Published in final edited form as:

J Am Soc Mass Spectrom. 2019 September ; 30(9): 1643–1653. doi:10.1007/s13361-019-02253-z.

Identification of crosslinked peptides using isotopomeric crosslinkers

Jie Luo, Jacob Bassett, Jeff Ranish*

Institute for Systems Biology, 401 Terry Ave North, Seattle, WA 98109, USA

Abstract

Chemical crosslinking combined with mass spectrometry (CL-MS) is a powerful method for characterizing the architecture of protein assemblies and for mapping protein–protein interactions. Despite its proven utility, confident identification of crosslinked peptides remains a formidable challenge, especially when the peptides are derived from complex mixtures. MS cleavable crosslinkers are gaining importance for CL-MS as they permit reliable identification of crosslinked peptides by whole proteome database searching using MS/MS information. Here we introduce a novel class of MS cleavable crosslinkers called isotopomeric crosslinkers (ICLs), which allow for confident and efficient identification of crosslinked peptides by whole proteome database searching. ICLs are simple, symmetrical molecules that asymmetrically incorporate heavy and light stable isotopes into the two arms of the crosslinker. As a result of this property, ICLs automatically generate pairs of isotopomeric crosslinked peptides, which differ only by the positions of the heavy and light isotopes. Upon fragmentation during MS analysis, these isotopomeric crosslinked peptides generate unique isotopic doublet ions that correspond to the individual peptides in the crosslink. The doublet ion information is used to determine the masses of the two crosslinked peptides from the same MS² spectrum that is also used for peptide spectrum matching (PSM) by sequence database searching. Here we present the rationale for and mechanism of crosslinked peptide identification by ICL-MS. We describe the synthesis of the ICL-1 reagent, the ICL-MS workflow, and the performance characteristics of ICL-MS for identifying crosslinked peptides derived from increasingly complex mixtures by whole proteome database searching.

Introduction

Chemical crosslinking combined with mass spectrometry (CL-MS) has emerged as a powerful method to study protein-protein interactions (PPIs) and the architecture of protein complexes. CL-MS utilizes chemical crosslinkers to fix the physical state of proteins/complexes, and uses MS technology to identify the two crosslinked peptides/proteins and the

*Address reprint requests to: Jeff Ranish, Institute for Systems Biology, 401 Terry Ave North, Seattle, WA 98109, USA, (p) 206-732-1357, (f) 206-732-1200, jranish@systemsbiology.org.

Supporting Information

Supporting methods for RPB1-TAP purification, crosslinking and mass spectrometry (PDF)

Comparison of database search results for ICL-1 crosslinked peptides from BSA using Nexus2.0 and pLink (.xlsx file).

Summary of the database search results for the Pol II crosslinking experiment (.xlsx file), including the Comet search results for the proteins in the sample, pLink and Nexus2.0 search results for crosslinked peptides, and the output of all Nexus2.0 search results

sites of crosslinking. The identified crosslinked residues then provide site-specific distance constraints that can be used to model the PPIs and the architecture of protein complexes [1, 2].

One of the main challenges associated with CL-MS studies of complex mixtures of proteins is the confident identification of the resulting crosslinked peptides. Two main issues contribute to this problem. The first is that, with most crosslinkers, it is not possible to distinguish MS2 spectra derived from crosslinked peptides from MS2 spectra derived from all other types of peptides including monolinks, looplinks and unmodified peptides. All MS2 spectra are considered as candidate spectra from crosslinked peptides during database searching which is computationally intensive and can result in increased false-positive identification of crosslinks. The second issue is that the masses of the two crosslinked peptides are not typically measured during MS. As a result one is confronted with the challenge of considering large numbers of candidate crosslinked peptide pairs during database searching. The quadratic increase in database search space that occurs to account for all combinations of candidate crosslinked peptide pairs presents a formidable challenge for both crosslinked peptide spectrum matching (PSM) and estimation of false-discovery rate (FDR) [3, 4]. For example, the 32 proteins that comprise the RNA polymerase II pre-initiation complex (PIC) generate half a billion candidate peptide pairs upon sequence recombination [5]. To address these issues, various heuristic strategies or the use of limited databases are commonly employed to identify the crosslinked peptides. The pLink program performs a quick open modification database search with the mass difference as the modification mass and then fine scores the pairs from the top 500 peptides whose masses are larger or smaller than half of the precursor mass [3]. The xQuest program uses a *de-novo* strategy to generate ion-tags to limit the selection of peptide pairs [6]. These heuristic strategies were used to perform crosslinked peptide database searching using *E. coli* or *C. elegans* proteome databases [3, 7]. Other search programs, such as Kojak [8], xiNET [9], Nexus [1], and Protein Prospector [10], use open modification database searching or exhaustive peptide recombination, to search limited databases (usually < 50 proteins). In general, confident identification of crosslinked peptides by database searching is still challenging and most CL-MS studies have been performed on relatively simple mixtures of purified proteins or protein complexes.

One approach to overcome the issue of confident crosslinked peptide identification involves the use of “MS labile” crosslinkers, which can be readily fragmented during MS2 while leaving the peptides intact so that the crosslinked peptide is transformed into two or more modified peptides which can in turn be selected for collision-induced dissociation (CID) during MS3 for identification. Various MS labile crosslinkers have been described with one or more of the following MS labile bonds: the Aspartyl-Prolyl-bond (D-P) [11, 12], various forms of a carbon-sulfur (C-S) bond [13, 14], a urea moiety [15], and a Rink moiety [16, 17]. The MS labile strategy allows for confident crosslinked peptide identification by whole proteome database searching, but suffers from long MS duty cycles due to multiple MS3 events, in-source decay, and poor MS3 signals. Recently, Liu *et al.* utilized the CID-labile but electron-transfer dissociation (ETD)-stable property of disuccinimidyl sulfoxide (DSSO) to deduce the masses of the two crosslinked peptides from the CID-MS2 spectrum and then performed PSM using the ETD-derived MS2 spectrum [18]. With the knowledge of the

individual masses of the two crosslinked peptides, they could efficiently limit the search space for the crosslinked peptides and applied the approach to identify crosslinked peptides derived from human cell lysates [18].

