy and Infectious Diseases, Bethesda, MD 20892, USA

⁴National Institute of Environmental Health Sciences – National Institutes of Health, Research Triangle Park, NC 27709, USA

Abstract

Chemical cross-linking combined with mass spectrometry provides a powerful method for identifying protein-protein interactions and probing the structure of protein complexes. A number of strategies have been reported that take advantage of the high sensitivity and high resolution of modern mass spectrometers. Approaches typically include synthesis of novel cross-linking compounds, and/or isotopic labelling of the cross-linking reagent and/or protein, and label-free methods. We report Xlink-Identifier, a comprehensive data analysis platform that has been developed to support label-free analyses. It can identify inter-peptide, intra-peptide, and deadend cross-links as well as underivatized peptides. The software streamlines data pre-processing, peptide scoring, and visualization and provides an overall data analysis strategy for studying protein-protein interactions and protein structure using mass spectrometry. The software has been evaluated using a custom synthesized cross-linking reagent that features an enrichment tag. Xlink-Identifier offers the potential to perform large-scale identifications of protein-protein interactions using tandem mass spectrometry.

Keywords

Chemical cross-linking; mass spectrometry; peptide identification; protein-protein interaction; protein structure

*To whom correspondence should be addressed. Xiuxia Du, 500 Laureate Way, Suite 2350, Kannapolis, NC 28081, Phone: (704) 250-5754, Fax: (704) 250-5759, xiuxia.du@uncc.edu .

Supporting Information Available

Figure S1: Flow chart of the overall data analysis procedure. Figure S2: Flow chart of the data analysis procedure for identifying each of the four types of peptides in Xlink-Identifier. Figure S3: Effect of denoising on MS/MS spectra. File S1 (tab-delimited text files): all of the identifications of inter-peptide cross-links generated by Xlink-Identifier from CID and ETD spectra with charge states 4+ and 5+. File S2: raw mass spectrometry data. File S3: CID and ETD MS/MS spectra in .dta format with charge states 4+ and 5+. File S4: an html file that displays match details for the CID spectrum that corresponds to scan 10894. This information is available free of charge via the Internet at <http://pubs.acs.org>.

Introduction

Chemical cross-linking combined with mass spectrometry is a powerful technique for the identification of protein-protein interactions (PPIs) and study of the structure of proteins complexes. Compared to other techniques, such as nuclear magnetic resonance,¹ X-ray crystallography,² yeast two-hybrid systems,³⁻⁵ affinity chromatography,³ and coimmunoprecipitation,⁵ the cross-linking method has many advantages that include: (1) both interacting partners and interacting sites can be identified in one experiment; (2) protein complexes may be studied *in vitro* or *in vivo*;^{6, 7} (3) broad applicability when bottom-up proteomics analysis approaches are employed;⁸ (4) the availability of cross-linking reagents with different lengths and amino acid specificities, and (5) high analysis sensitivity.

As a result, a number of techniques have been developed recently to take full advantage of the high mass measurement accuracy and high sensitivity of modern mass spectrometers.⁹⁻¹⁷ Most of these techniques employ a bottom-up proteomics strategy wherein cross-linked proteins are digested into peptides that are subsequently analyzed. Identifying a cross-linked peptide using tandem mass spectra benefits from a database of all possible cross-linked peptides from the candidate proteins. Its identification can then be accomplished by querying the database to generate a list of candidate cross-linked peptides having molecular masses within a specified precursor mass tolerance, comparing the experimental spectrum with the theoretical spectrum of each candidate, and ultimately selecting the best match and determining its statistical confidence.

For simple samples containing only a small number of known proteins, this approach works very well and cross-linked proteins can be identified confidently and quickly from the sequence information of cross-linked peptides obtained from the tandem mass spectrometry data. A number of data analysis packages have been developed to analyze spectra data including ASAP,¹⁸ X!Link,¹⁹ GPMW,^{20, 21} MS-Bridge,^{22, 23} CLPM,^{24, 25} VirtualMSLab,²⁶ MS2Assign,²⁷ X-Link,²⁸ MS3D,²⁹ FindLink,³⁰ CrossSearch,³¹ SearchXLinks (optimized for ISD spectra of peptides with disulfide bonds).^{32, 33} These programs compute and list all of the possible peptide masses.

However, when the list of candidate proteins grows, the total number of possible inter-peptide cross-links (two peptides connected with a cross-linker) grows in proportion to N^2 where N is the total number of individual peptides. As a result, the computational complexity is enormous for any sample consisting of more than just a few proteins. An additional challenge results from the very low abundance of cross-linked peptides in samples compared to underivatized peptides. Therefore, identifying inter-peptide cross-links from tandem mass spectrometry data is extremely challenging, and often this search is considered a looking-for-a-needle-in-a-haystack type of problem. Since the aforementioned software packages were not designed to handle this complexity, various experimental approaches have been developed to reduce data complexity or simplify the analysis, including: (1) isotopically labelling cross-linkers or proteins;³⁴⁻⁴² (2) designing special MS/MS cleavable cross-linkers;⁴³⁻⁴⁶ (3) incorporating affinity tags into the cross-linker for enrichment of cross-linked peptides.^{12, 40, 42, 47, 48} Some of these techniques for performing cross-linking experiments have been reviewed by Sinz.⁸

Isotopic labelling produces a characteristic mass shift easily detected by mass spectrometry. Software that has been developed for analysing the resultant data include Pro-Crosslink,⁴⁹ iXLink/doXLink/XLinkViewer,^{36, 50} and xQuest.³⁵ However, labelling requires complex sample preparation methods and increased sample concentration compared to a label-free approach.⁵¹ In addition to experimental limitations, there are limitations in data analysis as well. Rinner et al. have performed the most comprehensive cross-linking analysis to date

against the entire *E. Coli* proteome by ^{18}O -labeling the cross-linker and developed xQuest.³⁵ To identify cross-linked peptides, the tandem (MS/MS) mass spectra that correspond to the light and heavy version of the same precursor are compared. Common ions (fragment ions that are present in both spectra) and cross-linked ions (fragment ions that show a characteristic isotopic shift between the two spectra) are extracted and then used to identify the cross-linked peptide. However, fragmentation of an inter-peptide cross-link often produces more ions than a underivatized peptide and many of these ions can have similar m/z values. Therefore, it is difficult to unambiguously differentiate common and cross-linked ions.

