Published in final edited form as:

J Proteome Res. 2014 February 7; 13(2): 898–914. doi:10.1021/pr400953w.

Enhanced identification of zero-length chemical crosslinks using label-free quantitation and high-resolution fragment ion spectra

Sira Sriswasdi^{1,2}, Sandra L. Harper¹, Hsin-Yao Tang¹, and David W. Speicher^{1,2,*}

¹Center for Systems and Computational Biology and Molecular and Cellular Oncogenesis Program, The Wistar Institute, Philadelphia, Pennsylvania, USA

²Genomics and Computational Biology Graduate Group, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Abstract

Chemical crosslinking coupled with mass spectrometry provides structural information that is useful for probing protein conformations and providing experimental support for molecular models. “Zero-length” crosslinks have greater value for these applications than longer crosslinks because they provide more stringent distance constraints. However, this method is less commonly utilized because it cannot take advantage of isotopic labels, MS-labile bonds, or enrichment tags to facilitate identification. In this study, we combined label-free precursor ion quantitation and targeted tandem mass spectrometry with a new software tool, Zero-length Crosslink Miner (ZXMiner), to form a multi-tiered analysis strategy. A major, critical objective was to simultaneously achieve very high accuracy with essentially no false positive crosslink identifications, while maintaining a good depth of analysis. Our strategy was optimized on several proteins with known crystal structures. Comparison of ZXMiner to several existing crosslink analysis software showed that other algorithms detected less true positive crosslinks and were far less accurate. Although prior use of zero-length crosslinking was typically restricted to small proteins, ZXMiner and the associated strategy enables facile analysis of very large protein complexes. This was demonstrated by identification of zero-length crosslinks using purified 526 kDa spectrin heterodimers and intact red cell membranes and membrane skeletons.

Keywords

chemical crosslinking; mass spectrometry; software

INTRODUCTION

Chemical crosslinking coupled with mass spectrometry (CX-MS) is a valuable tool for probing protein complexes that complements protein biophysical methods and high-resolution structural determinations, as it provides distance constraints between specific protein residues. Particular advantages are that CX-MS can be applied to heterogeneous protein mixtures and very large protein complexes where high-resolution techniques such as X-ray crystallography and NMR cannot be directly exploited. Similarly, structural

*Corresponding author: David W. Speicher, tel. (215) 898-3972, fax (215) 898-0664, speicher@wistar.org.

SUPPORTING INFORMATION

Supporting information and representative annotated MS/MS spectra of crosslinked peptides are available. These materials are available free of charge via the Internet at <http://pubs.acs.org>.

information can be obtained from flexible or disordered proteins or physiologically-important conformational changes on proteins that are too large for NMR. Determining proximity for specific amino acid side chains can identify or confirm protein-protein interactions or support structural model predictions by both distinguishing between alternative predicted models and providing distance constraints for further model refinement.

CX-MS emerged as an important experimental tool in the late 1990s^{1, 2} and has been applied to diverse biological problems^{3–7}. The extensive progress in this field over the past decade has been recently reviewed^{8–14}. Historically, the biggest challenge in crosslinking experiments has been identification of crosslinked products, which are usually present at very low stoichiometry, often produce weak MS signals, and yield complex MS/MS spectra. The availability of high-resolution MS for analysis of proteolyzed crosslinked proteins or complexes has been a key advance in the field¹⁵. Progress has been further facilitated by development of a substantial number of crosslinker reagents that facilitate MS identification of crosslinked products¹¹, including incorporation of isotopic labels¹⁶, MS-labile bonds^{17, 18}, and enrichment tags^{19, 20}. In addition, a number of software tools have been developed to take advantage of the unique MS and MS/MS signatures produced by these specialized crosslinkers. However, the main drawback of crosslinkers that utilize these MS-friendly properties is that the spacer arm length is usually substantial. This greatly reduces the stringency, and therefore the value, of the resulting distance constraints⁹. Bifunctional crosslinkers also give rise to dead-end and self-loop products²¹ – linear peptides connected to only one end or both ends of the crosslinker reagent. The presence of products from these side reactions further increases the complexity of peptide mixtures and data analysis.

In contrast, “zero-length” crosslinkers such as EDC, eliminate a water molecule when a bond is formed between an amine and a carboxyl group, rather than adding any extra atoms to the crosslinked products. Hence, they yield the most stringent distance constraints, which are optimal for identifying contact sites in multi-protein complexes and for aiding computational molecular modeling⁹. Another advantage of zero-length crosslinkers is that they generally do not form stable dead-end and self-loop crosslinked products. But, the lack of a spacer arm does not allow incorporation of isotope labels or affinity tags, and relatively few studies have utilized zero-length crosslinkers with the majority of systems studied having less than 100 kDa of unique protein sequence^{22–27}. The unique protein sequence size is the most appropriate measure of data analysis complexity in a crosslinking experiment as complexity scales with the number of unique peptides and not the actual complex size. That is, a 400 kDa heterodimer is a far more challenging problem than a 400 kDa complex comprised of 10 copies of a 40 kDa polypeptide chain.

The most important factor that has restricted the use of zero-length crosslinkers to relatively small protein sequence-sized problems is the paucity of robust software tools that have been optimized for zero-length crosslink data. As a result, many prior studies involving zero-length crosslinks required extensive manual data analysis, which is undesirable because it is subjective and tedious. Most existing software tools were either developed specifically for non-zero-length crosslinkers or optimized based on non-zero-length crosslink datasets, and the few that are capable of processing zero-length crosslink datasets do so with critical limitations²⁸. Also, most existing software tools do not consider the important fact that a zero-length crosslink between adjacent peptides will have an identical precursor mass to an incomplete proteolysis of these adjacent peptides, and the MS/MS spectra are often quite similar. Another issue that is often not fully considered when interpreting datasets from crosslinking experiments is that very low false discovery rates are essential and maximum depth of analysis is highly desirable. Accurate crosslink assignment is critical because even a single false positive identification can be quite detrimental, leading to incorrect

assignments of protein-protein interactions or incorrect distance constraints for molecular modeling that will result in inaccurate structures. At the same time, failure to assign a significant portion of the actual crosslinked peptides present in a sample will usually diminish the value of the dataset due to missed distance constraints. Hence, a dedicated analytical strategy and software tools that are optimized for zero-length crosslinking is needed to fully realize the potential offered by these more precise distance constraints.

In this study, we developed software and an analytical strategy that was optimized for highly accurate identification of zero-length crosslinked peptides. The analytical method integrated two common proteomics techniques – namely label-free quantitation and targeted mass spectrometry – into a multi-tiered strategy yielding high-resolution MS/MS data with high depth-of-analysis for crosslinked peptides. An accompanying software tool, Zero length Crosslink Miner (ZXMiner), facilitates data acquisition processes and provides an optimized computational analysis of high-resolution MS/MS data in zero-length crosslink datasets. The data acquisition and analysis strategy was validated on proteins with known crystal structures. The performance of ZXMiner compared favorably with alternative software tools currently used for crosslinked peptide analysis. We demonstrate that good depth of analysis can be achieved with false discovery rates (FDR) of less than 1%. Importantly, this new method is not significantly limited by size of protein complex as even intact cell membranes can now be probed by zero-length crosslinking.

EXPERIMENTAL PROCEDURES

GST and myoglobin crosslinking

GST was purified after cleavage from a fusion protein as previously described²⁹ and myoglobin was purchased from Sigma-Aldrich. Crosslink reactions of GST (0.3 mg/ml) and myoglobin (0.4 mg/ml) were performed at 0°C using 25 mM EDC /12.5 mM sulfo-NHS and 10 mM EDC / 5 mM sulfo-NHS, respectively. Aliquots of the crosslinked products were removed after 60 min of reaction time, quenched by the addition of dithiothreitol, and subsequently buffer exchanged and concentrated on 10K MWCO filters. Crosslinked and untreated control samples were separated on SDS gels stained with colloidal Coomassie Blue. Bands of interest were excised, alkylated, and digested with trypsin as previously described³⁰. Tryptic digests were lyophilized and resuspended in 0.1% formic acid in MilliQ water at an estimated concentration of 0.1 µg/µL based on gel stain intensity.

Spectrin heterodimer crosslinking

Spectrin heterodimers were isolated as previously described³¹ from fresh blood obtained from healthy volunteer human donors using informed consent and under a protocol approved by an institutional review board. Prior to crosslinking, spectrin dimers were isolated in 10mM sodium phosphate, 130 mM sodium chloride at pH 7.4 by gel filtration using three G5000PWXL columns connected in series and maintained at 4°C. Sample concentration was then adjusted to 0.2 mg/ml. Crosslinking reactions were carried out at 0°C with 5 mM EDC / 2.5 mM sulfo-NHS, respectively. Aliquots were removed at 15, 30, 60, and 120 min and quenched by addition of 20 mM dithiothreitol (final concentration) with incubation for 15 min at reaction temperature. Samples were buffer exchanged, concentrated using an Amicon Ultra 50K filtration unit, and separated on 3–8% Tris-Acetate mini-gels in Tris acetate running buffer. Bands corresponding to the molecular weight of dimer were excised, alkylated, and digested with trypsin. Tryptic digests were lyophilized and resuspended in 0.1% formic acid in MilliQ water at an estimated concentration of 0.1 µg/µL based on gel stain intensity.

Crosslinking of human red cell membranes

Membrane cytoskeletons and intact human red cell membranes were isolated essentially as previously described^{31, 32}. Crosslinking of the intact membrane was performed in cell lysis buffer (5 mM sodium phosphate, 1 mM EDTA, 0.1 mM DFP, pH 8.0). Crosslinking of the cytoskeleton was performed in 20 mM HEPES, 125 mM NaCl, 3.75 mM CaCl₂, 2.5 mM MgCl₂, at pH 7.4. Both sets of reactions were carried out at 0.5 mg/ml protein concentration at 0°C using 10 mM EDC / 5 mM sulfo-NHS. Aliquots were removed and quenched with dithiothreitol at 15 min and 30 min for the intact membrane sample, and at 15 min, 60 min, and 120 min for the cytoskeleton sample. 200 µg of each of the control and crosslinked samples was digested with trypsin using the FASP method³³.