Recognizing the issues associated with MS labile crosslinkers, a number of groups have recently described “MS cleavable” crosslinkers that fragment in the same energy regime as the peptide backbone. During MS2, characteristic fragment ions pairs that are generated due to bond breakage in the crosslinker are used to determine the masses of the two crosslinked peptides, while peptide-specific fragment ions are used for PSM along with the deduced masses of the two peptides [18–20]. Confident and efficient identification of crosslinked peptides is achieved by using the MS2 fragment ion information to both distinguish spectra derived from crosslinked peptides from spectra derived from other types of peptides, and to search large sequence databases in a peptide mass-restricted manner.

Here we describe a new class of MS cleavable crosslinkers, called isotopomeric crosslinkers (ICLs), which allow confident and efficient identification of crosslinked peptides from their MS2 spectra by whole proteome database searching. Like other MS cleavable crosslinkers, ICLs require only a single MS2 event to obtain the fragment ion information needed to determine the masses of the two crosslinked peptides as well as the peptide-specific fragment ion information needed for PSM by whole proteome database searching. That is, the duty cycle of ICL-MS is the same as the duty cycle used for identification of linear peptides. ICL-MS is based on the novel design of our ICL crosslinker, a simple symmetrical molecule that asymmetrically incorporates stable isotopes into the two arms of the reagent. ICLs generate pairs of isotopomeric crosslinked peptides differing only by the positions of the heavy and light isotopes. During MS2 analysis, these ICL-crosslinked isotopomeric peptides generate unique isotopic doublet ions which are used to determine the masses of the two crosslinked peptides, as well as peptide-specific fragment ions which are used for PSM. We describe the synthesis of ICL-1 and present the fragmentation characteristics of ICL-modified peptides that permit confident and efficient identification of the peptides by whole proteome database searching. We also demonstrate the ability of ICL-MS to identify crosslinked peptides derived from increasingly complex mixtures using whole proteome database searches.

Methods

Materials

Methyliminodiacetic acid, N,N,N',N'-tetramethylchloroformamidinium hexafluorophosphate (TCFH), N,N-diisopropylethylamine (DIEA), N,N,N',N'-Tetramethyl-O-(N-succinimidyl)uronium tetrafluoroborate (TSTU), Trifluoroacetic acid (TFA), dichloromethane (DCM), Dimethylformamide (DMF), were purchased from Sigma-Aldrich. 2-chlorotriyl chloride resin, (benzotriazol-1-yloxy) tripyrrolidinophosphonium hexafluorophosphate (PyBOP), 1-hydroxybenzotriazole (HBot), were purchased from EMD Millipore. N-Fmoc-Gly-OH ($^{13}\text{C}_2$, ^{15}N) was purchased from Cambridge Isotope Laboratories, Tewksbury, MA.

Synthesis of the ICL-1 crosslinker

Asymmetric synthesis of ICL-1 using methyliminodiacetic acid follows the protocol for peptide synthesis using Fmoc-iminodiacetic acid (IDA) synthesis [21]. A mixture of methylimino diacetic acid (**1**) (1eq), N,N,N',N'-tetramethylchloroformamidinium hexafluorophosphate (TCFH) (1eq), and N,N-diisopropylethylamine (DIEA) (2eq) in dry DMF was stirred for one hour at room temperature. A solution of glycine tert-butyl ester (**2**) (1eq) and DIEA (1eq) in DMF was added to the mixture and stirred overnight to obtain product [**2**-[(2-tert-butoxy-2-oxoethyl)amino]-2-oxoethyl](methyl)amino]acetic acid (**3**). N-Fmoc-Gly-OH ($^{13}\text{C}_2$, ^{15}N) is loaded onto a 2-chlorotrityl chloride resin (2eq) in DMF with DIEA (4eq) and the Fmoc protection is removed. A solution of **3** (3eq) was added directly to the resin with (benzotriazol-1-yloxy) tripyrrolidinophosphonium hexafluorophosphate (PyBOP) (3eq) and 1-hydroxybenzotriazole (HBot) (3eq) under nitrogen bubbling for 3 hours. The resin was washed extensively with DMF/Methanol/DCM and 1% TFA/DCM was used to release the product from the resin. The cleavage solution was evaporated under vacuum to obtain a sticky and oily product, which was dissolved in water and precipitated by ether to obtain the white powder of product 2,2,9-trimethyl-4,7,11-trioxo(13,14- $^{13}\text{C}_2$,12- ^{15}N)-3-oxa-6,9,12-triazatetradecan-14-oic acid (**4**). Ninety-five percent (95%) TFA/water was added to the product **4** to remove the tert-butyl protection group. Ether was then added to precipitate the final product [**2**-[(carboxymethyl)amino]-2-oxoethyl](methyl)amino]acetyl(^{15}N)amino($^{13}\text{C}_2$)acetic acid (**5**) or the ICL-1 acid as a white powder. The yield was 68%, based on the percentage of isotopically heavy glycine that was incorporated into the final product, and the purity was 90%. The free acid of the final product **5** (100 mM) was activated using N,N,N',N'-Tetramethyl-O-(N-succinimidyl)uronium tetrafluoroborate (TSTU) (2eq) and DIEA (6eq) in DMF for one hour. The activated crosslinker was directly added to the protein/peptide solutions at a final concentration of 2–5mM for the crosslinking reactions.