Another very important aspect is that fragmentation of cross-linked peptides result in difficult-to-interpret fragment ion mass spectra. An alternative experimental approach to reduce data analysis complexity is to employ cross-linkers that allow cleavage by low energy collision induced dissociation (CID) and then release of cross-linked peptide chains without disrupting peptide backbones. PIR (Protein Interaction Reporter), developed by Bruce and coworkers, is a cross-linker that has two MS/MS labile bonds between each cross-linked peptide, allowing the production of signature fragments in CID.⁴³⁻⁴⁵ Data analysis is accomplished by using X-links.⁵² A different cross-linker that was developed by the Goshe and coworkers contains a labile Asp-Pro in the linker, which could be fragmented in-source prior to MS/MS analysis.^{53, 54} This approach enables data analysis using commercial software (SEQUEST) that is used for identifying underivatized peptides. A third cleavable cross-linker was recently reported by Schafer and coworkers.⁵⁵ In spite of the success of these approaches, a critical challenge that remains is unambiguous assignment of the location of cross-linking sites, especially for inter-peptide cross-links.

Due to the limitations of the aforementioned experimental approaches, we decided that a label-free approach would hold the most promise for large-scale identifications of PPIs as long as the data analysis bottleneck could be overcome. A custom-designed cross-linker called CLIP (click-enabled cross-linker for interacting proteins) opens doors to applying label-free CXMS in large-scale analysis.⁵⁶ CLIP is small in size and enables enrichment of cross-linked species by use of click chemistry. The latter functionality reduces the data complexity considerably. In conjunction with the development of CLIP, we have developed a data analysis platform, Xlink-Identifier, to identify peptides from tandem mass spectra. Compared to existing algorithms and software tools, Xlink-Identifier has the following advantages: 1) It is equipped with the capability to process and search both CID and ETD tandem mass spectra; 2) Unlike the approach that is adopted by Maiolica et al.³⁹ and Panchaud et al. (The software tool is called xComb)⁵⁷ by searching a database of linearized sequences of two cross-linked peptides, Xlink-Identifier directly fragments inter-peptide cross-links and can take into account fragment ions that result from multiple fragmentation events; 3) Xlink-Identifier can identify inter-peptide, intra-peptide (two amino acids within the same peptide are connected by the cross-linker), and deadend cross-links (one end of the cross-linker is connected to a peptide and the other end is hydrolyzed), and underivatized peptides (Figure 1). It employs a single, universal scoring mechanism for identifying these four types of peptides, which enabled the selection of the best peptide when multiple types of peptides match a single tandem mass spectrum; 4) Xlink-Identifier can identify peptides with high charge states. Inter-peptide cross-links can carry more charges than underivatized peptides, and intra-peptide and deadend cross-links. Unlike X!Link^{19, 58} that only accepts doubly and triply charged fragment ions, Xlink-Identifier produces fragment ions whose charge state can be as high as that of the precursor ion. As a result, it can identify inter-peptide cross-links of high charge states; 5) Xlink-Identifier was designed for general-purpose analyses in that the cross-linker can be amine-reactive or specific to other amino acid(s), which is an advantage compared to X!Link that only considers lysine reactive cross-linkers,^{19, 58} and 6) It features a denoising algorithm that effectively removes noise peaks

from the spectra and a visualization component that displays the quality of the match between the experimental and theoretical spectra in the form of an HTML page. In summary, Xlink-Identifier streamlines data pre-processing, peptide scoring, and visualization and provides an overall data analysis strategy for studying PPIs and protein structure using mass spectrometry. We describe its algorithms and data analysis workflow in this report. Xlink-Identifier is available at <http://www.du-lab.org>.

Experimental procedures

Cross-linking of ubiquitin, digestion, and LC-MS analysis

The cross-linking reagent CLIP (chemical structure depicted in Figure 2) was utilized. Detailed experimental procedures for synthesizing CLIP, preparing samples, and performing mass spectrometry analysis have been reported by Chowdhury et al.^{56, 59} In brief, cross-linking and enrichment reagent stock solutions were prepared in DMSO (100 mM). The cross-linking reaction was performed utilizing a 1:25 protein-to-cross-linker ratio. The reaction was allowed to proceed for 30 min after which the reaction was quenched with 50 μ L of 50mM Tris-HCL (pH 8.0). Excess cross-linker was removed. Affinity purification of cross-linked peptides was achieved after click labelling with a biotinylated azide and subsequent purification with biotin-avidin affinity chromatography.

In-solution digestion was performed by adding trypsin (Promega, Madison, WI) to the solution at a 1:50 protease-to-protein ratio. Proteolysis was conducted at 37 °C for 6 h, and the trypsin digestion was stopped by using 1% trifluoroacetic acid (TFA) in water.

LC-MS/MS analyses were performed using a custom LC platform coupled to an LTQ mass spectrometer (Thermo Fisher Scientific, San Jose, CA). Data-dependent data sets were collected for the four most abundant species after each MS scan using sequential CID and ETD collision modes.

Methods and Results

Types of cross-linked peptides

Three distinct types of cross-linked peptides can result from the proteolysis of proteins after reaction with cross-linking reagents. These are inter-peptide, intra-peptide, and deadend cross-links. Figure 1C depicts the cross-linking of two independent peptide chains called an inter-peptide cross-link. For inter-peptide cross-links, Schilling et al.²⁷ proposed a nomenclature for fragments generated from dissociation of cross-linked peptides, based on the previous nomenclature for underivatized peptides proposed by Roepstorff and Fohlman⁶⁰ and modified by Biemann.⁶¹ Schilling proposed that the longer peptide chain is named the α -chain and the shorter peptide chain as the β -chain. In cases where both peptide chains contain the same number of amino acids, the chain with the higher molecular weight is called the α -chain in contrast to the lighter β -chain. In cases where the two peptides have exactly the same sequence and same cross-linking site, either one can be named the α -chain with the other one being the β -chain without causing any confusion. Figure 1B depicts a linear peptide with two modified residues called an intra-peptide cross-link. Figure 1A depicts a deadend cross-linked peptide where a linear peptide is singly modified with a hydrolyzed or unreacted cross-linker (i.e., only one end of the cross-linker is connected to a peptide).

