research articles Journal of proteome research

Protein Cross-Linking Analysis Using Mass Spectrometry, Isotope-Coded Cross-Linkers, and Integrated Computational Data Processing

Jan Seebacher,[†,‡] Parag Mallick,[‡,§] Ning Zhang,[‡] James S. Eddes,[‡] Ruedi Aebersold,[‡,||] and Michael H. Gelb*[,†]

Departments of Chemistry and Biochemistry, University of Washington, Seattle, Washington 98195, Institute for Systems Biology, Seattle, Washington 98103-8909, Louis Warschaw Prostate Cancer Center, Cedars-Sinai Medical Center, Los Angeles, California, and Institute for Molecular Systems Biology, ETH, Zürich, and Faculty of Sciences, University of Zürich, Switzerland

Received April 8, 2006

Distance constraints in proteins and protein complexes provide invaluable information for calculation of 3D structures, identification of protein binding partners and localization of protein−protein contact sites. We have developed an integrative approach to identify and characterize such sites through the analysis of proteolytic products derived from proteins chemically cross-linked by isotopically coded cross-linkers using LC-MALDI tandem mass spectrometry and computer software. This method is specifically tailored toward the rapid analysis of low microgram amounts of proteins or multimeric protein complexes cross-linked with nonlabeled and deuterium-labeled bis-NHS ester cross-linking reagents (both commercially available and readily synthesized). Through labeling with [^{18}O]water solvent and LC-MALDI analysis, the method further allows the possible distinction between Type 0 and Type 1 or Type 2 modified peptides (monolinks and looplinks or cross-links), although such a distinction is more readily made from analysis of tandem mass spectrometry data. When applied to the bacterial Colicin E7 DNAse/Im7 heterodimeric protein complex, 23 cross-links were identified including six intersubunit cross-links, all between residues that are close in space when examined in the context of the X-ray structure of the heterodimer. In addition, cross-links were successfully identified in five single subunit proteins, beta-lactoglobulin, cytochrome *c*, lysozyme, myoglobin, and ribonuclease A, establishing the generality of the approach.

Keywords: protein structure • protein structure prediction • protein−protein interaction • cross-linking • mass spectrometry • proteomics • protein complex

Introduction

The locations of covalently bound cross-links of defined length in proteins or in multiprotein complexes can provide useful information about their tertiary and quaternary structure.[1−4] Protein cross-linking reagents typically contain two reactive ends, each of which is capable of forming a covalent bond with protein amino acid side chains.

Figure 1 gives four possible reaction products[5] of a cross-linking reagent that reacts with a heterodimeric protein complex. Protein cross-links of the type shown in Figure 1 (panels A and B) provide distance constraints, which provide invaluable information about protein structure and protein complex topology. In contrast, cross-links of the type shown in Figure 1 (panels C and D) indicate that the protein reactive group is solvent accessible. Possible uses of such experimentally determined distance constraints include the refinement of computed 3D protein structures by excluding predicted structures that are inconsistent with the measured distance constraints,[6] the identification of interfaces in protein complexes and the resolution of mixtures of protein complex samples generated by immunoprecipitation or affinity tagging into structures with distinct composition.

The use of mass spectrometry to identify the cross-linked peptides and to determine the location of the covalent modification is an advancing bioanalytical area.[3,4,7−10] The choice for cross-linking reagents (length and chemical specificity) and the respective experimental cross-linking workflow are as important as the availability of computational tools for mass spectrometry data analysis as recently presented in a review by Sinz.[11] In general, the distance range within which a bifunctional cross-linker will covalently connect two amino acid residues (distance constraint) within a protein is more rigidly defined the shorter the cross-linker spacer arm. The rigidity of

* To whom correspondence should be addressed. E-mail: gelb@chem.washington.edu.
† University of Washington.
‡ Institute for Systems Biology.
§ Cedars-Sinai Medical Center.
|| Institute for Molecular Systems Biology, ETH, Zürich, and University of Zürich.