Sample preparation, mass spectrometry analysis and data processing

An *RPB1* [(the largest subunit of RNA polymerase II (Pol II))-TAP tagged yeast strain was used to purify the Pol II complex. The Pol II sample was crosslinked by ICL-1, reduced and alkylated by iodoacetamide and trypsin digested. The peptides were desalted by C18 cartridge and fractionated into 5 fractions by microcapillary strong cation exchange (SCX) chromatography. The peptides were analyzed on a Thermo Scientific Orbitrap Elite with higher-energy collisional dissociation (HCD) fragmentation. Details of sample preparation and mass spectrometry setting are in the supplementary methods. The RAW files were converted to mzXML files using RawConverter [22] software. We used the Comet search engine [23] for the analysis of the unmodified, monolinked and looplinked peptides. Then we used our in-house developed Nexus2.0 algorithm for ICL doublet ion feature extraction and database searches (executables can be obtained upon request). The forward and reverse sequences from a yeast or bovine whole proteome database were used for ICL-1 analyses with the following parameter settings: (a) up to three miscleavages; (b) static modification on Cysteines (+57.0215 Da); (c) differential oxidation modification on Methionines (+15.9949 Da); (d) differential modification on the peptide N-terminal Glutamic acid residues (−18.0106 Da) or N-terminal Glutamine residues (−17.0265 Da); (e) differential

mono-ICL-1 modification on Lysines (+246.0892Da); (f) MS1 mass tolerance 20 ppm, MS2 fragment mass tolerance 15 ppm. We also used pLink, [3] with restricted databases, to identify the crosslinked peptides in the samples.

Results and Discussion

Rationale for MS cleavable ICL crosslinkers

A key issue that restricts the effectiveness of most CL-MS approaches is their limited ability to confidently and efficiently identify crosslinked peptides derived from complex mixtures using whole proteome database searches. This is primarily due to the lack of information about the masses of the two linked peptides and the resulting quadratic increase in database search space that occurs to account for all combinations of peptides in a database that sum to the measured precursor mass minus the crosslinker mass. This issue is compounded by the inability to distinguish MS2 spectra derived from crosslinked peptides from spectra derived from all other peptides. While data from MS labile crosslinking experiments can be used to efficiently search whole proteome databases, the reliance on multiple fragmentation events per peptide can limit the efficiency and sensitivity of the analysis. To overcome the limitations of CL-MS for identification of crosslinked peptides from complex mixtures, we designed ICLs, which allow MS2 spectra derived from crosslinked peptides to be distinguished from spectra derived from other types of peptides, and permit the masses of the individual peptides in the crosslink to be determined from the MS2 spectrum that is also used for PSM to identify the peptides. The common feature and the essence of ICLs is a chemically symmetrical molecule that asymmetrically incorporates heavy and light stable isotopes into the two arms of the reagent (Fig. 1A). ICLs are homo-bi-functional crosslinkers (such as amine- or sulfhydryl-reactive) containing two MS cleavable bonds. Our design also permits the incorporation of a chemical moiety in the middle of the molecule, which can be used to enrich crosslinker-modified peptides, such as biotin or an alkyne/azide group.

Because of the chemically symmetrical nature of ICLs, two peptides (peptide A and peptide B, Fig. 1B) crosslinked with an ICL crosslinker will have an equal chance to react with the light or heavy isotopic arm of the crosslinker. Thus, there will be two forms of a crosslinked peptide with exactly the same chemical composition but a different orientation of the crosslinker (Fig. 1B, designated A→B or A←B). These two forms of a crosslinked peptide are isotopomers, differing only by the positions of isotopes. ICLs contain MS-cleavable bonds, so that upon fragmentation by CID/HCD (also compatible with other fragmentation methods), the backbone peptide bonds of peptide A and peptide B generate b and y daughter ions composed of the peptide, or the peptide plus the reacted crosslinker and the other peptide. These ions appear as singlets regardless of the orientation of the crosslinker. However, peptide bond breakage within the crosslinker generates doublet ions for peptide A and peptide B due to the presence of the light or heavy crosslinker moiety (Fig. 1B). These doublets have ~1:1 ratio of intensities. Thus, the isotopic doublet ions are only generated from bond breakage within the crosslinker and breakage at all other peptide bonds will generate singlets. The isotopic doublet ions allow the masses of the two linked peptides to be determined, so that for a given crosslinked peptide, one will know the precursor mass from

MS1, and both the product ion m/z 's and the masses of the two peptides from MS2. Knowing the exact masses of peptide A and peptide B, a whole proteome database search can be performed to identify each peptide, without the need to consider all peptide combinations that sum to the precursor mass minus the crosslinker. Furthermore, because spectra from crosslinked peptides are distinguished from spectra derived from other peptides, only crosslinked spectra are used for the database search thus reducing computational burden and false positive identifications. These features can be easily incorporated into other database search algorithms designed to identify peptides crosslinked with MS cleavable/labile reagents.

Synthesis and features of ICL-1

To synthesize the ICL reagent, we utilized the symmetric iminodiacetic acid (IDA) as a building block (Fig. 2A). The isotopically heavy and light arms were produced by linking isotopically light or heavy glycine residues through peptide bonds to IDA. N-Fmoc-IDA was used for the asymmetric synthesis [21]. For proof-of-principle, we used methyliminodiacetic acid to synthesize the acid form of ICL-1 (Fig. 2B, detailed description of synthesis in Methods). ICL-1 was then activated to an NHS ester by reaction with TSTU and DIEA. ICL-1 has a theoretical maximum distance of 15 Å between the reactive groups and yields a molecular weight modification of 228.0787 Da for ICL-1 crosslinked peptides.