An inter-peptide cross-link is potentially the most valuable because it gives information about two amino acids that are close in proximity but could be far apart in primary sequence on the same protein or that are even on two different proteins.⁶² Deadend cross-links do not provide any amino acid-to-amino acid distance information about a protein structure, but can

yield important information concerning relative reactivities at various sites on a protein. In addition, indirect structural information could still be obtained from the fact that only surface-exposed amino acids are reactive to cross-linkers.

Data analysis procedure

Supplementary Figure S1 depicts a flow chart of the data analysis pipeline designed to efficiently identify cross-linked and underivatized peptides. The raw data is first processed to obtain each MS/MS spectrum as well as the corresponding precursor ion molecular mass and charge state. Each MS/MS spectrum is then denoised (to be described) and the identification of underivatized, deadend, intra-peptide, and inter-peptide cross-links follows.

For each MS/MS spectrum, multiple identifications can occur, including identifications of peptides having different cross-linking types. The search result analysis component of the pipeline then compares all of the matches and selects the most confident identifications. Criteria for selecting the best identification include the mass difference between the theoretical and experimental molecular weight and match scores that are indicative of the similarity between experimental and theoretical spectra.

The basic procedure, illustrated in Supplementary Figure S2, to identify each of the four types of peptides, is very similar and we use the identification of inter-peptide cross-links to explain. First, all of the candidate proteins are digested *in silico*. Digestion parameters that need to be specified include: the specificity of the enzyme used in digestion (default is trypsin), amino acids that the cross-linker can react with, the molecular masses of the reacted (for inter-peptide and intra-peptide cross-links) and hydrolyzed (for deadend cross-links) cross-linker, the maximum number of allowed missed proteolytic cleavages, the precursor mass error tolerance, and the fragment ion mass error tolerance. Candidate proteins are digested based on these preset parameters. For cross-linked species, all possible combinations are enumerated. Therefore, as the protein database increases, the search space for identifying inter-peptide cross-links increases rapidly.

All possible inter-peptide cross-links are then tabulated and indexed by molecular mass using the peptides from the *in silico* protein digestion. Relevant information for each entry includes the sequences and originating protein of both peptides, the location of the two peptides within their corresponding proteins, and the total molecular weight. If one or both of the peptides contain(s) multiple amino acids reactive with the cross-linker, different combinations of the cross-linking sites will produce different inter-peptide cross-links. This combinatorial nature of linking sites and peptide pairs is the source of the enormous computational complexity.

Next, for each denoised tandem mass spectrum, the top peptide matches are determined. Specifically, a list is assembled of the candidate peptides that have molecular masses within a pre-specified tolerance of the precursor mass. For each candidate a theoretical fragmentation spectrum is generated by calculating the masses of all possible ions resulting from a single fragmentation event. For a deadend cross-linked peptide, cross-linking is treated as a modification. For an intra-peptide cross-link, the circular sub-sequence between the two linking sites is treated as one single unit because a single fragmentation anywhere along the peptide backbone will result in species of equal mass. The molecular mass of the unit is the mass of the constituent amino acids plus the cross-linker, and then the peptide is treated as an (otherwise) underivatized peptide. For an inter-peptide cross-link, the theoretical fragmentation is more complex and we present the details in the next section. For intra-peptide and deadend cross-links and underivatized peptides, the theoretical fragmentation spectrum is determined by calculating the b and y ion series for CID fragmentation and by calculating the c and z ion series from ETD fragmentation.^{63, 64} With

the theoretical spectrum generated, a score is calculated to quantify the match quality between the experimental and the theoretical spectrum. For the top matching peptides, additional theoretical fragment ions are calculated and checked for matching with peaks in the experimental spectrum. For CID spectra, these ions include a type ions that are degraded from b ions, H₂O and NH₃ neutral-losses from b and y ions, and ions corresponding to loss of the neutral-loss reporter tag from the precursor if the cross-linker features this type of a tag. This post-scoring checking is a critical step because experimental peaks that are not matched with any of the b or y ions might be explained by these additional ions.

Theoretical fragment ions of inter-peptide cross-links

In fragmenting an inter-peptide cross-link, one or multiple cleavage(s) can occur along the peptide backbone.²⁷ Figure 3 depicts ions resulting from one or two cleavage events of a 13-amino acid peptide coupled to a 9-amino acid peptide. In the example shown in Figure 3A, two ions, b_{2α} and y_{11α}, are generated when a single cleavage occurs. y_{11α} contains the C-terminal ion of peptide α, the entire β peptide, and the cross-linker. When two cleavages occur simultaneously, three fragments are produced as depicted in Figure 3B. Ions b_{2β} and y_{4α} are formed from sub-sequences of peptide β and α, respectively. Ion b_{9α}y_{7β} is formed from part of peptide α, part of peptide β, and the cross-linker. This nomenclature was proposed by Schilling et al.²⁷

To determine the theoretical spectrum, the series of b and y ions are then calculated by fragmenting the two peptides along their backbone. Ions produced from one cleavage are used in scoring whereas ions produced from two cleavages are not, in order to avoid overestimating the match score. This overestimate can occur because a higher number of theoretical ions increase the probability that a peak in the experimental spectrum is matched. After the top scoring peptides are selected, auxiliary ions that are produced from two simultaneous cleavages, a type ions that are degraded from b-ions, and fragment ions created by neutral losses of water and ammonia are checked for matching with the experimental spectrum in addition to the b- and y-ion series. This post-scoring checking provides a more comprehensive picture about the set of theoretical ions that are observed in the experimental spectrum and facilitates comparison of peptide candidates that give rise to similar matching scores. Because this post-scoring checking is performed only for peptide candidates having a sufficient number of theoretical b- and y-type ions that are observed in the experimental spectrum, a cross-linking product will not be assigned if only auxiliary ions, but no b- and y-type ions, are observed.