Construction of protein databases for red cell membrane samples

Each untreated control of the intact membrane and isolated membrane cytoskeleton was searched using MaxQuant version 1.3.0.5³⁴ against a Uniref100 human database (March 2013, Protein Information Resource, Georgetown University, Washington DC) combined with a list of common contaminant (trypsin, keratins, etc.) and a decoy database prepared by reversing each protein sequence. The combined database has a total of 234,648 entries. MS/MS spectra were searched using trypsin specificity without the Pro restriction with up to three missed cleavages, a 10 ppm precursor mass tolerance, 0.5 amu fragment ion mass tolerance, static modification of Cys (carboxamidomethylation +57.02146 Da), and up to three variable modifications for Met oxidation (+15.99492 Da) and protein N-terminal acetylation (+42.01056 Da). Minimum peptide length was set at 7 residues. Consensus protein lists were generated with false discovery rates (FDR) of <1% at both peptide and protein levels. Peptides common between multiple protein groups were re-assigned solely to the group with the highest number of peptides. Proteins were required to be identified by at least 2 peptides.

Mass spectrometry

Tryptic digests of crosslinked and untreated control samples were analyzed on an LTQ-Orbitrap XL (Thermo Scientific, Waltham, MA) coupled to a NanoACQUITY UPLC system (Waters, Milford, MA) with a column heater maintained at 40°C. The specification of the UPLC system, the buffer, and the 85-min gradient used for GST and myoglobin samples was set as previously described³⁵. Details for the 4-hr gradient used for spectrin heterodimer samples were as described in³⁶. The injection volume for each crosslinked sample was adjusted to maintain a consistent load of about 0.5 µg total peptides. The mass spectrometer was set to scan over the 400–2000 m/z range. During the first round of LC-MS/MS acquisition (discovery, low-resolution MS/MS), full scans were acquired in the Orbitrap in profile mode at 60,000 resolution followed by data dependent MS/MS scans in the LTQ ion trap for the six most abundant ions. Monoisotopic precursor selection was enabled and charge state screening was activated to reject +1 and +2 precursors. Dynamic exclusion was set for the duration of 45 sec on an 85-min run or 120 sec on a 4-hr run after a repeat count of 1. AGC targets were set at 10⁶ for full MS scans with a maximum injection time of 500 ms, and at 10⁴ with a maximum injection time of 100 ms for MS/MS scans. Isolation window width was set at 2.5 amu. For the subsequent targeted high-resolution LC-MS/MS runs, both full MS and dependent MS/MS scans were acquired at 15,000 resolution. CID fragmentation was performed in the ion trap followed by detection in the Orbitrap. Monoisotopic precursor selection was disabled and the dynamic exclusion duration was reduced to 12 sec. Isolation window width was raised to 4.0 amu to compensate for increased space charging. Minimum signal threshold for triggering MS/MS scans was set at 50,000 ion counts. The AGC target for MS/MS scans in the Orbitrap was set at 10⁵ with a maximum injection time of 500 ms.

Label-free comparison

In this study, Rosetta Elucidator software (version 3.1, Rosetta Biosoftware, Seattle, WA) was utilized to perform label-free quantitative comparisons of discovery LC-MS/MS data, although any label-free quantitative comparison software capable of aligning and comparing LC-MS data should be applicable. First, LC-MS data was trimmed to the region containing significant peptides. Generally this involved deleting the first 10–15 min and the last 10–15 min, which contained mostly noise. Parameters for the retention time alignment and feature (discrete precursor m/z signal) identification were set as previously described^{35, 37}, with the addition of the following filters: peak time score 0.7, peak m/z score 0.8, and charge state from 3 to 5. Then, the label-free quantitation data at isotope group level, where each isotopic envelope was collapsed into a monoisotopic m/z and charge state, and precursor signals that are at least 10-fold enriched in a crosslinked sample compare to the corresponding control were designated as putative crosslinked peptides.

ZXMiner software

The ZXMiner was developed to expedite data acquisition stages, to process LC-MS/MS data, and to identify crosslinked peptides, as highlighted in the diagram in Figure 1.

To match observed m/z precursors to theoretical crosslinks, the masses of putative crosslinked peptides defined by label-free quantitation were compared to the list of theoretical crosslink masses by ZXMiner using 5 ppm mass tolerance when the LC-MS data was obtained at 60,000 resolution, and 15 ppm for 15,000 resolution data. Theoretical crosslinked peptides were generated *in silico* by ZXMiner based on an input amino acid sequence database, protease reactivity, and expected crosslinker chemistry (trypsin and EDC, i.e. amines to carboxyl groups, in our case). A database consisted of only target protein sequences was considered for the purpose of determining putative crosslinked peptides, and a decoy database was later added when final crosslink identifications were made, as indicated in Figure 1. Full tryptic specificity was used. A static Carbamidomethyl modification for cysteine (+57.02146 Da) and variable oxidation of methionine (+15.99492 Da) were considered. Two or three incomplete cleavages were allowed and individual peptide size was limited to 5–50 amino acids for linear peptides prior to considering crosslinking thereof.

To match MS/MS spectra to theoretical peptides, putative crosslinked peptides where precursor ions matched to theoretical crosslink precursor ion m/z values were further evaluated by ZXMiner. Low-resolution MS/MS spectra from the initial discovery LC-MS/MS run were pre-processed by applying a peak intensity threshold of 10 ion counts. High-resolution MS/MS spectra were subjected to two preprocessing steps: applying a peak intensity threshold of 1,000 ion counts and de-isotoping. Our deisotoping strategy was implemented as described in^{34, 38}. Mass tolerance for the isotopic window spacing was set at 20 ppm and a cutoff of 0.6 was used for the Chi-square test when comparing observed intensity profile of an isotopic envelope to the expected pattern derived from averaging. For linear peptides, all theoretical b-ions and y-ions were generated and compared to observed spectra. For crosslinked peptide, all possible locations of the crosslinked site and their corresponding b-ions and y-ions were calculated. Ions containing less than six amino acids were assigned a charge state of +1. Ions containing more than 12 amino acids and ions containing the crosslinked site with intact partner peptide were assigned a minimum charge state of +2. Ions not containing the crosslinked site were not allowed to attain the precursor charge state. All other ions were allowed to assume any charge state from +1 up to the precursor charge state. The list of theoretical b-ion and y-ion m/z values generated using these rules was then compared to the m/z peaks in the preprocessed MS/MS spectrum. Mass tolerance was set to 0.5 Da for low-resolution data and 15 ppm for high-resolution data. If

multiple theoretical ions matched up to the same observed m/z peak, the alternative with the smallest mass error was selected. After all possible b-ion and y-ion matches had been assigned, neutral losses of the matched ions were generated and compared to the remaining unmatched observed m/z peaks. For low-resolution MS/MS data, up to only one neutral loss of water and one neutral loss of ammonia were considered. Up to two neutral losses were allowed for the precursor ions. For high-resolution data, these limits were doubled. For a-ions, a1 to a5 were considered. Neutral loss of CH₃SOH from the precursor ion was taken into account when oxidized methionine was present, as this loss was frequently observed. In addition, the m/z for the ¹³C ion was generated for low-resolution data since isotopic envelopes could not be detected and collapsed and for larger peptides, the ¹³C ion sometimes predominated.

To score the quality of matches between observed and theoretical MS/MS spectra, three scoring functions were used to evaluate the correlation between theoretical b-ion and y-ions and observed m/z peaks in the MS/MS spectrum, including: “Peak Coverage” which defines the proportion of observed peaks that matched to some theoretical ions, “Intensity Coverage” which describes the fraction of the total observed peak intensity that were explained by the theoretical ions, and “Ion Coverage” which corresponds to the proportion of the b-ions and y-ions that were found in the observed spectrum. For the purpose of computing Peak Coverage and Intensity Coverage from high-resolution MS/MS spectra, only peaks with intensity at least 5,000 ion counts (after de-isotoping) were considered. All peaks passing the 1,000 ion count threshold were considered when calculating Ion Coverage. The corresponding peak intensity thresholds for low-resolution MS/MS spectra were 50 and 10 ion counts, respectively. The geometric mean of these three scores (GM score) was then calculated and used to rank the quality of identification. For each observed MS/MS spectrum, the matched theoretical peptide with the highest GM score was reported as the identification. To collapse the result at the peptide level, the corresponding MS/MS spectrum with the highest GM score was reported for each identified crosslinked peptide (distinct sequence, charge state, or variable modification).

The FDR was estimated using a decoy database of reversed target protein sequences. The formula for calculating the FDR, which takes into account the inherent bias in GM scores toward “hybrid” crosslinks – crosslinks consisting of one peptide from a target protein and another from a decoy sequence – compared to crosslinks involving two decoy sequences, was adapted from³⁹.

The source code and compiled version of ZXMiner will be available for download from <https://shiek-db.wistar.upenn.edu/proteomics/>.

Input parameters for StavroX, Crux, pLink, and MassMatrix

StavroX v2.3.4—The definition for EDC and trypsin without the proline restriction was added to the software. Up to two missed cleavages were allowed. A static Carbamidomethyl modification for cysteine and variable oxidation of methionine were selected. Mass tolerance was set at 10 ppm for the precursor and 0.5 Da for the fragment. Mass limit was set to be between 300–4000 Da. Signal-to-noise threshold was set at 1. A score cutoff of 50 was selected to control the FDR to within 5%.

pLink (Feb 2013 release)—Enzyme, amino acid modifications, and precursor mass tolerance were set as described above. However, the number of missed cleavage could not be specified. Fragment mass tolerance was set at 15 ppm for high-resolution MS/MS data. Instrument fragmentation mode was set to CID. Maximal E-value was set at 1.00. FDR was automatically fixed at 5% within the software.

Crux v1.40—Enzyme and precursor mass tolerance were set as described above. Cysteine residues were considered as Carbamidomethyl modified. Minimal peptide length was set to five residues and up to two missed cleavages were allowed. Monoisotopic mass option was selected (as opposed to using average masses). Fragment mass tolerance and variable modification could not be specified. Also, crosslinking rules involving the protein N- and C-terminus could not be defined. The outputted Bonferroni corrected FDR values were used to control the FDR to within 5%.