We tested ICL-1 using a synthetic peptide:

acSDERAKIGGLNDPRTVKEVQGLFSPEEVR, which has two Lys residues and an acetylated N-terminus. Upon reaction with ICL-1 and trypsin digestion, the peptides were analyzed using LC-ESI-MS on an Orbitrap-Elite Fourier Transform Mass Spectrometer (FTMS) with different HCD energy settings without dynamic exclusion. The ICL-1 modified peptides can be readily identified from their masses and MS2 spectra. Fig. 3 shows an MS2 spectrum derived from the monolinked peptide AK*IGGLNDPR, where K* is modified by ICL-1. The corresponding b and y ions resulting from fragmentation along the peptide backbone are labeled. Two isotopic ion pairs corresponding to breakage at the two peptide bonds in the crosslinker are shown in the enlarged inserts (1097.5911/1100.6045, 1168.6451/1171.6501). The two ions that comprise each doublet have similar intensities, as expected. The ion pair 1097.5911/1100.6045 is a y ion pair corresponding to the peptide modified with a glycine residue (modification masses of 57.0214 and 60.0244, respectively). The less intense ion pair 1168.6451/1171.6501 is a c ion pair due to breakage at the C-N bond which gives a modification mass of 128.0586 and 131.0616 to the peptide ion. The b ion pair 1196.94/1199.65 was not observed. A unique ion pair 159.0766/ 162.0894 corresponds to a or b ions (187.07/190.07) after neutral loss of CO from the free end of the crosslinker. This unique ion pair is a characteristic feature observed in all MS2 spectra derived from ICL-1 monolinked peptides. The y ion pair of 1097.5911/1100.6045 was observed in 148 out of 163 (91%) MS2 spectra derived from the monolinked peptide during its elution peak. In nearly all of the cases where this y ion pair was observed, a singlet ion corresponding to the mass of the naked peptide MH⁺ (1040.5752) was also observed due to the additional loss of the Glycine (57.0214/60.0244 Da) moiety.

The MS2 spectrum of the crosslinked peptide composed of AKIGGLNDPR and TVKEVQFGLFSPEEVR is shown in Figure 4. The y ions corresponding to fragmentation of peptide AKIGGLNDPR are labeled as lower case 'y's. The y ions corresponding to fragmentation of peptide TVKEVQFGLFSPEEVR are labeled as capital 'Y's. Two ion pairs, 1097.6031/1100.6077 and 1922.0267/1925.0056, corresponding to y ions of the intact peptides AKIGGLNDPR and TVKEVQFGLFSPEEVR modified with a glycine residue respectively, are shown in the enlarged inserts. The y ion pair corresponding to the peptide AKIGGLNDPR (ion pair 1097.6031/1100.6077), was observed in 87 out of 93 (94%) MS2 spectra derived from the crosslinked peptide during its elution peak and the singlet ion corresponding to the loss of Glycine (1040.5752) was observed in all but two of these spectra. The y ion pair for the corresponding peptide TVKEVQFGLFSPEEVR (1922.0267/1925.0056) was only observed in 14 out of 93 (15%) MS2 spectra derived from the crosslinked peptide during its elution peak. This may be attributable to the large m/z of these ions. Because HCD fragmentation primarily produces 1+ ions, we seldom observe doublet pairs with charge states of 2+ or higher. Higher HCD energy produces more low m/z fragment ions and we found that doublet ion pair intensity was lower when 40% normalized collision energy (NCE) was used compared to 30% or 35% NCE. Since some intact precursor ions were observed at 30% NCE, 35% NCE was used for the rest of the experiments.

This example shows that two peptides crosslinked with ICL-1 can 1) fragment efficiently during HCD-MS2, as expected, and 2) can produce doublet ions that correspond to y ions of the two peptides modified with either light or heavy glycine. Due to the orientation of the peptide bonds in IDA, the reporter doublet ions are all y ions, which have a better chance of being observed in MS2 spectra. We rarely observe the corresponding b ion pairs. However, the presence of doublet ions in the MS2 spectra will depend on the precursor ion intensity and the physico-chemical properties of the individual peptides. Prior to using the MS2 spectra for database searching, we applied two filters to determine whether an MS2 spectrum corresponds to a potential crosslinked peptide: 1. If both doublet ion pairs are found and their summed mass is equal to the precursor mass minus the linker mass (113.0477 Da in case of ICL-1), or 2. If only one doublet is found, we search for the presence of a singlet y ion, corresponding to the loss of glycine from the doublet ion pair. If found, the doublet is identified as peptide A, and the other peptide's mass is deduced from the precursor mass and the mass of peptide A. Among all the spectra derived from the ICL-1 crosslinked peptide in the previous example, 85 out of 93 (91%) could be extracted out as potential crosslinked spectra when these two filters were applied. We also applied these criteria to a previously published dataset in which the BS3 crosslinker was used to crosslink the yeast TFIIF complex [1]. Out of 48499 MS2 spectra, only 417 (~0.86%) spectra passed the filters and searching these spectra against a yeast proteome database resulted in zero identified crosslinks. This test shows that the filters provide a good way to reduce false positive identification of ICL-1 crosslinked peptides by separating spectra derived from ICL-1 crosslinked peptides from other spectra, which is important for efficient and confident crosslinked peptide identification using whole proteome database searches.