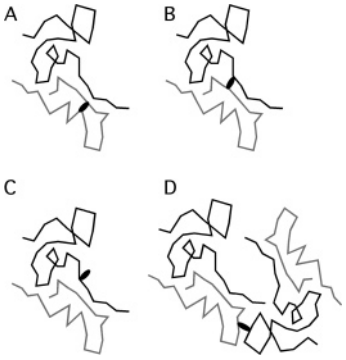

Figure 1. Possible types of covalent cross-links formed with a heterodimeric protein. Panel A shows a cross-link (black oval) of the intra-subunit type between residues of the same polypeptide chain. Panel B is an intra-heterodimer cross-link of the intersubunit type involving a residue from one subunit (light gray chain) linked to a residue in the other subunit (dark gray chain). Panel C shows a monolink in which only one end of the cross-linker is attached to the protein. Panel D shows an inter-heterodimer cross-link in which two heterodimers are linked together.

distance constraints is critical for their benefit for protein structure prediction. Unfortunately, more sophisticated cross-linking reagents with a combination of functionalities such as cleavage sites, affinity handles, isotope labels, etc. (see i.e., Petrochenko et al.,[10] Trester-Zedlitz et al.,[12] Hurst et al.,[13] Müller et al.,[14] Collins et al.,[15] and the Pierce catalog for cross-linking reagents) require elaborate syntheses, and, more importantly, result in increased cross-linking spacer lengths and therefore distance constraints that are less useful for structure prediction. Another consideration regarding the "ideal" cross-linking reagent for a protein sample is the target amino acid. Typical functional groups for which specific cross-linkers are available are sulfhydryls from cysteines, which are rare (2.3% according to the human IPI database Version 3.17[16]) and can be involved in disulfide bridges, and the more frequent ϵ-amino groups from lysines (5.6%[16]), which are often located on the surface of proteins, and are therefore expected to be accessible to cross-linking reagents. Therefore, amino-reactive, bis-NHS esters are arguably the most commonly used protein cross-linking reagents.

Experimentally, a protein or multiprotein complex is generally treated with a cross-linking reagent and then digested with one or more proteolytic enzymes to generate a mixture of protein-derived peptides. The large amount of spectral data typically obtained when analyzing biomolecules requires methods to distinguish cross-linker-modified peptides from unmodified peptides, since cross-linker modified peptides are usually infrequent in the population of peptides generated. If the cross-linking reagent contains an incorporated affinity handle, then it becomes possible to use affinity chromatography to enrich for cross-linked species.[12,13] As an alternative, cross-linked peptides can be selectively marked by a mass spectrometry-observable, signature feature that is unique to such peptides, i.e., heavy isotopic labeling.[10,14,15] Ideally the isotopic signature should be diagnostic for the type of cross-links shown in Figure 1.

Singly modified protein residues are the major products of the protein/cross-linker reaction; most surface reactive amino acid side chains can react with cross-linker reagent, but a second protein reactive group will only rarely lie in close enough proximity to the attached reagent to form a second cross-linker/protein linkage. Even when two protein reactive groups are positioned so that cross-linking is sterically possible, competition from hydrolysis of the cross-linking reagent's reactive ester contributes to monolink formation. Consequently, any effective method must detect rare cross-links in the presence of a vast excess of monolinks and nonmodified peptides.

We have developed a new integrative method for identifying protein cross-links, monolinks and nonmodified peptides that relies on the collection of MALDI mass spectrometry data of protein samples that have been modified with isotopically substituted cross-linking reagents in the presence of isotopically substituted solvent water. We have also developed powerful computational tools that analyze the experimental data and rapidly lead to the detection and identification of cross-linked, monolinked, and nonmodified peptides. In brief, the isotopic substitution leads to mass spectrometry signature features of cross-linked and monolinked peptides that can be automatically identified by computer analysis.