Denoising of tandem mass spectra

Efficient removal of noise peaks is essential for accurately identifying inter-peptide cross-links. The reasons are two-fold. Inter-peptide cross-links contain two independent peptides with a total of two basic tryptic C-termini and thus tend to carry more charges. Consequently, tandem mass spectra from inter-peptide cross-links are generally noisier than those from underivatized, inter-peptide, or deadend cross-links.

A number of algorithms for denoising mass spectra have been reported.⁶⁵⁻⁶⁷ However, most of the algorithms were not specifically tailored for tandem mass spectra from highly charged species. For example, the algorithm reported by Ding et al. considered only charge states of 1+ and 2+ for peaks in a tandem mass spectrum⁶⁷ while the MEND algorithm was mainly designed for denoising LC-MS spectra.⁶⁵ Therefore, there is a need to develop an algorithm for denoising tandem mass spectra from inter-peptide cross-links with high charge states.

Considering that different m/z regions in tandem mass spectra usually have different noise backgrounds, we divided each spectrum into regions of 100 A.M.U each. Within each

region, the signal-to-noise ratio (SNR) was estimated in an iterative fashion until it converges. During each iteration, peaks exceeding one standard deviation from the mean were removed and the mean of the remaining peaks was calculated. After all of the major signal peaks have been removed, the mean of the remaining peaks should not change significantly during subsequent iterations compared to the changes calculated during the previous iterations. The iteration stops when this change is less than a specified threshold. All peaks that were removed are considered signal peaks and all those remaining were considered noise. The SNR is then estimated as the intensity ratio of the highest intensity signal peak to the mean of the noise peaks. If this SNR exceeds a specified threshold, all of the signal peaks will be used for scoring. Otherwise denoising is assumed to have failed and only peaks that exceed a predetermined percentage of the overall median of all of the raw peaks across the entire spectrum are used for scoring. Supplementary Figure S3 depicts the effect of denoising on a particular MS/MS spectrum.

Denoising ETD experimental spectra requires an additional step. In an ETD MS/MS spectrum, precursor ions that are partially neutralized by transferred electrons can show up as intense peaks. These ions do not provide any information about the peptide sequence and thus are excluded from scoring. However, these peaks are later checked for matching with theoretical ions to add confidence to the identification.

Scoring of match quality between experimental and theoretical tandem mass spectra

For each candidate peptide corresponding to an experimental tandem spectrum, its theoretical spectrum is produced and a score is calculated to quantify the similarity between the theoretical and experimental spectra. The theoretical spectrum consists of peaks of b and y ion series with the same intensity. The intensity of the experimental spectrum is linearly scaled so that the maximum intensity equals 100. Both spectra are binned with a bin width of 1 Da and the cross-correlation between them is computed. The cross-correlation is a function of the offset (e.g. shift) between the two spectra and the cross-correlation at each offset is basically the dot product between the two spectral signals. The final score, termed XlinkScore, is calculated as the dot product at offset equal to zero after subtracting the mean of the dot product at offsets from -75 to 75 excluding offset zero. The subtraction functions as a baseline correction of the cross correlation around offset zero and was used in the calculation of XCorr in Sequest.⁶⁸

This calculation of XlinkScore bears similarities with that of XCorr in Sequest on two aspects. Firstly, both of them use dot product as a measure of similarity between the experimental and theoretical spectrum. Dot product is a commonly used technique in signal processing to quantify the similarity between two signals. The higher the similarity, the larger the dot product. This measure of similarity has also been used very widely to identify compounds from gas chromatography mass spectrometry data, i.e. GC-MS spectra.⁶⁹ Secondly, both of the final scores result from subtracting the mean of dot products between offset -75 and 75 from the dot product at offset zero. This subtraction facilitates selecting the peptide candidate that truly stands out from other possible candidates. An alternative scoring method is to calculate the expectation score that compares the top-scoring peptide scores with the complete distribution of such scores for the MS/MS spectrum. However, the latter approach will further add a considerable amount of computation to the scoring algorithm that is already computationally challenged for inter-peptide cross-link searches and therefore has not been explored in the current version of the software.

Results

Xlink-Identifier identified two inter-peptide and three dead-end cross-links, respectively, using search parameters listed in Table 1. Figure 4A,B depicts spectra from an inter-peptide

cross-link that was identified from two sequential CID and ETD spectra, respectively. The two constituent peptides have the same sequence and presumably originated from a homomultimer. In the ETD spectrum (Figure 4B), peaks that corresponded to partially neutralized precursor ions are of high intensity compared to the other peaks and thus it is important to exclude them during scoring. We have also identified an inter-peptide cross-link LIFAGK[^]QLEDGR--TSLDYNIQK[^]ESTLHLVLR from two sequential CID and ETD spectra with charge state 5+ and more details about this identification can be found in Chowdhury et al.⁵⁶ The dead-end cross-links that were identified are: LIFAGK[^]QLEDGR, TSLDYNIQK[^]ESTLHLVLR, and MQIFVK[^]TLTGK. The distance constraints obtained from these inter-peptide and dead-end cross-links are consistent with the crystal structure of ubiquitin.⁵⁶ The symbol “^” in the peptide sequences indicates that the preceding lysine residue has reacted with the cross-linking reagent.

Both of the two aforementioned inter-peptide cross-links have the highest normalized XlinkScore among the identifications with the same precursor charge state (all of the identifications of inter-peptide cross-links from CID and ETD spectra with charge states 4+ and 5+ are provided in Supplementary File S1). Table 2 shows the actual numeric values of the normalized scores for these two identifications. The normalization is accomplished by dividing the natural logarithm of the XlinkScore by the natural logarithm of the total peptide length of the two cross-linked peptides. This type of normalization has been applied by PeptideProphet⁷⁰ and other software tools⁷¹ to correct the bias that longer peptides tend to produce larger matching scores.