MassMatrix v1.3.2—Amino acid modifications and mass tolerances were set as generally described above. Peptide length limit was set to be 5–50 residues. Up to two missed cleavages were allowed. Minimal pp and pp_tag scores were set at 0.1 and 0.01, respectively as per the suggestion in the software’s manual for analyzing crosslink datasets. The definition of trypsin without the Proline restriction was selected. Due to the limitation on the allowed definition of crosslinker, crosslinking rules involving the protein N- and C-terminus could not be defined. Crosslink searches also had to be split into two separate runs – one with Lys-Asp crosslinking and another with Lys-Glu – and manually combined.

RESULTS

Development and Optimization of the multi-tiered MS analysis for zero-length crosslinks

An overview of the multi-tiered LC-MS/MS data acquisition strategy and associated software for analysis of zero length crosslinked peptides is summarized in Figure 1. This approach involves: 1) label-free comparison of crosslinked and control samples, 2) matching of theoretical crosslinked peptides to putative crosslinked peptides using low-resolution MS/MS data, 3) targeted mass spectrometry to acquire high-resolution MS/MS data on potential crosslinked peptides, and 4) automated identification of crosslinked peptides. Individual steps in the procedure were optimized primarily using standard proteins with known crystal structures with further testing on larger protein complexes. Because zero-length crosslinking does not allow incorporation of isotope tags or affinity tags to aid in identification of crosslinked peptides, we found that label-free comparison of high-resolution LC-MS data was the most effective method for filtering out many non-crosslinked precursors. For each protein or protein complex of interest, LC-MS data for an untreated control sample were aligned against data from all crosslinking reactions for that protein (typically multiple reaction times and temperatures). Ideally, all crosslinked peptides would produce signals that are unique to the crosslinking reactions and completely absent in the control. However, due to the nature of the LC-MS peak alignment algorithm implemented in the Rosetta Elucidator software, unrelated signals in the control sample that overlap with non-monoisotopic peaks in the isotopic envelopes of true crosslinked peptides, produce non-zero intensities in the control. To compensate for this shortcoming, any precursor signals at least 10-fold enriched in at least one of the crosslinked samples compared to the untreated control were designated as candidate crosslinked peptides (Supplementary Figure S1). Larger fold differences can be set to minimize the number of signals that need to be considered, albeit at the risk of losing a few crosslinked peptide signals. For simplicity, peptide ions either unique to the crosslinked samples or enriched by at least 10-fold based on the label-free comparison will be referred to below as “enriched” signals or candidate crosslinked peptides. It is important to note that when isotopic envelopes were detected and collapsed into representative monoisotopic peaks prior to label-free peak alignment – the strategy described in³⁴, all identified crosslinked peptides in our datasets had zero intensities in the control sample (data not shown).

The list of candidate crosslinked peptides from the label-free comparison is subsequently narrowed down by comparing precursor masses enriched in the crosslinked samples to those

of theoretical crosslinks of peptides from the protein(s) analyzed. The utility of this step is dependent upon the size of the protein or protein complex being analyzed because the number of theoretical crosslinked peptides increases with the square of the amount of unique protein sequence, resulting in more random matches between observed and theoretical precursor masses. For GST (26 kDa unique sequence; 52 kDa homodimer), about 10% of the enriched signals matched a theoretical crosslinked peptide, but for large protein complexes like red cell spectrin heterodimers (526 kDa unique sequence) more than half of the enriched signals matched a theoretical crosslinked peptide. Furthermore, while each enriched signal from GST usually matched a single theoretical crosslinked peptide, enriched signals from spectrin heterodimer matched an average of about 30, and as many as 50 distinct theoretical crosslinked peptides. This rapid increase in complexity, even when label-free comparisons were used, illustrated the need for improved data acquisition and data analysis tools.

Even when moderate-sized proteins were analyzed and a stringent 5 ppm precursor mass tolerance was applied, most matches between enriched precursor signals and theoretical crosslinked peptides occurred by chance. Less than 5% of these initial matches were subsequently identified as true crosslinked peptides, illustrating further refinement of the candidate crosslinked peptide list was needed. As the initial discovery LC-MS/MS analysis also produced low-resolution MS/MS spectra, these data were further evaluated by comparison to the expected fragment spectra of theoretical crosslinked peptides to further distinguish the best candidate crosslinked peptides. Three fundamental scoring functions were used to evaluate the quality of such comparisons, including: Peak Coverage, Intensity Coverage, and Ion Coverage (see Experimental Procedures). The geometric mean of these three scores (GM score) was used to rank crosslinked peptide identifications. Based upon initial tests, a preliminary GM score threshold for filtering low-resolution MS/MS spectra was set at 0.4 to ensure high crosslinked peptide coverage. Lastly, to ensure that every candidate crosslinked peptide would be targeted at least once in the subsequent high-resolution LC-MS/MS runs, no more than 80–100 precursors were generally assigned to each 85-min run. To ensure comprehensive detection of crosslinked peptides in initial evaluations, four targeted LC-MS/MS runs were used for GST and one targeted run for myoglobin sample. However, subsequent analyses showed that the number of targeted runs could be reduced with minimal impact on depth of analysis. Another effective way of reducing the number of required runs and required crosslinked sample was to use a longer gradient so that more targeted MS/MS acquisitions could be fit into a single run. For example, a 4-hr LC gradient could easily accommodate 300–400 targeted precursors per LC-MS/MS run on an Orbitrap XL instrument, and only two targeted 4-hr runs were required for each spectrin heterodimer sample to analyze all candidates (see below). Also, newer, faster instruments should be able to accommodate even more targeted precursors per run.

Analysis of GST and myoglobin

To evaluate our strategy and refine scoring schemes, we performed chemical crosslinking experiments using EDC/sulfo-NHS on GST and myoglobin, whose crystal structures are available (PDB: 1GTA and 1YMB, respectively). Using high-resolution targeted LC-MS/MS data, each putative crosslinked peptide was annotated as a true crosslink or non-crosslink based on alpha Carbon-alpha Carbon (C α -C α) distances in the pertinent crystal structure and review of MS/MS spectra annotations. Also, the influence of mobile regions in the crystal structures (higher b-factors) and involvement of likely flexible regions of the molecule such as loops, subunit interfaces, and the protein termini on crosslinkable residue distances were examined. Based on a typically accepted C α -C α distance limit of 12 Å for zero-length crosslinks⁴⁰, crosslinks occurring between residues whose alpha Carbons are at most 12 Å apart, and where good matches between observed and theoretical spectra occurred,

were labeled as true positive crosslinks. To further verify these assignments, the fit between expected and observed fragment ions was manually inspected. This 12 Å C α -C α distance limit corresponds roughly to the sum of fully-extended lengths of the side chains of lysine and glutamate or aspartate (K $\approx$ 6.3 Å, E $\approx$ 3.8 Å, and D $\approx$ 2.5 Å) plus an additional approximate 2 Å that might be contributed by uncertainty due to molecular dynamics in solution, or crystal structure resolution, or both. Crosslinks whose C α -C α distances exceeded 20 Å were automatically labeled as non-crosslinks, as the high level of molecular flexibility in solution needed to account for such events seemed unlikely to yield resolvable structures in crystallographic analyses. C α -C α distances between 12 and 20 that involved residues located in regions likely to be flexible were tentatively considered true positives. By visual inspection of the structure, each crosslink was also required to occur between residues with a clear path between the two side chains to qualify as a true positive crosslink, and as noted above, MS/MS spectra were carefully scrutinized to confirm the assignment as a positive crosslink.

With less than 1% FDR (GM score > 0.38), we identified 25 crosslinked peptides (counting different methionine oxidation states and charge states) in the GST sample (Table 1), all of which were confirmed to be true positives by the annotation and spectral review criteria described above. These crosslinked peptides correspond to 13 unique sequences and 10 distinct crosslinked sites on the molecule (Figure 2A). The fact that most of the crosslinked peptides were observed at multiple charge states and/or oxidation states of methionine further supports the reported crosslink assignments. Complete annotation of the GST datasets also revealed that the GM score derived from high-resolution MS/MS data is a powerful indicator of true crosslinked peptides (Figure 2B) that is far superior to scores from low-resolution data (Figure 2C). Specifically, the area under the curve for the corresponding ROC curve is 0.99 for GM score derived from high-resolution data and as high as 80% sensitivity can be achieved without any false positives (Figure 2C). Furthermore, all of the true crosslinked peptides with intermediate GM scores (0.2 to 0.4) were also detected at other charge states or with different methionine modification state where the GM score was higher than 0.4 (Figure 2B). For the myoglobin dataset, ZXMiner identified 15 true positive crosslinked peptides corresponding to 7 unique sequences and 6 different crosslinked sites (Figure 3A and Table 1). The GM score derived from high-resolution MS/MS data also performed very well on this dataset (Figure 3B and 3C) with a perfect area under the ROC curve of 1.00.

Impact of high-resolution MS/MS data

One major limitation of low-resolution MS/MS data that is particularly a problem for both manual and automated crosslinked peptide assignments is the fact that it is impossible to confidently identify isotopic envelopes and charge states. Most crosslinked peptides have charge states of at least +3 and many are +4 or +5 (Table 1), and therefore MS/MS spectra of crosslinked peptides generally contain b- and y-ions with multiple charge states or at least multiple possible charge states. For example, fragmentation of the +5 precursor ion of a crosslinked peptide can produce b- and y-ions that are +1, +2, +3, or +4. In low-resolution spectra, the combined uncertainty of charge state and low mass accuracy for each ion (about ± 0.5 amu) results in numerous random matches between observed and expected fragment ions. High-resolution data acquisition greatly reduces ambiguity in the MS/MS spectra by both reducing mass error from 0.5 amu to low ppm levels and allowing isotopic envelopes to be resolved with accurate determination of charge state and monoisotopic m/z for the majority of the peaks (Figure 4A). We considered discarding all peaks with unidentifiable charge states from the analysis to further reduce uncertainty, but results were very similar to those obtained when ions with unassigned charge states were retained (Supplementary Figure S2). Importantly, the GM scores of most non-crosslinks were reduced to zero when

high-resolution MS/MS spectra was evaluated, indicating that the majority of the matches between observed and theoretical ions in the low-resolution scans occurred by random chance for these non-crosslinks (Figure 4B).