Experimental Section

Synthesis of Deuterated Cross-Linking Reagents. Disuccinimidyl suberate containing 12 deuteriums in the diacyl unit, [d_{12}]DSS, (Figure 2, panel B) and disuccinimidyl glutarate containing 6 deuteriums in the diacyl unit, [d_6]DSG were prepared as follows.

A mixture of *N*-hydroxysuccinimide (50.0 mg, 434 μmol, Fluka) and [d_{12}]suberic acid (53.9 mg, 290 μmol, C/D/N Isotopes, Pointe-Claire, Quebec, Canada) in 1 mL of dry tetrahydrofuran was added to *N*-cyclohexylcarbodiimide, *N*′-methyl polystyrene resin (724 mg, 1.16 mmol, 200−400 mesh, 2% DVB, 1.3−1.7 mmol/g, Novabiochem), which had been swollen in 10 mL of dry tetrahydrofuran for 5 min. The reaction mixture was gently stirred overnight at room temperature under nitrogen. The mixture was filtered under nitrogen through a rubber septum-capped, glass frit filter, and the resin was washed with 10 mL of dry tetrahydrofuran. The filtrates were combined, and solvent was removed in vacuo to yield a pale yellow solid (60 mg, 54% yield). In the same way, [d_6]DSG was made from d_6-glutaric acid (40 mg, C/D/N Isotopes) (75% yield). Both cross-linkers were characterized by ^{1}H NMR in d_6-DMSO (2.509 ppm): δ 2.81 (s, 8H, CH_2, bis-NHS ester) and shown to be free of mono active ester or diacid starting material. [d_0]DSS and [d_0]DSG were obtained from Pierce. It may be noted that just after the completion of this work, the d_4-labeled bis-sulfo-NHS esters of glutaric and suberic acid have become commercially available (Pierce Chemicals, Inc.).

Cross-Linking Reactions. Protein concentrations in stock solutions were measured using the Bradford dye binding assay (BioRad) using bovine serum albumin as a standard. One cross-linking experiment was carried out in four separate reaction vials, I−IV, each with 50 μg of Colicin E7 DNAse/Im7 heterodimer. (obtained as a generous gift from D. Baker, University of Washington[17]): Reaction I, [^{16}O]H_2O + [d_0]DSS; II, [^{16}O]H_2O + [d_{12}]DSS; III, [^{16}O]H_2O/[^{18}O] H_2O (1/1) + [d_0]DSS; IV, [^{16}O]-H_2O/[^{18}O] H_2O (1/1) + [d_{12}]DSS. Reactions were prepared by mixing 50 μL of [^{16}O]H_2O or [^{18}O]H_2O (isotope enrichment: >95 at. %, Isotopes Inc. in Pelham, NH), 50 μL of 2X phosphate buffered saline (568 mg Na_2HPO_4, 1.754 g NaCl per 100 mL, adjusted to pH 8.5 with 0.1 M HCl) and 10 μL of heterodimer (5 mg/mL in phosphate buffered saline, pH 8.5). Cross-linker (2.3 μL of 14.7 mM in dry dimethylformamide) was added in