In order to estimate the false discovery rate (FDR) of the identifications of inter-peptide cross-links, we applied the target-decoy approach by searching the experimental MS/MS spectra against the reverse ubiquitin sequence.⁷² The resulting identifications are provided in Supplementary File S1. A comparison between the identifications from forward and reverse search results revealed that: 1) the normalized XlinkScores of the two aforementioned inter-peptide cross-links are higher than those of the top-most identifications in the corresponding reverse search. This is indicative of the capability of the scoring algorithm of Xlink-Identifier to select high quality matches, 2) when the difference of normalized XlinkScores is small between the top-most matches from the forward and reverse search, the identifications from the forward search explain more of the fragment ions than the top-most matches in the reverse searches. This is reflected in the total number of matched b and y ions in the case of CID fragment spectra or c and z ions in the case of ETD spectra.

Figure 5 depicts the histogram of the normalized XlinkScore for inter-peptide cross-links under CID and ETD fragmentation mechanisms with charge states 4+ and 5+. The green and red curves correspond to searches against the forward and reverse ubiquitin sequences, respectively. All of the inter-peptide cross-linked peptides were used except those identifications with delta mass (i.e. the mass difference between the experimental and theoretical molecular mass) exceeding 300ppm. Based on these search results, FDR of inter-peptide cross-link identifications can be readily estimated for different threshold of the normalized XlinkScore using the equation described by Elias et al.⁷² Since we deem only the top-most identifications are of high confidence and they have been verified using the crystal structure of ubiquitin, the FDR would be zero when only the top-most identifications are selected and would increase quickly when the threshold of normalized XlinkScore is gradually reduced.

For this particular experiment, Xlink-Identifier did not identify other inter-peptide cross-links that are of high confidence. The same sample was also analyzed using SDA-PAGE. The higher molecular band was excised from the gel, digested with trypsin, and analyzed by LC-MS/MS. The resulting identifications of cross-linked species have been reported in a

previous publication⁵⁶ and are almost the same as what have been identified from the in-solution digestion analysis reported in this paper. The total number of confident cross-linked species is small from both experiments and the reasons could be two-fold: 1) ubiquitin has only seven lysine residues and thus did not produce many cross-linked candidates, 2) the efficiency of CLIP to react with lysine residues was not very high.

Performance

Xlink-Identifier was prototyped in Matlab and then recoded into C++. It took the C++ version about 50 sec to process the 19,429 MS/MS spectra and search them against the single protein ubiquitin in the fasta file. Additionally, the C++ version of XlinkIdentifier could handle over a hundred proteins.

Xlink-Identifier is equipped with the capability to display match details between the experimental and theoretical fragment spectra. The display (Supplementary File S4) includes the annotated MS/MS experimental spectrum and tables of matched fragment ions. This capability greatly facilitates visual verification of identifications. In particular, users can quickly check the continuity of matched fragment ions by looking at the tables.

Currently, Xlink-Identifier takes MS/MS spectra in the .dta format. As a result, it is able to handle experimental data from other mass spectrometers as long as .dta files can be produced from the raw data. If MS/MS spectra are acquired with high measurement resolution, the m/z tolerance in the scoring process should be reduced to take advantage of the higher quality of the raw data.

Xlink-Identifier was originally developed for analyzing data from cross-linking experiment with CLIP. However, Xlink-Identifier could be readily applied when other cross-linking reagents are used and this can be accomplished by changing the molecular weight of the reagent in the parameter file.

Conclusions

Chemical cross-linking combined with mass spectrometry provides a very powerful approach for identifying PPIs and for studying the structures of proteins in general. Currently the major challenges to the use of this approach include: (1) synthesis of efficient cross-linkers that are small in size, membrane permeable, and feature an enrichment tag, and (2) efficient data analysis software tools that can handle the computational complexity associated with searching against vast number of candidates of inter-peptide cross-linked species. This report presents our efforts to address the second challenge by developing Xlink-Identifier. Xlink-Identifier is a comprehensive software package that can identify deadend, intra-peptide, and inter-peptide cross-links and underivatized peptides without manual intervention. It streamlines data pre-processing, peptide scoring, and visualization and provides an overall data analysis strategy for studying protein-protein interactions and protein structures using mass spectrometry.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported, in part, by the startup fund from the University of North Carolina at Charlotte, the Laboratory Directed Research and Development program at Pacific Northwest National Laboratory (PNNL), and the NIH National Center for Research Resources (RR18522). This work utilized data generated on instrumentation and capabilities developed under support from the National Center for Research Resources (Grant RR 018522 to

RDS) and the DOE's Office of Biological and Environmental Research. Part of this work was performed in the Environmental Molecular Science Laboratory, a U.S. Department of Energy (DOE) national scientific user facility located at PNNL (Richland, WA). Battelle Memorial Institute operates PNNL for the DOE under Contract DE-AC05-76RLO01830.