Comparison to StavroX, Ccrux, pLink, and MassMatrix

To evaluate the performance of our software developed specifically for zero-length crosslink datasets, we compared ZXMiner to several common existing crosslink analysis programs, namely StavroX⁴¹, Crux^{42, 43}, pLink⁴⁴, and MassMatrix^{45, 46}. StavroX was considered because it had been shown to outperform a number of other software packages. Crux was chosen since it best represents the concept of adapting well-established database search algorithms for crosslinking, as it improved upon prominent alternatives. Furthermore, Crux was developed based on zero-length crosslink datasets with EDC. pLink was included due to its unique capability to utilize ppm-level fragment mass tolerance and hence it is most suitable for high-resolution MS/MS data. Lastly, MassMatrix was pertinent because it was utilized in several recent studies that utilized zero-length crosslinking^{25–27}. One important special situation when considering zero-length crosslinks is the possibility of crosslinks between adjacent tryptic peptides vs. an incomplete cleavage, as both will have identical precursor masses. Because there is generally not enough MS/MS evidence to reliably distinguish between these two cases, we expected that including these “adjacent-peptide” crosslink identifications will result in high false positive rates. Furthermore, it is important to note that many true adjacent-peptide crosslinks will involve proximal residue positions that provide less crucial structural information. In fact, examination of 11 adjacent-peptide crosslinks collectively reported by StavroX and Crux (Supplementary Table S1) revealed that two of them are false positives (Cα-Cα distances > 26.7 Å) and seven others occur between nearby residues on the protein sequence and in the crystal structure and are therefore uninformative. Hence, reporting adjacent-peptide crosslinks provides very little additional structural information and is not worth the risk of misidentification. In the following comparison of different software, all such crosslinks were excluded for consistency. Exact crosslinked site information within an identified crosslinked complex was also not considered because some programs either do not automatically assign crosslinked sites or do not rank the relative fit of alternative linkages.

GST crosslink datasets were used to compare the performance of ZXMiner against existing crosslinking analysis software tools. Raw files containing low-resolution MS/MS data were input into Crux^{42, 43}, StavroX⁴¹, and MassMatrix^{45, 46}. High-resolution MS/MS runs of the same crosslinked sample were used for pLink⁴⁴ as this is the only program other than ZXMiner that can utilize high resolution MS/MS data. Raw file format conversions were performed using ProteoWizard v3.0.4472⁴⁷. FDRs were set to be within 5% in all programs for consistency because this level is automatically set in some programs and cannot be adjusted. Overall, a total of 15 unique true crosslink sequences were identified (Table 2) and 8 of them were reported by at least four software tools. However, variations across programs were much greater at the unique precursor level where charge states and methionine oxidation states are considered as different, as out of 43 true positive crosslinked peptides, 16 were uniquely identified by a single software and 12 were reported by only two software (Supplementary Figure S3). Using an FDR of <5%, ZXMiner provided the highest number of identifications at the unique precursor level with 30 positive crosslinked peptides, followed by StavroX (20 crosslinked peptides), Crux (19 crosslinked peptides), and MassMatrix (15 crosslinked peptides), while pLink yielded only 6 crosslinked peptides. Interestingly, when the FDR for each dataset was recalculated based upon the actual observed true positive crosslinks, the precursor level FDR was much higher than that estimated by the other programs with FDRs ranging from 21 to 33%, whereas the actual FDR for ZXMiner increased moderately to 6.3% (Table 2). Furthermore, when the FDR at

the unique sequence level rather than unique precursor level was considered, all programs, including ZXMiner, showed an unacceptably high actual FDR when an input FDR of <5% was used. Because all of these FDRs were unacceptably high, ZXMiner was also run using an FDR of <1%. Importantly, at this level, no false positives assignments were made, and while the number of identifications at the unique sequence level decreased from 15 to 13, this is an acceptable moderate reduction in depth of analysis considering that false positive assignments were eliminated (Table 2)

In addition to high FDRs, StavroX, Crux, and MassMatrix had other significant limitations when analyzing zero-length crosslink datasets that may have contributed to their poor performance in this comparison. For example, MassMatrix only considers crosslinks between a single pair of residues, which meant that for EDC crosslinks, two separate runs were needed for Lys-Glu and Lys-Asp and crosslinks involving protein termini were not considered. On the other hand, Crux could not search for crosslinked peptides with variable modifications and the fragment mass tolerance cannot be specified. Lastly, StavroX reported a number of crosslinks that involved a lysine residue that was located at the C-terminal end of a tryptic peptide as one of the crosslinked sites – a situation that should not occur since crosslinked lysine residues should not be cleaved by trypsin.

Feasibility of using zero-length crosslinking on 526 kDa spectrin heterodimers and intact red cell membranes

To evaluate the feasibility of using our software and data acquisition strategy to analyze large protein complexes, crosslinking experiments using EDC/sulfo-NHS were performed using freshly isolated spectrin heterodimers purified from human red cells. Reactions were quenched at 15, 30, 60, and 120 min timepoints to generate a time-series intensity profile for each peptide precursor. Enriched putative crosslink signals from all four timepoints were combined and analyzed together through the rest of the multi-tiered pipeline. A total of 18 crosslinked peptides were identified that were present by either the 15 or 30 min timepoint (Table 3, Figure 5A). The spectrin heterodimer is quite complex with 4,556 residues of unique sequence (526 kDa), and not surprisingly each observed precursor could be matched to an average of about 30 different theoretical crosslinked peptides when a 5 ppm mass error was used. Typically many of these theoretical crosslinked complexes yielded similar low-resolution GM scores (Table 4). However, when high-resolution MS/MS spectra with their reduced charge state ambiguity and increased mass accuracy were available, the correct crosslink assignment was usually readily singled out as illustrated by the example in Table 4.

In subsequent proof-of-principle analyses of even larger complexes, EDC/sulfo-NHS was used to crosslink intact human red cell membranes as well as the membrane cytoskeleton, which is a subset of the intact membrane containing spectrin, actin, and other associated proteins. Importantly, the entire crosslinked membrane or cytoskeleton sample was digested with trypsin and analyzed by LC-MS/MS, in contrast to the experiments above, where crosslinked complexes were greatly enriched by isolation on 1D SDS gels followed by excision of the pertinent band prior to trypsin digestion. LC-MS/MS data from the untreated controls of both samples were searched against a Uniref100 human database to identify the major protein components in these samples (see Experimental Procedures). Based on these results, the 29 major proteins found in the intact membrane control (Supplementary Table S2) and the 19 major proteins in the isolated membrane cytoskeleton (Supplementary Table S3) were incorporated into databases for analysis of crosslinked membranes and cytoskeletons, respectively. Relative proteins abundances in the control samples were ranked by spectral counts normalized to molecular weight. Due to the greater complexity of these samples, 4-hr LC gradients were used for both the discovery and targeted LC-MS/MS runs. Only two targeted runs were needed for the intact membrane sample and one targeted

run was used for the isolated membrane cytoskeleton. Despite the lack of a crosslinked protein enrichment step and the much greater protein complexity, we were able to identify 9 spectrin-spectrin crosslink precursors corresponding to 6 unique peptide sequences in the intact membrane sample (Table 3) and 20 additional crosslinked peptides involving other membrane proteins (data not shown). For the isolated membrane cytoskeleton sample, 6 spectrin-spectrin crosslinked precursors corresponding to 6 unique peptide sequences were identified (Table 3) and 6 additional crosslinked peptides involving other proteins were identified (data not shown). The incomplete overlap of identified spectrin crosslinked peptides among these complex membrane and cytoskeleton samples and the purified spectrin samples was most likely due to the decreased depth of analysis and the stochastic nature of under-sampling of very complex peptide mixtures by LC-MS/MS. The FDR for both datasets were estimated to be less than 1% (Supplementary Figure S4) and all spectrin-spectrin crosslinked peptides identified in either the intact membranes or the membrane cytoskeleton are in good agreement with current knowledge of the molecular topography of spectrin (Figure 5B).

Graphical evaluation of identified crosslinked peptides

Finally, the ability to determine the exact crosslinked residues in crosslinked complexes containing more than one potential crosslinked site candidate is essential for converting each identified crosslinked peptide complex into distance constraints to support molecular modeling and docking models. This issue is especially important for zero-length crosslink datasets induced by EDC because often at least one of the peptides and sometimes both peptides contain multiple internal crosslinkable residues. For example, the crosslinked complex shown in Figure 6A has six alternative linkages and a number of these will have very similar MS/MS spectra that differ by only a few ions. To facilitate manual review of assigned crosslinked peptides, ZXMiner annotates crosslinked sequences and MS/MS spectra with color-coded b- and y-ions to distinguish between those that contain the crosslinked residues and those that do not. Figure 6B shows the annotated spectra for the correct linkage of the peptides shown in Figure 6A. Actual annotated spectra for selected crosslinked peptides are shown in Supplementary Figure S5 and S6. To further aid manual validation of alternative linkage sites, ZXMiner provides a graphical comparison mode where common and unique b- and y-ions and their relative intensities are color-coded for the highest scoring alternative linkages (Figure 6C).