Application of ICL-MS to bovine serum albumin (BSA)

To test whether ICL-MS can be used to confidently identify crosslinked peptides from whole proteome database searches, we used ICL-1 to crosslink BSA and analyzed the crosslinked sample directly by LC-ESI-MS using the Orbitrap-Elite after trypsin digestion and C18 purification. We first used the pLink database search algorithm[3] to identify ICL-1 crosslinked peptides using a database composed of the two BSA (P02769 and Q3SZR2) sequences. We reasoned that this search would identify most of the spectra derived from ICL-1 crosslinked peptides and would thus serve as a reference to evaluate the performance of our Nexus2.0 algorithm. Nexus2.0 applies the two filters described above to identify potential spectra derived from ICL-crosslinked peptides. Once MS2 spectra corresponding to ICL-crosslinked peptides are identified, Nexus2.0 deduces the masses of the two peptides in the crosslink and then uses the mass information and fragmentation spectrum to identify each peptide by whole proteome database searching.

pLink identified 6 unique BSA crosslinks from 16 spectra. We then used Nexus2.0 to identify potential spectra derived from ICL-1 crosslinked peptides. Out of 6290 MS2 spectra, 471 spectra (7.5%) passed the filters for ICL-1 crosslinked spectra, and individual peptide masses were deduced from these MS2 spectra. These spectra were used to search a bovine proteome database (9396 entries) and its reverse decoy database, and the search was completed within one hour (Table S-1). Importantly, and as expected for a crosslinking reaction consisting of BSA alone, no interlinked peptides were identified, even though the bovine proteome was searched. 6 intralinked peptides were identified from 13 spectra by Nexus2.0. All of these crosslinks were also identified by pLink. Extraction of doublet ion information failed for 4 spectra identified by pLink because either the isotopic envelope was not observed, thus preventing charge state determination, or contaminating ions interfered with doublet ion identification. One spectrum that lacked doublet ion features could be a false positive identification by pLink; it has a large E-value and the C α -C α distance between crosslinked lysines is larger than expected for ICL-1. In addition, Nexus2.0 identified a crosslink derived from a collagen-like protein (Q2KIT0) from two spectra. Q2KIT0 was also identified in the BSA sample using the Comet search engine to search for unmodified peptides. Upon performing a pLink search against a database containing the two BSA proteins and Q2KIT0, pLink also identified this intralink from Q2KIT0 (Table S-1). The results show that ICL-MS can be used for confident identification of crosslinked peptides from whole proteome database searches at a low false discovery rate.

Application of ICL-MS to the RNA polymerase II complex

To evaluate the effectiveness of ICL-MS for confidently identifying crosslinked peptides derived from complex samples and for mapping PPIs in protein complexes, we applied ICL-MS to a sample containing a highly purified yeast RNA polymerase II complex (Pol II, Fig. 5A). One difficulty for CL-MS experiments, especially for those using whole proteome database searches, involves distinguishing true-positive identifications from false positive identifications. This is particularly challenging for crosslinks between two different proteins compared to crosslinks within the same protein because the likelihood of identifying two peptides from the same protein from a whole proteome database search by random chance is much lower than the likelihood of identifying two peptides from two different proteins.

Previous studies that used whole proteome database searches to identify crosslinked peptides/proteins in very complex mixtures such as whole cells or cell extracts[7, 18, 24], identified hundreds of inter-protein crosslinks, but in many cases it is not possible to know whether the identified crosslinks represent true positives, *i.e.*, *bona fide* PPIs. By using a highly purified sample of Pol II it is reasonable to assume that we should only identify crosslinked peptides between and within Pol II subunits (interlinks and intralinks, respectively). Any identified crosslink that contains a peptide from a protein that is not detectable in the sample, or a peptide from the decoy database, can be considered a false positive.

After crosslinking the Pol II sample with ICL-1, and preparing the sample for MS analysis, the peptides were analyzed by LC-ESI-MS using an Orbitrap-Elite mass spectrometer with HCD fragmentation. 101 proteins were identified by a Comet search for unmodified and mono-linked peptides at 0.5% FDR, with the majority of identified spectra (89%) corresponding to peptides from the 12 subunits of Pol II (Table S-2). 5965 spectra (~18.2%) out of 32,688 MS2 spectra contained features consistent with an ICL-crosslinked peptide. These spectra were searched against the *S. cerevisiae* proteome database and its reverse decoy database (total of 13,600 entries) using Nexus2.0. The search was completed in 12 hours. Nexus2.0 identified 105 unique intralinks from 333 spectra, and 42 unique interlinks from 119 spectra (Fig. 5B, Fig. 5C, Fig. 5D and Table S-2). Two false positive interlinks were identified by Nexus2.0 involving a Pol II subunit and a non-Pol II protein (Rpb2 to YEL019C and Rpb11 to YNR012W), that were not identified in the sample by the Comet search; all of the other identified interlinks are between Pol II subunits (Fig. 5B). One intralink from TDH3 was identified. The TDH3 protein was identified when Comet was used to identify proteins in the sample and the Ca-Ca distance of the crosslinked Lysine residues is 8.9 Å when mapped onto the modelled structure; therefore this intralink was not considered a false positive. All of the other identified intralinks involved Pol II subunits. Among them, 4 spectra corresponding to 4 intralinks involving peptides with the same sequences were considered false positives. Crosslinks involving peptides with the same or overlapping peptide sequences could indicate the presence of a homodimer, but this is not expected for Pol II. We estimate that the false positive rate is about 2.5% for interlinks and 1.2% for intralinks. We also used pLink to search for ICL-1 crosslinked peptides using a database containing only sequences from the 12 Pol II subunits. pLink identified 100 unique intralinks within Pol II subunits from 376 spectra and 66 unique interlinks between Pol II subunits from 174 spectra. 62 intralinks and 30 interlinks were identified by both pLink and Nexus2.0 (Fig. 5C). 38 intralinks and 36 interlinks were only identified by pLink by searching against a database composed of the 12 Pol II subunit sequences. Nexus2.0 did not identify these crosslinks because no doublet ion features could be identified in the MS2 spectra. Strikingly, 43 intralinks and 10 interlinks were only identified by Nexus2.0 by searching against the whole yeast proteome (Fig. 5C). The reason that pLink failed to identify these crosslinks may be due to the failure of these spectra to pass the pLink FDR cutoff when a limited database search was used. We mapped all identified crosslinks onto the Pol II crystal structure (PDB 1WCM). Most Ca-Ca distances between crosslinked residues were within or close to the theoretical crosslinking distance (< 35Å) of ICL-1 (Fig. 5D and Table S-2). These results demonstrate the ability of ICL-MS to permit confident and