References

1. Zuiderweg ER. Mapping protein-protein interactions in solution by NMR spectroscopy. *Biochemistry* 2002;41(1):1–7. [PubMed: 11771996]
2. Scott EE, White MA, He YA, Johnson EF, Stout CD, Halpert JR. Structure of mammalian cytochrome P450 2B4 complexed with 4-(4-chlorophenyl)imidazole at 1.9-Å resolution: insight into the range of P450 conformations and the coordination of redox partner binding. *J Biol Chem* 2004;279(26):27294–301. [PubMed: 15100217]
3. Shen Z, Cloud KG, Chen DJ, Park MS. Specific interactions between the human RAD51 and RAD52 proteins. *J Biol Chem* 1996;271(1):148–52. [PubMed: 8550550]
4. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 2002;417(6887):399–403. [PubMed: 12000970]
5. Howell JM, Winstone TL, Coorsen JR, Turner RJ. An evaluation of in vitro protein-protein interaction techniques: assessing contaminating background proteins. *Proteomics* 2006;6(7):2050–69. [PubMed: 16518870]
6. Guerrero C, Tagwerker C, Kaiser P, Huang L. An integrated mass spectrometry-based proteomic approach: quantitative analysis of tandem affinity-purified in vivo cross-linked protein complexes (QTAX) to decipher the 26 S proteasome-interacting network. *Mol Cell Proteomics* 2006;5(2):366–78. [PubMed: 16284124]
7. Orlando V, Strutt H, Paro R. Analysis of chromatin structure by in vivo formaldehyde cross-linking. *Methods* 1997;11(2):205–14. [PubMed: 8993033]
8. Sinz A. Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein-protein interactions. *Mass Spectrom Rev* 2006;25(4):663–82. [PubMed: 16477643]
9. Gingras AC, Gstaiger M, Raught B, Aebersold R. Analysis of protein complexes using mass spectrometry. *Nat Rev Mol Cell Biol* 2007;8(8):645–54. [PubMed: 17593931]
10. Back JW, de Jong L, Muijsers AO, de Koster CG. Chemical cross-linking and mass spectrometry for protein structural modeling. *J Mol Biol* 2003;331(2):303–13. [PubMed: 12888339]
11. Sinz A. Chemical cross-linking and mass spectrometry for mapping three-dimensional structures of proteins and protein complexes. *J Mass Spectrom* 2003;38(12):1225–37. [PubMed: 14696200]
12. Sinz A, Kalkhof S, Ihling C. Mapping protein interfaces by a trifunctional cross-linker combined with MALDI-TOF and ESI-FTICR mass spectrometry. *J Am Soc Mass Spectrom* 2005;16(12):1921–31. [PubMed: 16246579]
13. Novak P, Haskins WE, Ayson MJ, Jacobsen RB, Schoeniger JS, Leavell MD, Young MM, Kruppa GH. Unambiguous assignment of intramolecular chemical cross-links in modified mammalian membrane proteins by Fourier transform-tandem mass spectrometry. *Anal Chem* 2005;77(16):5101–6. [PubMed: 16097745]
14. Dihazi GH, Sinz A. Mapping low-resolution three-dimensional protein structures using chemical cross-linking and Fourier transform ion-cyclotron resonance mass spectrometry. *Rapid Commun Mass Spectrom* 2003;17(17):2005–14. [PubMed: 12913864]
15. Huang BX, Kim HY, Dass C. Probing three-dimensional structure of bovine serum albumin by chemical cross-linking and mass spectrometry. *J Am Soc Mass Spectrom* 2004;15(8):1237–47. [PubMed: 15276171]
16. Pearson KM, Pannell LK, Fales HM. Intramolecular cross-linking experiments on cytochrome c and ribonuclease A using an isotope multiplet method. *Rapid Commun Mass Spectrom* 2002;16(3):149–59. [PubMed: 11803535]
17. Trester-Zedlitz M, Kamada K, Burley SK, Fenyo D, Chait BT, Muir TW. A modular cross-linking approach for exploring protein interactions. *J Am Chem Soc* 2003;125(9):2416–25. [PubMed: 12603129]

18. Young MM, Tang N, Hempel JC, Oshiro CM, Taylor EW, Kuntz ID, Gibson BW, Dollinger G. High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proc Natl Acad Sci U S A* 2000;97(11):5802–6. [PubMed: 10811876]
19. Lee YJ, Lackner LL, Nunnari JM, Phinney BS. Shotgun cross-linking analysis for studying quaternary and tertiary protein structures. *J Proteome Res* 2007;6(10):3908–17. [PubMed: 17854217]
20. Bennett KL, Kussmann M, Bjork P, Godzwon M, Mikkelsen M, Sorensen P, Roepstorff P. Chemical cross-linking with thiol-cleavable reagents combined with differential mass spectrometric peptide mapping--a novel approach to assess intermolecular protein contacts. *Protein Sci* 2000;9(8):1503–18. [PubMed: 10975572]
21. GPMaw. <http://gpmaw.com/GPMaw/gpmaw.html>
22. Chu F, Shan SO, Moustakas DT, Alber F, Egea PF, Stroud RM, Walter P, Burlingame AL. Unraveling the interface of signal recognition particle and its receptor by using chemical cross-linking and tandem mass spectrometry. *Proc Natl Acad Sci U S A* 2004;101(47):16454–9. [PubMed: 15546976]
23. MSBridge. <http://prospector.ucsf.edu>
24. Tang Y, Chen Y, Lichti CF, Hall RA, Raney KD, Jennings SF. CLPM: a cross-linked peptide mapping algorithm for mass spectrometric analysis. *BMC Bioinformatics* 2005;6(Suppl 2):S9. [PubMed: 16026606]
25. CLPM. <http://bioinformatics.ualr.edu/mbc/services/CLPM.html>
26. de Koning LJ, Kasper PT, Back JW, Nessen MA, Vanrobaeys F, Van Beeumen J, Gherardi E, de Koster CG, de Jong L. Computer-assisted mass spectrometric analysis of naturally occurring and artificially introduced cross-links in proteins and protein complexes. *FEBS J* 2006;273(2):281–91. [PubMed: 16403016]
27. Schilling B, Row RH, Gibson BW, Guo X, Young MM. MS2Assign, automated assignment and nomenclature of tandem mass spectra of chemically crosslinked peptides. *J Am Soc Mass Spectrom* 2003;14(8):834–50. [PubMed: 12892908]
28. Taverner T, Hall NE, O'Hair RA, Simpson RJ. Characterization of an antagonist interleukin-6 dimer by stable isotope labeling, cross-linking, and mass spectrometry. *J Biol Chem* 2002;277(48):46487–92. [PubMed: 12235153]
29. Yu ET, Hawkins A, Kuntz ID, Rahn LA, Rothfuss A, Sale K, Young MM, Yang CL, Pancerella CM, Fabris D. The collaboratory for MS3D: a new cyberinfrastructure for the structural elucidation of biological macromolecules and their assemblies using mass spectrometry-based approaches. *J Proteome Res* 2008;7(11):4848–57. [PubMed: 18817429]
30. Back JW, Notenboom V, de Koning LJ, Muijsers AO, Sixma TK, de Koster CG, de Jong L. Identification of cross-linked peptides for protein interaction studies using mass spectrometry and ¹⁸O labeling. *Anal Chem* 2002;74(17):4417–22. [PubMed: 12236350]
31. Nadeau OW, Wyckoff GJ, Paschall JE, Artigues A, Sage J, Villar MT, Carlson GM. CrossSearch, a user-friendly search engine for detecting chemically cross-linked peptides in conjugated proteins. *Mol Cell Proteomics* 2008;7(4):739–49. [PubMed: 18281724]
32. Schnaible V, Wefing S, Resemann A, Suckau D, Bucker A, Wolf-Kummeth S, Hoffmann D. Screening for disulfide bonds in proteins by MALDI in-source decay and LIFT-TOF/TOF-MS. *Anal Chem* 2002;74(19):4980–8. [PubMed: 12380820]
33. Wefing S, Schnaible V, Hoffmann D. SearchXLinks. A program for the identification of disulfide bonds in proteins from mass spectra. *Anal Chem* 2006;78(4):1235–41. [PubMed: 16478117]
34. Petrotchenko EV, Olkhovik VK, Borchers CH. Isotopically coded cleavable cross-linker for studying protein-protein interaction and protein complexes. *Mol Cell Proteomics* 2005;4(8):1167–79. [PubMed: 15901824]
35. Rinner O, Seebacher J, Walzthoeni 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]