DISCUSSION

It is important to facilitate the use of zero-length crosslink analysis because the tightest distance constraints between amino acid residues are the most advantageous for structural model verification and refinement. So far, longer crosslinking reagents have been more frequently used, and most studies that utilized zero-length crosslinking were restricted to relatively small protein complexes due to difficulty of analyzing zero-length crosslink data. This hurdle has been removed with the development of our multitiered LC-MS/MS strategy and ZXMiner software as it can automatically analyze data and assign sequences with very high accuracy and good depth of analysis. Graphical tools facilitate optional review of assigned sequences and alternative complexes for confirmation but are not needed for initial crosslink assignments. Examples of annotated crosslinked peptides include those identified in: GST that are located more than 12 Å apart (Supplementary Figure S5) and purified spectrin heterodimers that crosslink between distant sites on the molecule (Supplementary Figure S6). The scoring system derived here, the GM score, is relatively simple but effective, even when megadalton complexes are analyzed. Attempts to construct a better scoring function using linear discriminant analysis (LDA) – a technique which has been previously utilized for Lys-Lys crosslink datasets⁴⁸ – achieved only marginal improvements

here. With LDA, the area under the ROC curve for the GST dataset increased to 0.996, and 86.67% sensitivity could be achieved with no false positives. On the other hand, logistic regressions and nested model analyses of the three coverage scores that comprise the GM score as well as the GM score revealed that the optimal predictor was obtained when only Peak Coverage, Ion Coverage, and GM score were considered. This resulted in 0.999 area under the ROC curve and as high as 96.67% sensitivity was achieved without any false positives. Nested model analyses using deviance⁴⁹ showed that this improvement over the GM score-only model is significant with a p-value of 0.0051. Further evaluation of this program using diverse protein problems and other mass spectrometers are required to determine the generalizability of this alternative scoring model as some differences in optimal scoring strategies may occur when different types of mass spectrometers are utilized.

One potential limitation of the workflow proposed here is the need for multiple LC-MS/MS acquisitions, which can be an important consideration when the amount of crosslinked sample is limited, as might be the case when low yield crosslinked complexes are cut out of 1D SDS gels. However, a number of strategies can be utilized to limit the number of LC-MS/MS runs to one discovery run and one or two targeted LC-MS/MS runs. First, while we typically use a conservative preliminary GM score cutoff of 0.4 to avoid loss of potential crosslinked peptide identifications, higher scores can be used. For example, the low-resolution MS/MS data resulted in four targeted LC-MS/MS runs in the case of the GST in early studies, but subsequent analyses showed that this threshold could be raised to 0.5 without any negative impact on the crosslink identification performance. Raising the preliminary low resolution MS/MS score from 0.4 to 0.5 cuts the number of required targeted runs by half for most datasets. In general, this GM score threshold can be freely adjusted to achieve a desired balance between throughput and depth of analysis. Additionally, the gradient length can be increased to reduce the number of targeted runs needed and thereby conserve samples available in limited amounts, as longer gradients can accommodate more target precursors per run. Noticeably, when a 4-hr gradient was used for the spectrin heterodimer, isolated membrane cytoskeleton, and intact red cell membrane samples, no more than two targeted LC-MS/MS runs were required per sample. Another possibility for improving the throughput of the LC-MS/MS step is to utilize mass spectrometers that can acquire high resolution MS/MS spectra with faster scan speeds such as Thermo Electron's Q Exactive or Orbitrap Fusion. These instruments have the potential to obtain high-resolution LC-MS and in-depth high resolution MS/MS data in a single experiment involving one LC-MS/MS run each for the crosslinked and control samples. One can then use label-free quantitation to compare the MS data from the crosslinked and untreated control samples to identify crosslinked peptide candidates, followed by evaluation of MS/MS spectra from the same run to identify crosslinked peptides.

The comparison with existing crosslink analysis software tools revealed that, at least in the context of the GST dataset, ZXMiner shows very favorable performance with more true positive identifications at both the unique precursor and unique sequence levels and a lower actual false discovery rate (Table 2). Possible reasons why these alternative software packages did not perform well here is because most of them were optimized using datasets generated with different crosslinkers and in a few cases with different types of mass spectrometers. Specifically, MassMatrix and pLink were developed using BS³ crosslinker, while StavroX considered BS²G crosslinker, disulfide bonds, and SBC – an amine-reactive photo-crosslinker. Although Crux was optimized using an EDC zero-length crosslink dataset, due to software design limitations, it was unable to identify any of the ten true positive crosslinked peptides in the GST dataset that had oxidized methionine. Interestingly, most crosslinked peptides identified by a single program were false positives (13 of 15 at the unique sequence level) while most of those identified by at least three programs were correct

identifications (12 out of 14 at the unique sequence level) (Supplementary Figure S3). Hence, as with identification of linear peptides, one can improve depth of analysis and simultaneously reduce false positives by reporting crosslinked peptides identified across multiple software packages. In this regard, all of the crosslinked peptides reported by ZXMiner were also identified by at least one other program, and all but one were identified by at least two others.

Although good depth of analysis is highly desirable in crosslinking experiments, for most applications, the most critical factor is a very low FDR, because even a single false positive crosslink can have profound negative effects. Incorrect crosslink assignments will usually erroneously indicate protein-protein proximity, including an interaction between two proteins that does not exist, and when used to support structural models, incorrect distance constraints will reduce the accuracy of structural model. As noted above, some programs use a fixed 5% FDR and with programs where this is adjustable, an FDR of 5% is often used. Furthermore, as shown in Table 2, the actual FDR at the unique sequence level is far higher than the FDR selected in the analysis. Given the negative impact of incorrect assignments, we feel that the actual FDR must be less than 1%, and preferably should be 0%. The results shown here indicate that our multi-tier analysis strategy with ZXMiner can effectively achieve such low FDR levels while retaining excellent depth of analysis.

The feasibility of applying our zero-length crosslink analysis strategy to interrogate large protein complexes was demonstrated by analyses of 526 kDa spectrin heterodimers, isolated red cell membrane cytoskeletons, and the intact red cell membranes. The latter two samples represent by far the most complex zero-length crosslink dataset to date, as their analysis involved databases with more than 2,000 kDa of unique sequence. A low false discovery rate (< 1%) was maintained throughout, and the majority of the crosslinked peptides identified in the spectrin heterodimer dataset and all spectrin-spectrin crosslinked peptides in the other two are in good agreement with the known general topography of the protein (Figure 5). Interestingly, three crosslinked peptides were observed in the spectrin heterodimer experiment between sites on the molecule that are thought to be large distances from each other (blue dashed lines in Figure 5A). The most likely explanation for these unexpected crosslinks is that because isolated spectrin heterodimers are highly flexible, extended, worm-like molecules, such folding back upon itself can occur with isolated molecules in solution due to the increased molecular freedom relative to proteins in a two-dimensional lattice on the membrane. Another possibility is that these crosslinks are due to random interactions between two separate dimer molecules, although this is not likely as crosslinking of other proteins at similar protein concentrations did not typically show significant intermolecule crosslinking and crosslinked higher oligomers were not observed in these samples.

Finally, the decreased depth of analysis achieved in the far more complex purified spectrin, intact membrane, and cytoskeleton samples and stochastic detection of very weak signals in very complex peptide mixtures are the most likely reasons for the incomplete overlap of identified crosslinked peptides. Specifically, one might expect that most if not all detectable spectrin crosslinks in intact membrane samples would also be detectable in the less complex purified spectrin samples. However, purified spectrin is more than 500 kDa and the linear peptides from a tryptic digest of this protein produce a quite complex background for detection of low abundance crosslinked products. Furthermore, spectrin is the most abundant protein in red cell membrane where it contributes about 25% of the total membrane protein content. Our strategy relies on MS/MS data from the initial discovery LC-MS/MS run to prioritize candidate crosslinks for targeted analysis, and under-sampling of the typically weak crosslinked peptide precursor signals in these complex samples will result in stochastic detection of crosslinked peptides. Inspection of the discovery LC-MS/MS data for the

purified spectrin and intact membrane samples revealed that several of the crosslinks that were uniquely identified in the intact membrane samples were also present in the purified spectrin samples. However, the intensity levels of these crosslinks were very low (in the 10^4 ion count range) in the purified spectrin samples and therefore only poor quality or no MS/MS spectra were obtained for these precursors in the purified protein experiments. Future experiments will develop methods for enhancing the depth of analysis of crosslinked peptides in very complex samples such as intact membranes by fractionating tryptic digest so that larger molar amounts of crosslinked peptides can be injected without overloading the system. Preliminary results using high-pH reverse-phase fractionation of digested crosslinked and control samples show substantial promise. In these experiments, fraction boundary affects in label-free comparisons are avoided by pooling overlapping and broader fractions for the control sample.

CONCLUSIONS

Recent progress in the development and use of non-zero-length crosslinkers with identification-enhancing properties for analysis of large protein complexes and biological systems demonstrate the great utility of CX-MS. However, it is also important to expand the capacities of zero-length crosslinkers because they provide the tightest distance constraints and are therefore potentially the most valuable crosslinkers for structural studies. Our multi-tiered strategy for acquiring high resolution MS/MS spectra of crosslink candidates coupled with the software tool, ZXMiner, compare favorably with existing crosslink analysis programs. Specifically, ZXMiner provides excellent depth of analysis for identifying crosslinks, while maintaining a very low FDR, which is the most critical parameter in crosslink identification. In the context of structural studies where even a single misidentification can lead to an incorrect protein-protein interaction assignment and inaccurate molecular models, in depth identification of crosslinked peptides with essentially no false positives is especially invaluable. Furthermore, we have demonstrated the scalability of our strategy by analyzing the 526 kDa spectrin heterodimer and the multimegadalton-scale intact red cell membrane and membrane cytoskeleton. These samples represent some of the largest zero-length crosslink datasets analyzed to date.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank P. Hembach for technical assistance and the Wistar Institute Proteomics Core for mass spectrometry analyses. This work was supported by the US National Institute of Health grants R01HL038794 (to D.W.S.) and P30CA010815 (NCI core grant to the Wistar Institute), as well as a Philadelphia Health Care Trust Fellowship (to S.S.).