efficient identification of crosslinked peptides derived from complex mixtures by searching their MS2 spectra against whole proteome databases, and its utility for mapping PPIs and the sites of the PPIs in protein complexes.

In this paper, we report the novel design of ICL crosslinkers which allow efficient and confident identification of crosslinked peptides using mass spectrometry and whole proteome database searching of MS2 spectra. The unique features of our ICL crosslinkers that enable confident crosslink identification are the following: First, ICLs are based on the novel design concept of a symmetrical bi-functional crosslinker that incorporates heavy and light isotopes into the two arms of the molecule (Figs. 1A and 2A). This feature ultimately enables the masses of the individual peptides to be determined from the MS2 data. This information not only enables whole proteome database searches but also increases the confidence and reduces the false discovery rate (FDR) of crosslinked peptide identification. Because the two peptide masses are determined from the MS2 spectra and are not deduced from the mathematical combination of the masses of the best matching peptide pairs, ICL-MS improves the confidence of crosslinked peptide identification. Whole proteome database searches also allow FDR estimation using whole proteome reverse sequences (as is done for FDR estimation during standard peptide identification). Most CL-MS methods only use the reverse sequences from a limited database (< 100 proteins). Second, the doublet features in the crosslinked MS2 spectra easily distinguish spectra derived from crosslinked peptides from spectra derived from linear, mono-linked, or loop-linked peptides, which increases search speed by limiting the analysis to spectra derived from crosslinked peptides and further increases confidence by excluding the possibility of falsely identifying a crosslinked peptide from a spectrum derived from a non-crosslinked peptide. Third, unlike other CL-MS approaches that employ isotopically labeled crosslinkers, ICL-MS involves only one crosslinker, so that there is only one peak in the MS1 scan per crosslinked peptide. As a result, the MS1 spectral space is simplified, which may further improve the ability to detect and identify crosslinked peptides in complex mixtures. In addition, data acquisition is more efficient because there is no need to acquire MS2 spectra for both the heavy and light modified peptides. Fourth, ICLs are small, inexpensive molecules that can be analyzed by most high accuracy mass spectrometers; features that will facilitate the widespread use of the technology. Finally the design of ICLs permits easy incorporation of chemical moieties that can be used to enrich ICL-modified peptides from complex mixtures, which should improve detection and identification of crosslinked peptides. This shall be easily achieved using Fmoc-IDA as the starting material.

While the doublet ion information provided by ICLs is attractive because it allows confident and efficient identification of ICL-crosslinked peptides using whole proteome database searching, a loss in sensitivity may occur if the crosslinked spectra fail to generate useful doublet ions. We observed several scenarios that account for the failure to generate useful doublet ions. The most common scenario involves situations in which the charge state of the doublet ion cannot be determined due to the absence of its naturally occurring isotopic ions. The overall fragment ion intensity is generally lower for many of these spectra. The second scenario involves situations in which the m/z of the doublet ion(s) is large, and, as a result, is either outside the scan range or the intensity of the doublet ions is low. HCD generates mostly 1+ ions. Alternative fragmentation methods such as CID that are more likely to

produce fragment ions with higher charge states may alleviate this issue. A third scenario involves situations where the isotopic distributions of the isotopically heavy and light doublet ions appear to overlap due to the presence of unrelated ions. This can make it difficult to distinguish the doublet ions from other ions. In the future, the use of heavy and light glycine residues that have a 5 Da mass difference, as opposed to 3 Da for ICL-1, will generate doublet ions that are better separated on the m/z scale and may improve identification of the doublet ions. While Nexus2.0 failed to identify some crosslinks in the Pol II sample that were identified by pLink because of the issues described above, Nexus2.0 identified many crosslinks that pLink failed to identify even though pLink used a 12 protein database compared to the yeast proteome database used by Nexus2.0. It is likely that pLink failed to identify these crosslinks because the corresponding spectral matches didn't pass pLink's FDR cut-off. Nonetheless, ICL-MS was able to confidently identify these crosslinks from whole proteome searches because the two peptide masses were specified. The ability of ICL-MS to confidently identify crosslinked peptides from their MS2 spectra using whole proteome database searches is expected to complement existing CL-MS approaches for large scale PPI studies.