36. Seebacher J, Mallick P, Zhang N, Eddes JS, Aebersold R, Gelb MH. Protein cross-linking analysis using mass spectrometry, isotope-coded cross-linkers, and integrated computational data processing. *J Proteome Res* 2006;5(9):2270–82. [PubMed: 16944939]
37. Ihling C, Schmidt A, Kalkhof S, Schulz DM, Stingl C, Mechtler K, Haack M, Beck-Sickinger AG, Cooper DM, Sinz A. Isotope-labeled cross-linkers and Fourier transform ion cyclotron resonance mass spectrometry for structural analysis of a protein/peptide complex. *J Am Soc Mass Spectrom* 2006;17(8):1100–13. [PubMed: 16750914]
38. Muller DR, Schindler P, Towbin H, Wirth U, Voshol H, Hoving S, Steinmetz MO. Isotope-tagged cross-linking reagents. A new tool in mass spectrometric protein interaction analysis. *Anal Chem* 2001;73(9):1927–34. [PubMed: 11354472]
39. Maiolica A, Cittaro D, Borsotti D, Sennels L, Ciferri C, Tarricone C, Musacchio A, Rappsilber J. Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching. *Mol Cell Proteomics* 2007;6(12):2200–11. [PubMed: 17921176]
40. Chu F, Mahrus S, Craik CS, Burlingame AL. Isotope-coded and affinity-tagged cross-linking (ICATXL): an efficient strategy to probe protein interaction surfaces. *J Am Chem Soc* 2006;128(32):10362–3. [PubMed: 16895390]
41. Lamos SM, Krusemark CJ, McGee CJ, Scalf M, Smith LM, Belshaw PJ. Mixed isotope photoaffinity reagents for identification of small-molecule targets by mass spectrometry. *Angew Chem Int Ed Engl* 2006;45(26):4329–33. [PubMed: 16733838]
42. Sinz A. Isotope-Labeled Photoaffinity Reagents and Mass Spectrometry To Identify Protein-Ligand Interactions. *Angew Chem Int Ed Engl* 2006;46(5):660–662. [PubMed: 17167803]
43. Tang X, Munske GR, Siems WF, Bruce JE. Mass spectrometry identifiable cross-linking strategy for studying protein-protein interactions. *Anal Chem* 2005;77(1):311–8. [PubMed: 15623310]
44. Zhang H, Tang X, Munske GR, Zakharova N, Yang L, Zheng C, Wolff MA, Tolic N, Anderson GA, Shi L, Marshall MJ, Fredrickson JK, Bruce JE. In vivo identification of the outer membrane protein OmcA-MtrC interaction network in *Shewanella oneidensis* MR-1 cells using novel hydrophobic chemical cross-linkers. *J Proteome Res* 2008;7(4):1712–20. [PubMed: 18303833]
45. Zhang H, Tang X, Munske GR, Tolic N, Anderson GA, Bruce JE. Identification of protein-protein interactions and topologies in living cells with chemical cross-linking and mass spectrometry. *Mol Cell Proteomics* 2009;8(3):409–20. [PubMed: 18936057]
46. Chowdhury SM, Munske GR, Tang X, Bruce JE. Collisionally activated dissociation and electron capture dissociation of several mass spectrometry-identifiable chemical cross-linkers. *Anal Chem* 2006;78(24):8183–93. [PubMed: 17165806]
47. Hurst GB, Lankford TK, Kennel SJ. Mass spectrometric detection of affinity purified crosslinked peptides. *J Am Soc Mass Spectrom* 2004;15(6):832–9. [PubMed: 15144972]
48. Liu B, Archer CT, Burdine L, Gillette TG, Kodadek T. Label transfer chemistry for the characterization of protein-protein interactions. *J Am Chem Soc* 2007;129(41):12348–9. [PubMed: 17894490]
49. Gao Q, Xue S, Doneanu CE, Shaffer SA, Goodlett DR, Nelson SD. Pro-CrossLink. Software tool for protein cross-linking and mass spectrometry. *Anal Chem* 2006;78(7):2145–9. [PubMed: 16579592]
50. iXLINK. <http://tools.proteomecenter.org/XLink.php>
51. Patel VJ, Thalassinou K, Slade SE, Connolly JB, Crombie A, Murrell JC, Scrivens JH. A Comparison of Labeling and Label-Free Mass Spectrometry-Based Proteomics Approaches. *J Proteome Res*. 2009
52. Anderson GA, Tolic N, Tang X, Zheng C, Bruce JE. Informatics strategies for large-scale novel cross-linking analysis. *J Proteome Res* 2007;6(9):3412–21. [PubMed: 17676784]
53. Soderblom EJ, Goshe MB. Collision-induced dissociative chemical cross-linking reagents and methodology: Applications to protein structural characterization using tandem mass spectrometry analysis. *Anal Chem* 2006;78(23):8059–68. [PubMed: 17134140]
54. Soderblom EJ, Bobay BG, Cavanagh J, Goshe MB. Tandem mass spectrometry acquisition approaches to enhance identification of protein-protein interactions using low-energy collision-induced dissociative chemical crosslinking reagents. *Rapid Commun Mass Spectrom* 2007;21(21):3395–408. [PubMed: 17902198]