REFERENCES

1. Rappsilber J, Siniosoglou S, Hurt EC, Mann M. A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry. *Anal Chem.* 2000; 72(2):267–275. [PubMed: 10658319]
2. 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):5802–5806. [PubMed: 10811876]
3. Chen ZA, Jawhari A, Fischer L, Buchen C, Tahir S, Kamenski T, Rasmussen M, Lariviere L, Bukowski-Wills JC, Nilges M, Cramer P, Rappsilber J. Architecture of the RNA polymerase II-

- TFIIF complex revealed by cross-linking and mass spectrometry. *EMBO J.* 2010; 29(4):717–726. [PubMed: 20094031]
4. Lasker K, Forster F, Bohn S, Walzthoeni T, Villa E, Unverdorben P, Beck F, Aebersold R, Sali A, Baumeister W. Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. *Proc Natl Acad Sci U S A.* 2012; 109(5):1380–1387. [PubMed: 22307589]
 5. Leitner A, Joachimiak LA, Bracher A, Monkemeyer L, Walzthoeni T, Chen B, Pechmann S, Holmes S, Cong Y, Ma B, Ludtke S, Chiu W, Hartl FU, Aebersold R, Frydman J. The molecular architecture of the eukaryotic chaperonin TRiC/CCT. *Structure.* 2012; 20(5):814–825. [PubMed: 22503819]
 6. Herzog F, Kahraman A, Boehringer D, Mak R, Bracher A, Walzthoeni T, Leitner A, Beck M, Hartl FU, Ban N, Malmstrom L, Aebersold R. Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry. *Science.* 2012; 337(6100):1348–1352. [PubMed: 22984071]
 7. Jacobsen RB, Sale KL, Ayson MJ, Novak P, Hong J, Lane P, Wood NL, Kruppa GH, Young MM, Schoeniger JS. Structure and dynamics of dark-state bovine rhodopsin revealed by chemical cross-linking and high-resolution mass spectrometry. *Protein Sci.* 2006; 15(6):1303–1317. [PubMed: 16731966]
 8. 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):303–313. [PubMed: 12888339]
 9. Leitner A, Walzthoeni T, Kahraman A, Herzog F, Rinner O, Beck M, Aebersold R. Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. *Mol Cell Proteomics.* 2010; 9(8):1634–1649. [PubMed: 20360032]
 10. Mouradov D, King G, Ross IL, Forwood JK, Hume DA, Sinz A, Martin JL, Kobe B, Huber T. Protein structure determination using a combination of cross-linking, mass spectrometry, and molecular modeling. *Methods Mol Biol.* 2008; 426:459–474. [PubMed: 18542884]
 11. Paramelle D, Miralles G, Subra G, Martinez J. Chemical cross-linkers for protein structure studies by mass spectrometry. *Proteomics.* 2013; 13(3–4):438–456. [PubMed: 23255214]
 12. Rappsilber J. The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *J Struct Biol.* 2011; 173(3):530–540. [PubMed: 21029779]
 13. Sinz A. Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein-protein interactions. *Mass Spectrom Rev.* 2006; 25(4):663–682. [PubMed: 16477643]
 14. Sinz A. Investigation of protein-protein interactions in living cells by chemical crosslinking and mass spectrometry. *Anal Bioanal Chem.* 2010; 397(8):3433–3440. [PubMed: 20076950]
 15. 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):5101–5106. [PubMed: 16097745]
 16. 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):1927–1934. [PubMed: 11354472]
 17. Back JW, Hartog AF, Dekker HL, Muijsers AO, de Koning LJ, de Jong L. A new crosslinker for mass spectrometric analysis of the quaternary structure of protein complexes. *J Am Soc Mass Spectrom.* 2001; 12(2):222–227. [PubMed: 11212007]
 18. 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.* 2011; 10(1) M110 002212.
 19. Kang S, Mou L, Lanman J, Velu S, Brouillette WJ, Prevelige PE Jr. Synthesis of biotin-tagged chemical cross-linkers and their applications for mass spectrometry. *Rapid Commun Mass Spectrom.* 2009; 23(11):1719–1726. [PubMed: 19412923]
 20. Petrotchenko EV, Serpa JJ, Borchers CH. An isotopically coded CID-cleavable biotinylated cross-linker for structural proteomics. *Mol Cell Proteomics.* 2011; 10(2) M110 001420.

21. 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):834–850. [PubMed: 12892908]
22. Yamashiro S, Speicher KD, Speicher DW, Fowler VM. Mammalian tropomodulins nucleate actin polymerization via their actin monomer binding and filament pointed end-capping activities. *J Biol Chem.* 2010; 285(43):33265–33380. [PubMed: 20650902]
23. Kolenko P, Rozbesky D, Vanek O, Kopecky V Jr, Hofbauerova K, Novak P, Pompach P, Hasek J, Skalova T, Bezouska K, Dohnalek J. Molecular architecture of mouse activating NKR-P1 receptors. *J Struct Biol.* 2011; 175(3):434–441. [PubMed: 21600988]
24. Zhao C, Gao Q, Roberts AG, Shaffer SA, Doneanu CE, Xue S, Goodlett DR, Nelson SD, Atkins WM. Cross-linking mass spectrometry and mutagenesis confirm the functional importance of surface interactions between CYP3A4 and holo/apo cytochrome b(5). *Biochemistry.* 2012; 51(47):9488–9500. [PubMed: 23150942]
25. Bojja RS, Andrade MD, Merkel G, Weigand S, Dunbrack RL Jr, Skalka AM. Architecture and assembly of HIV integrase multimers in the absence of DNA substrates. *J Biol Chem.* 2013; 288(10):7373–7386. [PubMed: 23322775]
26. Ido K, Kakiuchi S, Uno C, Nishimura T, Fukao Y, Noguchi T, Sato F, Ifuku K. The conserved His-144 in the PsbP protein is important for the interaction between the PsbP N-terminus and the Cyt b559 subunit of photosystem II. *J Biol Chem.* 2012; 287(31):26377–26387. [PubMed: 22707728]
27. Liu H, Huang RY, Chen J, Gross ML, Pakrasi HB. Psb27, a transiently associated protein, binds to the chlorophyll binding protein CP43 in photosystem II assembly intermediates. *Proc Natl Acad Sci U S A.* 2011; 108(45):18536–18541. [PubMed: 22031695]
28. Mayne SL, Patterson HG. Bioinformatics tools for the structural elucidation of multi-subunit protein complexes by mass spectrometric analysis of protein-protein cross-links. *Brief Bioinform.* 2011; 12(6):660–671. [PubMed: 22101029]
29. Harper SL, Li D, Maksimova Y, Gallagher PG, Speicher DW. A fused alpha-beta "minispectrin" mimics the intact erythrocyte spectrin head-to-head tetramer. *J Biol Chem.* 2010; 285(14):11003–11012. [PubMed: 20139081]
30. Beer LA, Tang HY, Barnhart KT, Speicher DW. Plasma biomarker discovery using 3D protein profiling coupled with label-free quantitation. *Methods Mol Biol.* 2011; 728:3–27. [PubMed: 21468938]
31. Speicher DW, Weglarz L, DeSilva TM. Properties of human red cell spectrin heterodimer (side-to-side) assembly and identification of an essential nucleation site. *J Biol Chem.* 1992; 267(21):14775–14782. [PubMed: 1634521]
32. Sheetz MP. Integral membrane protein interaction with Triton cytoskeletons of erythrocytes. *Biochim Biophys Acta.* 1979; 557(1):122–134. [PubMed: 549630]
33. Wisniewski JR, Zougman A, Nagaraj N, Mann M. Universal sample preparation method for proteome analysis. *Nat Methods.* 2009; 6(5):359–362. [PubMed: 19377485]
34. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol.* 2008; 26(12):1367–1372. [PubMed: 19029910]
35. Beer LA, Tang HY, Sriswasdi S, Barnhart KT, Speicher DW. Systematic discovery of ectopic pregnancy serum biomarkers using 3-D protein profiling coupled with label-free quantitation. *J Proteome Res.* 2011; 10(3):1126–1138. [PubMed: 21142075]
36. Tang HY, Beer LA, Tanyi JL, Zhang R, Liu Q, Speicher DW. Protein isoform-specific validation defines multiple chloride intracellular channel and tropomyosin isoforms as serological biomarkers of ovarian cancer. *J Proteomics.* 2013; 89C:165–178. [PubMed: 23792823]
37. Neubert H, Bonnert TP, Rumpel K, Hunt BT, Henle ES, James IT. Label-free detection of differential protein expression by LC/MALDI mass spectrometry. *J Proteome Res.* 2008; 7(6):2270–2279. [PubMed: 18412385]
38. Senko MW, Beu SC, McLafferty FW. Determination of Monoisotopic Masses and Ion Populations for Large Biomolecules from Resolved Isotopic Distributions. *J Am Soc Mass Spectr.* 1995; 6(4):229–233.

39. 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*. 2012; 9(9):901–903. [PubMed: 22772729]
40. Rozbesky D, Man P, Kavan D, Chmelik J, Cerny J, Bezouska K, Novak P. Chemical crosslinking and H/D exchange for fast refinement of protein crystal structure. *Anal Chem*. 2012; 84(2):867–870. [PubMed: 22196380]
41. Gotze M, Pettelkau J, Schaks S, Bosse K, Ihling CH, Krauth F, Fritzsche R, Kuhn U, Sinz A. StavroX--a software for analyzing crosslinked products in protein interaction studies. *J Am Soc Mass Spectrom*. 2012; 23(1):76–87. [PubMed: 22038510]
42. McIlwain S, Draghicescu P, Singh P, Goodlett DR, Noble WS. Detecting cross-linked peptides by searching against a database of cross-linked peptide pairs. *J Proteome Res*. 2010; 9(5):2488–2495. [PubMed: 20349954]
43. Park CY, Klammer AA, Kall L, MacCoss MJ, Noble WS. Rapid and accurate peptide identification from tandem mass spectra. *J Proteome Res*. 2008; 7(7):3022–3027. [PubMed: 18505281]
44. Yang B, Wu YJ, Zhu M, Fan SB, Lin J, Zhang K, Li S, Chi H, Li YX, Chen HF, Luo SK, Ding YH, Wang LH, Hao Z, Xiu LY, Chen S, Ye K, He SM, Dong MQ. Identification of cross-linked peptides from complex samples. *Nat Methods*. 2012; 9(9):904–906. [PubMed: 22772728]
45. Xu H, Freitas MA. MassMatrix: a database search program for rapid characterization of proteins and peptides from tandem mass spectrometry data. *Proteomics*. 2009; 9(6):1548–1555. [PubMed: 19235167]
46. Xu H, Zhang L, Freitas MA. Identification and characterization of disulfide bonds in proteins and peptides from tandem MS data by use of the MassMatrix MS/MS search engine. *J Proteome Res*. 2008; 7(1):138–144. [PubMed: 18072732]
47. Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics*. 2008; 24(21):2534–2536. [PubMed: 18606607]
48. 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):315–318. [PubMed: 18327264]
49. McCullagh, P.; Nelder, JA. Generalized linear models. 2nd ed. London; New York: Chapman and Hall; 1989. p. 511
50. Li D, Tang HY, Speicher DW. A structural model of the erythrocyte spectrin heterodimer initiation site determined using homology modeling and chemical cross-linking. *J Biol Chem*. 2008; 283(3):1553–1562. [PubMed: 17977835]

FIGURE 1. The multi-tiered data acquisition and analysis pipeline for zero-length crosslinks
Steps highlighted in gold boxes are performed automatically in ZXMiner.