ICL crosslinkers are chemically symmetrical but asymmetrically incorporate heavy and light isotopes into the two arms of the crosslinker. As a result, only breakages within the ICL crosslinker generate isotopic doublets ions. In most cases, at the energy used for HCD, only one breakage within the crosslinked peptide occurs generating the corresponding b and y ions. As a result, peptide bond breakage within the ICL crosslinker produces a doublet ion that is composed of an intact peptide with a crosslinker-derived modification. Occasionally we observe more than two doublet ions in one MS2 spectrum, some of which are due to two or more breakages within the crosslinked peptide, with one breakage within the crosslinker and another breakage elsewhere along the peptide backbone. This is especially true for ICL-1 monolinked peptides. Multiple doublet ions can also occur due to co-isolation of an unrelated crosslinker-modified peptide. However, most of these doublet ions will be filtered out by the two criteria used for ICL-crosslinked peptide spectrum identification: 1) The sum of the masses of the two isotopic doublet ions plus the linker mass equals the precursor mass; or 2) A singlet ion which corresponds to the loss of a glycine moiety from a doublet ion is observed. If multiple potential isotopic doublets are identified in an MS2 spectrum, each pair is treated equally during database searching, and the best scoring PSM pairs are reported as the crosslinked peptide pair. For the Pol II ICL-1 data, the average number of isotopic doublet ion pairs per spectrum was 1.73. Even with multiple doublet pairs found in one spectrum, the potential search space is still negligible compared to all potential candidate peptide combinations in the database. In the Pol II crosslinking experiment, out of 5965 spectra containing doublets features, 2667 (~45%) spectra had a minimal PSM Xcorr score of 3. Out of these 2667 spectra, at least one peptide from a Pol II subunit was identified from 1276 (48%) spectra, suggesting that the doublet information correctly predicted the individual peptide masses from these spectra. There are many spectra for which only one Pol II peptide is identified that fail to pass the FDR filter. Future development of ICL-MS will seek to increase the identification rate of spectra derived from crosslinked peptides.

Conclusions

ICL-MS is a new way to perform MS-based crosslinked peptide identification. The unique design features of the technology allow confident identification of crosslinked peptides and permit efficient searches of whole proteome databases. These features can be easily incorporated into other database search algorithms designed to identify peptides crosslinked with MS cleavable/labile crosslinking reagents, which will facilitate the utility of ICL-MS for CL-MS studies. Further optimization of the ICL design and MS settings are expected to improve doublet ion identification which will improve crosslink identification rates. Future ICL crosslinkers will also contain chemical moieties that can be used to enrich crosslinker-modified peptides from complex mixtures. The combination of an enrichment handle with the features described here within a single crosslinker is expected to produce a CL-MS technology that will complement existing CL-MS approaches for large scale PPI studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Mark Gillespie and Isil Hamdemir for comments on the manuscript and the Fred Hutchinson Cancer Research Center proteomics facility (L.A. Jones and P.R. Gafken) for MS analyses. This work was supported by NIH grant R01GM110064 to JR.

References

1. Luo J, Cimermancic P, Viswanath S, Ebmeier CC, Kim B, Dehecq M, Raman V, Greenberg CH, Pellarin R, Sali A, Taatjes DJ, Hahn S, Ranish J: Architecture of the Human and Yeast General Transcription and DNA Repair Factor TFIIH. *Mol Cell.* 59, 794–806 (2015) [PubMed: 26340423]
2. Schweppe DK, Chavez JD, Bruce JE: XLmap: an R package to visualize and score protein structure models based on sites of protein cross-linking. *Bioinformatics.* 32, 306–308 (2016) [PubMed: 26411867]
3. 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.* 9, 904–906 (2012) [PubMed: 22772728]
4. Walzthoeni T, Claassen M, Leitner A, Herzog F, Bohn S, Forster F, Beck M, Aebersold R: False discovery rate estimation for cross-linked peptides identified by mass spectrometry. *Nat Methods.* 9, 901–903 (2012) [PubMed: 22772729]
5. Murakami K, Elmlund H, Kalisman N, Bushnell DA, Adams CM, Azubel M, Elmlund D, Levi-Kalishman Y, Liu X, Gibbons BJ, Levitt M, Kornberg RD: Architecture of an RNA polymerase II transcription pre-initiation complex. *Science.* 342, 1238724(2013) [PubMed: 24072820]
6. 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.* 5, 315–318 (2008) [PubMed: 18327264]
7. Tan D, Li Q, Zhang MJ, Liu C, Ma C, Zhang P, Ding YH, Fan SB, Tao L, Yang B, Li X, Ma S, Liu J, Feng B, Liu X, Wang HW, He SM, Gao N, Ye K, Dong MQ, Lei X: Trifunctional cross-linker for mapping protein-protein interaction networks and comparing protein conformational states. *Elife.* 5, (2016)
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.* 14, 2190–2198 (2015) [PubMed: 25812159]