55. Dreiocker F, Muller MQ, Sinz A, Schafer M. Collision-induced dissociative chemical cross-linking reagent for protein structure characterization: applied Edman chemistry in the gas phase. *J Mass Spectrom* 45(2):178–89. [PubMed: 19950134]
56. Chowdhury SM, Du X, Tolic N, Wu S, Moore RJ, Mayer MU, Smith RD, Adkins JN. Identification of cross-linked peptides after click-based enrichment using sequential collision-induced dissociation and electron transfer dissociation tandem mass spectrometry. *Anal Chem* 2009;81(13):5524–32. [PubMed: 19496583]
57. Panchaud A, Singh P, Shaffer SA, Goodlett DR. xComb: a cross-linked peptide database approach to protein-protein interaction analysis. *J Proteome Res* 9(5):2508–15. [PubMed: 20302351]
58. Singh P, Panchaud A, Goodlett DR. Chemical cross-linking and mass spectrometry as a low-resolution protein structure determination technique. *Anal Chem* 82(7):2636–42. [PubMed: 20210330]
59. Chowdhury SM, Du X, Tolic N, Wu S, Moore RJ, Mayer MU, Smith RD, Adkins JN. Identification of Cross-Linked Peptides after Click-Based Enrichment Using Sequential Collision-Induced Dissociation and Electron Transfer Dissociation Tandem Mass Spectrometry. *Anal Chem*. 2009
60. Roepstorff P, Fohlman J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom* 1984;11(11):601. [PubMed: 6525415]
61. Biemann K. Appendix 5. Nomenclature for peptide fragment ions (positive ions). *Methods Enzymol* 1990;193:886–7. [PubMed: 2074849]
62. Jin Lee Y. Mass spectrometric analysis of cross-linking sites for the structure of proteins and protein complexes. *Mol Biosyst* 2008;4(8):816–23. [PubMed: 18633483]
63. Good DM, Wirtala M, McAlister GC, Coon JJ. Performance characteristics of electron transfer dissociation mass spectrometry. *Mol Cell Proteomics* 2007;6(11):1942–51. [PubMed: 17673454]
64. Coon JJ, Ueberheide B, Syka JE, Dryhurst DD, Ausio J, Shabanowitz J, Hunt DF. Protein identification using sequential ion/ion reactions and tandem mass spectrometry. *Proc Natl Acad Sci U S A* 2005;102(27):9463–8. [PubMed: 15983376]
65. Andreev VP, Rejtar T, Chen HS, Moskovets EV, Ivanov AR, Karger BL. A universal denoising and peak picking algorithm for LC-MS based on matched filtration in the chromatographic time domain. *Anal Chem* 2003;75(22):6314–26. [PubMed: 14616016]
66. Satten GA, Datta S, Moura H, Woolfitt AR, Carvalho Mda G, Carlone GM, De BK, Pavlopoulos A, Barr JR. Standardization and denoising algorithms for mass spectra to classify whole-organism bacterial specimens. *Bioinformatics* 2004;20(17):3128–36. [PubMed: 15217815]
67. Ding J, Shi J, Poirier GG, Wu FX. A novel approach to denoising ion trap tandem mass spectra. *Proteome Sci* 2009;7:9. [PubMed: 19292921]
68. Eng JK, McCormack AL, Yates JR 3rd. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 1994;5:976–989.
69. Stein SE. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry* 1999;10(8):770–781.
70. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 2002;74(20):5383–92. [PubMed: 12403597]
71. Du X, Yang F, Manes NP, Stenoi DL, Monroe ME, Adkins JN, States DJ, Purvine SO, Camp DG 2nd, Smith RD. Linear discriminant analysis-based estimation of the false discovery rate for phosphopeptide identifications. *J Proteome Res* 2008;7(6):2195–203. [PubMed: 18422353]
72. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* 2007;4(3):207–14. [PubMed: 17327847]

Figure 1.

Illustration of three types of cross-linked peptides. Inter-peptide, intra-peptide, and deadend cross-links can all result from the interaction of a cross-linking agent with peptides. These types of cross-linked species are also called type 2, type 1, and type 0 cross-linked peptides, respectively.²⁷ “X” refers to any amino acid. “K” refers to the amino acid that the cross-linker reacts with. For the commonly used amine-reactive cross-linker, this K will be lysine. “N” and “C” refer to the N and C terminus of the peptide, respectively.

CLIP Cross-linker spacer chain 9-10Å

Figure 2.
Chemical structure of the cross-linker CLIP.

Figure 3.

Examples of ions that result from one single fragmentation event or from two simultaneous fragmentation events. Letters A, B, C, and D represent any amino acid. K represents lysine which can react with an amine-reactive cross-linker.

Figure 4. Experimental CID and ETD spectra that match with an inter-peptide cross-link from ubiquitin. Cross-linker CLIP was used. (A) The CID spectrum; (B) The ETD spectrum.

Figure 5. Histogram of the normalized XlinkScore for inter-peptide crosslinks under different fragmentation methods and with different charge states. x-axis corresponds to the normalized XlinkScore. Green and red curves correspond to search results from forward and reverse ubiquitin sequences, respectively. (A) CID, CS = 4+; (B) ETD, CS = 4+; (C) CID, CS = 5+, (D) ETD, CS = 5+.

Table 1

Search parameters

Parameters	Value
precursor mass tolerance	3 Da
fragment ion m/z tolerance	0.6
number of missed cleavages	3
maximum single peptide length	30

Table 2

Normalized XlinkScore

Peptide	CS	Normalized score from CID spectrum	Normalized score from ETD spectrum
LIFAGK^QLEDGR -- LIFAGK^QLEDGR	4+	1.099	0.901
LIFAGK^QLEDGR--TLSDYNIQK^ESTLHLVLR	5+	1.079	0.972