FIGURE 2. Crosslink analysis using GST

(A) Locations of identified crosslinks on the crystal structure of GST homodimer (PDB: 1GTA). Lys residues are highlighted in blue and Glu and Asp are in red. Black lines connect the two alpha Carbons of each crosslink. Crosslinks between residues whose Ca-Ca distances are significantly larger than 12 Å were highlighted in orange. (B) Scatter plot showing the relationship between GM scores derived from high-resolution MS/MS data and Ca-Ca distances for all crosslinked peptide candidates in the GST dataset. A few crosslinks located in regions likely to exhibit increased flexibility such as loops or inter-subunit interfaces exceeded the expected 12 Å maximum Ca-Ca distance. (C) ROC curves showing

the superior performance of high-resolution MS/MS data (area under the curve = 0.99) compared to low-resolution data (area under the curve = 0.80).

FIGURE 3. Crosslink analysis using myoglobin

(A) Locations of identified crosslinks on the crystal structure of myoglobin monomer (PDB: 1YMB). Lys residues are highlighted in blue and Glu and Asp are in red. Black lines connect the two alpha Carbons of each crosslink. (B) Scatter plot showing the relationship between GM scores derived from high-resolution MS/MS data and Ca-Ca distances for all crosslinked peptide candidates in the GST dataset. (C) ROC curves showing the perfect performance (area under the curve = 1.00) of high-resolution GM score in this dataset.

FIGURE 4. Impact of high-resolution MS/MS spectra

(A) At high-resolution, isotopic envelopes, monoisotopic m/z, and charge states can be determined for the majority of the MS/MS peaks. The bottom panels zoom in on a small m/z window to highlight how de-isotoping greatly reduces the complexity and ambiguity of the spectrum. (B) Bar plots illustrating the improved discriminatory power of GM scores obtained from high-resolution MS/MS data (right panel) compared to those based on low-resolution data (left panel). The plot for low-resolution data was truncated because only crosslinked peptide candidates with low-resolution GM score of at least 0.4 were further analyzed.

FIGURE 5. Zero-length crosslink analysis of spectrin as purified heterodimers and in red cell membranes

Because the crystal structure of the entire protein is unavailable, approximate locations of identified crosslinks are plotted on the widely-accepted schematic of spectrin domain structure^{29, 50}. As indicated, the long highly flexible α and β chains laterally associate laterally along the length of both chains. Crosslinks that fit the known domain structure and lateral alignment of the subunits are indicated by red lines, while those indicative of the protein folding back upon itself are shown by dashed blue lines. (A) Purified heterodimers in solution. (B) Intact membranes and isolated membrane cytoskeletons combined.

FIGURE 6. Graphical annotations of candidate crosslinked peptides

(A) Example of a crosslinked peptide with multiple possible linkages. (B) ZXMiner annotates each candidate crosslinked peptide to the matched MS/MS spectrum with identified b- and y-ions (red and blue labels, respectively). Contributions from each peptide are separated into two panels to individually annotate b-ions and y-ions from the two crosslinked peptides. Peaks that match the crosslinked peptide are highlighted in green. (C) A detailed graphical comparison mode of ZXMiner facilitates comprehensive side-by-side evaluation of alternative crosslinked sites (numbered 1 through 6) within the crosslinked peptide complex. Uninformative peak assignments shared by all possible linkages are colored in green while linkage-specific fragments are color-coded according to their relative

intensity (red: > 50,000 ion counts, blue: between 5,000 and 50,000 ion counts, and magenta: between 1,000 and 5,000 ion counts). Red arrow indicates the correct interpretation.

TABLE 1
GST and myoglobin crosslinked peptides identified at FDR of less than 1%

Unique Sequence ID	Charge	MH+	Crosslink Sequence	Ca-Ca Distance Å
<u>GST</u>				
1	3,4	2455.2460	YEEHLY[E]R-{MSPILGYW[K]IK	5.5
1#	5	2471.2400	YEEHLY[E]R-{M [#] SPILGYW[K]IK	5.5
2	4	3508.7050	FELGLEFPNLPYYIDG[D]VK-HNMLGGCP[K]ER	8.7
2#	3	3524.6970	FELGLEFPNLPYYIDG[D]VK-HNM [#] LGGCP[K]ER	8.7
3	4	3878.9360	HNMLGGCP[K]ER-NKKFELGLEFPNLPYYIDG[D]VK	8.7
3#	4	3894.9140	HNM [#] LGGCP[K]ER-NKKFELGLEFPNLPYYIDG[D]VK	8.7
4	3,4	3636.7980	KFELGLEFPNLPYYIDG[D]VK-HNMLGGCP[K]ER	8.7
4#	3,4	3652.7920	KFELGLEFPNLPYYIDG[D]VK-HNM [#] LGGCP[K]ER	8.7
5	3,4	2561.4420	LLL[E]YLEEK-IEAIPQID[K]YLK	9.2
6	4	2717.5410	LLL[E]YLEEK-RIEAIPQID[K]YLK	9.2
7	3	2734.3930	LLLEYL[E]K-YIAD[K]HNMLGGCPK	11.4
7#	3,4	2750.3840	LLLEYL[E]K-YIAD[K]HNM [#] LGGCPK	11.4
8	3,4	2048.0540	[D]F[E]TLK-IAYS[K]DFETLK	11.6
9	5	3318.7510	YIAWPLQGWQATFGGG[D]HPPK-I[K]GLVQPTR	11.8
10	4	1856.0400	LLL[E]YLEEK-YL[K]SSK	12.4
11	3	3068.6050	LP[E]MLK-[K]FELGLEFPNLPYYIDGDVK	12.4
11#	3	3084.6010	LP[E]M [#] LK-[K]FELGLEFPNLPYYIDGDVK	12.4
12	3	3726.9063	YIAWPLQGWQATFGGGDHPP[K]-V[D]FLSKLPEMLK	14.6 [*]
13	3	3035.4860	MFE[D]R-[K]FELGLEFPNLPYYIDGDVK	15.8 [*]
<u>Myoglobin</u>				
1	4,5	1925.068	LFTGHPETL[E]K-F[K]HLK	6.00
2	4	1327.690	ASE[D]LK-FD[K]FK	8.10
3	4	3778.946	{GLS[D]GEWQQVLNVWGK-[K]GHHEAELKPLAQSHATK	8.50
4	4,5	2339.208	LFTGHPETL[E]K-HL[K]TEAEMK	10.90
4#	3,4	2355.202	LFTGHPETL[E]K-HL[K]TEAEM [#] K	10.90
5	3,4,5	2757.448	NDIAA[K]YK-GHH[E]AELKPLAQSHATK	11.80
6	4,5	2885.542	NDIAA[K]YK-KGHH[E]AELKPLAQSHATK	11.80
7	3,5	3094.781	V[E]ADIAGHGQEVLR-HGTVVLTALGGIL[K]K	12.20

Asterisks indicate crosslinks involving flexible regions whose Ca-Ca distances significantly exceed 12 Å.

[#] indicates oxidized Methionine (+15.99492). All Cysteines are carboxyamidomethylated (+57.02146).

{ } indicates the protein N- and C-terminus, respectively. [] indicates crosslinked sites.

TABLE 2
Comparison of GST crosslink identifications using alternative software

For each peptide sequence, charge states of the crosslinked peptide identified by each software package are indicated. Estimated FDRs were derived from the decoy data that each software package provided (not available from MassMatrix). Number of identifications, false positives, and the actual FDRs were calculated directly from this table.

Unique Sequence ID	Peptide Sequence	Co-Co Distance Å	ZXMiner	pLink	StavroX	CruX ^d	Mass Matrix ^b	
True Positives								
1	YEEHLYER-[MSPILGYWKIK	5.5	3,4	3,4,5	3	3,4,5	3,4	5
1#	YEEHLYER-[M#SPILGYWKIK	5.5	5	4,5	-	3,4	N/A	5
2	FELGLEFPNLPYYIDGDVK-HNMLGGCPKER	8.7	4	4	-	-	3,4	4
2#	FELGLEFPNLPYYIDGDVK-HNM#LGGCPKER	8.7	3	3	-	-	N/A	3
3	HNMLGGCPKER-NKKFELGLEFPNLPYYIDGDVK	8.7	4	4,5	-	-	4,5	4
3#	HNM#LGGCPKER-NKKFELGLEFPNLPYYIDGDVK	8.7	4	4	-	-	N/A	5
4	KFELGLEFPNLPYYIDGDVK-HNMLGGCPKER	8.7	3,4	3,4	-	-	3,4	3
4#	KFELGLEFPNLPYYIDGDVK-HNM#LGGCPKER	8.7	3,4	3,4	-	-	N/A	5
5	LLLEYLEEK-IEAIPQIDKYLK	9.2	3,4	3,4	3	3,4	3,4	-
6	LLLEYLEEK-RIEAIPQIDKYLK	9.2	4	4	-	4	-	-
7	YEEHLYER-[M#SPILGYWK	10.0	-	-	-	3,4	N/A	N/A
8	LLLEYLEEK-YIADKHNMLGGCPK	11.4	3	3	3	-	3	3
8#	LLLEYLEEK-YIADKHNM#LGGCPK	11.4	3,4	3,4	-	-	N/A	-
9	DFETLK-IAYSKDFETLK	11.6	3,4	3,4	3	3	3,4	-
10	YIAWPLQGWQATFGGDHPPK]-IKGLVQPTR	11.8	5	5	-	5	5	5
11	LPEMLK-KFELGLEFPNLPYYIDGDVK	12.4	3	3,4	-	3,4	3,4	3
11#	LPEM#LK-KFELGLEFPNLPYYIDGDVK	12.4	3	3	-	4	N/A	4
12	LLLEYLEEK-YLKSSK	12.4	4	4	-	3,4	4	-
13	AEISMLEGAVLDIRYGVSR-YIADKHNM#LGGCPK	13.8 ^c	-	-	-	-	N/A	4
14	YIAWPLQGWQATFGGDHPPK]-VDFLSKLPEMLK	14.6 ^d	-	5	5	-	4,5	5
14#	YIAWPLQGWQATFGGDHPPK]-VDFLSKLPEM#LK	14.6 ^d	3	3	-	3,4	N/A	5
15	MFEDR-KFELGLEFPNLPYYIDGDVK	15.8 ^{c,d}	3	3	3	-	-	-