9. Combe CW, Fischer L, Rappsilber J: xiNET: cross-link network maps with residue resolution. *Mol Cell Proteomics*. 14, 1137–1147 (2015) [PubMed: 25648531]
10. Chu F, Baker PR, Burlingame AL, Chalkley RJ: Finding chimeras: a bioinformatics strategy for identification of cross-linked peptides. *Mol Cell Proteomics*. 9, 25–31 (2010) [PubMed: 19809093]
11. 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*. 78, 8059–8068 (2006) [PubMed: 17134140]
12. 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*. 8, 409–420 (2009) [PubMed: 18936057]
13. Lu Y, Tanasova M, Borhan B, Reid GE: Ionic reagent for controlling the gas-phase fragmentation reactions of cross-linked peptides. *Anal Chem*. 80, 9279–9287 (2008) [PubMed: 19551991]
14. Kao A, Chiu CL, Vellucci D, Yang Y, Patel VR, Guan S, Randall A, Baldi P, Rychnovsky SD, Huang L: Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes. *Mol Cell Proteomics*. 10, M110 002212 (2011)
15. Muller MQ, Dreier F, Ihling CH, Schafer M, Sinz A: Cleavable cross-linker for protein structure analysis: reliable identification of cross-linking products by tandem MS. *Anal Chem*. 82, 6958–6968 (2010) [PubMed: 20704385]
16. Tang X, Munske GR, Siems WF, Bruce JE: Mass spectrometry identifiable cross-linking strategy for studying protein-protein interactions. *Anal Chem*. 77, 311–318 (2005) [PubMed: 15623310]
17. Luo J, Fishburn J, Hahn S, Ranish J: An integrated chemical cross-linking and mass spectrometry approach to study protein complex architecture and function. *Mol Cell Proteomics*. 11, M111 008318 (2012)
18. Liu F, Rijkers DT, Post H, Heck AJ: Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nat Methods*. 12, 1179–1184 (2015) [PubMed: 26414014]
19. Arlt C, Gotze M, Ihling CH, Hage C, Schafer M, Sinz A: Integrated Workflow for Structural Proteomics Studies Based on Cross-Linking/Mass Spectrometry with an MS/MS Cleavable Cross-Linker. *Analytical chemistry*. 88, 7930–7937 (2016) [PubMed: 27428000]
20. Iacobucci C, Gotze M, Ihling CH, Piotrowski C, Arlt C, Schafer M, Hage C, Schmidt R, Sinz A: A cross-linking/mass spectrometry workflow based on MS-cleavable cross-linkers and the MeroX software for studying protein structures and protein-protein interactions. *Nat Protoc*. 13, 2864–2889 (2018) [PubMed: 30382245]
21. Khattab SN, El-Faham A, El-Massry AM, Mansour EME, Abd El-Rahman MM: Coupling of iminodiacetic acid with amine acid derivatives in solution and solid phase. *Letters in Peptide Science*. 7, 331–345 (2001)
22. He L, Diedrich J, Chu YY, Yates JR 3rd: Extracting Accurate Precursor Information for Tandem Mass Spectra by RawConverter. *Analytical chemistry*. 87, 11361–11367 (2015) [PubMed: 26499134]
23. Eng JK, Jahan TA, Hoopmann MR: Comet: an open-source MS/MS sequence database search tool. *Proteomics*. 13, 22–24 (2013) [PubMed: 23148064]
24. Schweppe DK, Harding C, Chavez JD, Wu X, Ramage E, Singh PK, Manoil C, Bruce JE: Host-Microbe Protein Interactions during Bacterial Infection. *Chem Biol*. 22, 1521–1530 (2015) [PubMed: 26548613]

Figure 1.

Schematic of the ICL cross-linker and its MS2 breakage pattern. (a) Key features of a ICL cross-linker are shown. (b) Schematic illustrating expected products from ICL fragmentation during MS2 analysis. mP = the mass of the crosslinked peptide.

Figure 2.
(a) General structure of ICL reagents. Heavy isotopes are in red. (b) Synthesis scheme for ICL-1.

Figure 3.

MS2 spectrum of peptide AKIGGLNDPR modified with ICL-1 using HCD 35% NCE. The precursor m/z is 643.8401, charge state 2+. K[374.18] indicates the mass of lysine modified with ICL-1. The corresponding b and y ions resulting from fragmentation along the peptide backbone are labeled. Two isotopic ion pairs, 1097.5911/1100.6045 and 1168.6451/1171.6501, corresponding to y+ and c+ ions due to bond breakage within the crosslinker, as well as the isotopic pair, 159.0766/162.0804, derived from loss of CO from the free end of the crosslinker are shown in the enlarged inserts. The masses of fragment ions derived from bond breakage within the crosslinker, as well as the structures of the unique ion pair (159.0766/162.0804) derived from the neutral loss of CO from the b ion

(187.07/190.07) at the free end of the crosslinker, are depicted at the bottom of the figure. The MH^+ of the naked peptide (np) is 1040.5752.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 4.

MS2 spectrum of an ICL-1 crosslinked peptide using HCD 35% NCE. The precursor m/z is 1044.8838, charge state 3+. The y ions corresponding to fragmentation of peptide AKIGGLNDPR are labeled as lower case 'y's, and the ion corresponding to the naked peptide (MH⁺ of 1040.5846) is labeled "np". The y ions corresponding to fragmentation of peptide TVKEVQGLFSPEEVR are labeled as capital 'Y's. Two ion pairs, 1097.6031/1100.6077 and 1922.0267/1925.0056, corresponding to y ions of the intact peptides AKIGGLNDPR and TVKEVQGLFSPEEVR modified with a glycine residue respectively, are shown in the enlarged inserts. The masses of fragment ions derived from peptide bond breakage within the crosslinker are depicted at the bottom of the figure.

Figure 5.

The ICL-1 crosslinking results for the yeast Pol II complex. (a) Silver stained gel of 1ug purified yeast Pol II. Yeast Pol II subunits are indicated. (b) Mapping of the ICL-1 interlinks (41) and intralinks (104) identified by the Nexus2.0 search against the *S. cerevisiae* proteome database onto the Pol II subunits sequences. Inter- and intra-links are shown as red and blue lines respectively. Lysine residues that were identified in a crosslink are indicated by red dots. (c) Comparison of pLink search results, using a database consisting of the 12 Pol II subunits, and Nexus2.0 search results, using the *S. cerevisiae* proteome. (d) Distribution of $C\alpha-C\alpha$ distances between crosslinked residues identified by pLink (P) and Nexus2.0 (N) based on the structure from 1wcm.pdb. The unknown group contains crosslinks whose distance cannot be measured because the corresponding residues were not resolved in the structure. FP indicates the obvious false positives: for example, intralinks involving peptides with the same sequences, which could indicate the presence of a homodimer, but is not expected for Pol II; or interlinks between a Pol II subunit and a non-Pol II protein.