Unique Sequence ID	Peptide Sequence	Cq-Cq Distance Å	ZXMiner	pLink	StavroX	CruX ^a	Mass Matrix ^b
15#	M#FEDR-KFELGLEPPNLPYYIDGDVK	15.8 ^{c,d}	-	-	3	N/A	3
False Positives							
1	LLLEYLEEK-LVCFKK	16.3	-	-	-	3	-
2	IKGLVQPTR-DEGDK	16.4	-	-	3	4	3
3	IKGLVQPTR-DEGDKWR	16.4	-	-	-	3	-
4	DEGDKWR-LPEMLK	18.5	-	-	-	3,4	3
5	DEGDKWRNK-LPEMLKMFEDR	18.5	-	-	-	-	3
6	VDFLSKLPPEMLKMFEDR-ER	19.0	-	-	-	-	3
6#	VDFLSKLPPEMLKMFEDR-ER	19.0	-	-	-	3	N/A
7	IKGLVQPTR-LLLEYLEEK	19.1	-	-	-	-	3
8	{MSPILGYWKIK-ERAEISMLEGA VLDIR	20.9	-	-	-	4	-
9	ERAEISMLEGA VLDIR YGVSR-WRNKK	21.5	-	-	-	-	3
10	IKGLVQPTR-DFETLK	22.7	-	3	3	-	3
11	HNMLGGCPKER-LVCFKK	23.9	-	-	-	-	3
12	LLLEYLEEK-DEGDKWR	25.2	-	3	-	3	-
13	YEEHL YERDEGDKWR-AEISM#LEGAVLDIR	27.2	-	-	-	-	N/A
14	VDFLSKLPPEMLKMFEDR- YIADKHNNMLGGCPK	28.5	-	-	-	-	5
15	IEAIPQIDKYLK-MFEDR	29.6	-	-	-	3	-
16	LLLEYLEEK-LPEMLKMFEDRLCHK	33.4	-	-	-	-	3
17	DFETLK-LVCFKK	34.2	-	-	-	3	-
Estimated FDR: <1% <5% <5% <5% <5% <5% N/A							
Unique Precursor Level							
True Positives:			25	30	6	20	19
False Positives:			0	2	3	8	9
Actual FDR:			0%	6.3%	33%	29%	32%
True Positives:			13	15	6	10	11
False Positives:			0	2	3	7	9
Actual FDR:			0%	12%	33%	41%	45%
Unique Sequence Level							
True Positives:			13	15	6	10	11
False Positives:			0	2	3	7	9
Actual FDR:			0%	12%	33%	41%	45%

^aCrux cannot identify peptides with variable modifications.

^b MassMatrix cannot identify crosslinks involving the protein terminus as one of the crosslinked sites.

^c Crosslinks between subunits.

^d Crosslinks involving flexible regions (as reflected by the elevated B-factors and/or loops).

TABLE 3
Spectrin crosslinked peptides

Charge	M+H	Sequence	Domain
Purified spectrin heterodimer			
4,5	2799.5039	ADVVEAWIADK-HLLEVEDLLQKHK	α 19- β 3
3	1763.9174	DFLEELEESR-ALGKK	β 5- β 5
4	3671.8910	DGLNEM [#] WADLLELIDTR-LLEVLSGEM [#] LPKPTK	β 14-ABD [*]
3	3486.6290	DLEELEEWISEM [#] LPTACDESYK-KLSGLER	α 15- β 7
5	3551.8268	EFSTIYK-AYFLDGSLLKETGTLESQLEANKR	EF-EF
4	4089.9774	EKEPIVDNTNYGADEEAAGALLKK-DLEDETLWVEER	α 9- β 12
3	2003.9920	ETDDLEQWISEK-PTKGK	β 14-ABD [*]
5	3209.5935	ETDDLEQWISEK-RKLENM [#] YHLFQLK	β 14- β 14
5	3498.7063	ETDDLEQWISEK-YFYTGAEILGLIDEKHR	β 14- β 15
4	2768.3413	FDEFQK-KAENTGVELDDVWELQK	α 11- α 11
3,4	2899.4578	GQQLVEAAEIDCQDLEER-AKLQISR	β 12- β 12
4	2409.2997	KHGLLESAAVAR-VDNVNAFIER	α 7- β 14
3	2083.0342	LADDEDYK-VQKQVFEK	α 6- α 6
5	3716.7245	LSSEHPDATEDLQR-FTEGKGYQPCDPQVIQDR	α 12- β 3 [*]
4	2515.3231	NWINKK-YFYTGAEILGLIDEK	α 6- β 15
5	3348.7182	QDTLDASLQSFQQR-HLLEVEDLLQKHK	α 19- β 3
3,5	3128.5745	SSDEIENAFQALAEGK-VGKVIDHAITEK	EF-ABD
4	3162.5890	YNEFLLAYEAGDMLEWIEK-M [#] LAKLK	α 11- β 11
Intact Red Cell Membrane			
4	3132.5832	QEAFLNEDLGNLSGSAEALLQK-EKAATR	α 5- α 5
3,4	3470.6341	DLEELEEWISEMLPTACDESYK-KLSGLER	α 15- β 7
4	2846.5044	VQKQVFEK-YFYTGAEILGLIDEK	α 6- β 15
4	3480.6015	SHLSGYDYVGFTNSYFGN]-ANNQKVYTPHDGK	EF-ABD
3,4	2931.4134	GQQLVEAAEIDCQDLEER-KQLESSR	β 12- β 12
3,4	2899.4578	GQQLVEAAEIDCQDLEER-AKLQISR	β 12- β 12
Isolated Membrane Cytoskeleton			
4	1845.0236	DIQNLK-VQKQVFEK	α 6- α 6
4	3146.5941	YNEFLLAYEAGDMLEWIEK-MLAKLK	α 11- β 11
4	2515.3231	NWINKK-YFYTGAEILGLIDEK	α 6- β 15
4	2899.4578	GQQLVEAAEIDCQDLEER-AKLQISR	β 12- β 12
4	2931.4134	GQQLVEAAEIDCQDLEER-KQLESSR	β 12- β 12
3	1546.8600	LSGLER-WITDKTK	β 7- β 7

Asterisks indicate crosslinks that are not consistent with the current understanding of spectrin domain structure and molecular shape.

[#] indicates oxidized Methionine (+15.99492). All Cysteines are carboxyamidomethylated (+57.02146).

TABLE 4
Power of high-resolution MS/MS data when applied to large protein complexes such as the 526kDa spectrin heterodimer

Out of 27 theoretical crosslinked peptides with m/z within 5ppm of the observed m/z of 603,0810 and with low-resolution GM score of at least 0.4, only one putative crosslinked peptide has a non-zero GM score when high-resolution data is available (boldfaced and italicized).

#	m/z (theoretical)	Sequence	Mass Error (ppm)	Low-Res GM	High-Res GM
1	603.078408	QIAER-HKLM EADIAIQGDKVK	3.82	0.50	0
2	603.078410	EDMK-DLASAGNLLKKHQLLER	3.81	0.53	0
3	603.078410	KNNEK-HKLM EADIAIQGDKVK	3.81	0.45	0
4	603.078745	RWEQLLEASAVHR-PTKGKMR	3.26	0.45	0
5	603.079368	DKAAVGQEEIQLR-EAAAGRLQR	2.22	0.54	0
6	603.079373	LEQLAR-EKTQHL SAA RSSLR	2.22	0.50	0
7	603.079415	EFRSCLR-FAALEKPTTLELK	2.15	0.44	0
8	603.080373	KHGLLESAAAR-VDNVNAFIER	0.56	0.70	0.48
9	603.080375	KLLNRHR-EADDTKEWIEKK	0.55	0.52	0
10	603.080378	DLQGVQNLKKHK-STASWAER	0.55	0.45	0
11	603.080378	LQAVKLER-ANNQKVYTPHDGK	0.55	0.48	0
12	603.080378	RQEVLTR-TWKHLSDIEER	0.55	0.43	0
13	603.081048	LGDYANLK-WITDGTKVVESTK	0.56	0.49	0
14	603.081218	QEVLTR-DEEGAIVMLKRHLR	0.85	0.52	0
15	603.081220	IQEITER-KHQLLEREMLAR	0.85	0.46	0
16	603.081383	YQSFKER-WKALK AQLIDER	1.12	0.49	0
17	603.081385	RQEVLTRYSFK-PPKFQEK	1.13	0.48	0
18	603.082183	QKALSNAANLQR-RVEDQVNVR	2.45	0.53	0
19	603.082850	TQLEQSK-RVGKVIDHAITEK	3.56	0.46	0
20	603.082850	VGKVIDHAITEK-TQLEQSKR	3.56	0.52	0
21	603.083183	EAAAGRLQR-LQGQVDKHYAGLK	4.11	0.52	0
22	603.083183	GLAEVQNRLRK-HKAFEDELR	4.11	0.48	0
23	603.083183	GLAEVQNRLR-KHKAFEDELR	4.11	0.48	0
24	603.083183	LRKHGLLESAAAR-STASWAER	4.11	0.57	0

#	m/z (theoretical)	Sequence	Mass Error (ppm)	Low-Res GM	High-Res GM
25	603.083185	KLNEASRQQR-DGLAFNALIHK	4.11	0.58	0
26	603.083185	QNEVNAAWERLR-EPLATRKK	4.11	0.46	0
27	603.083190	RQEVLTTRYQSEK-KDENVNKR	4.12	0.44	0