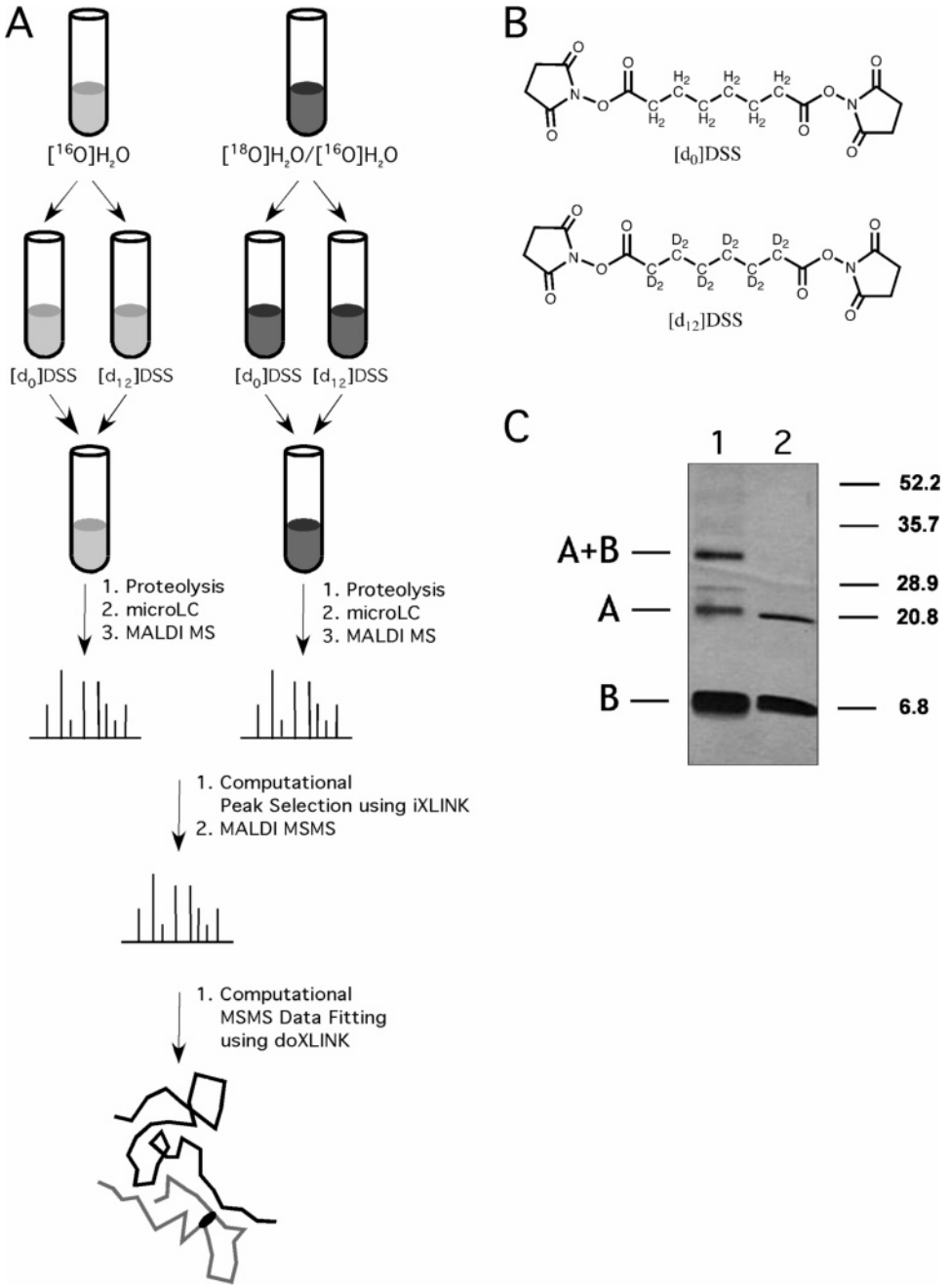

Figure 2. Schematic of the cross-linking experimental design and analysis. (panel A) A protein sample in $[^{16}O]H_2O$-buffer is split into 2 identical samples, and one is treated with light cross-linker (d_0[DSS], panel B) and the other with heavy cross-linker (d_{12}[DSS], panel B). The two samples are combined, proteolyzed and submitted to liquid chromatography; fractions are spotted onto a MALDI plate for mass spectrometry analysis. This process is repeated but starting with a sample of protein in $[^{16}O]/[^{18}O]$water-buffer to generate a second set of MALDI data. The mass spectrometry data sets are analyzed by the program iXLINK to identify MALDI spots containing monolinked and cross-linked peptides for subsequent tandem mass spectrometry analysis. Data from the latter are analyzed with doXLINK software and confirmed with XLinkViewer to infer the sequences of monolinked and cross-linked peptides. Panel C shows an SDS-PAGE gel of DNase/Im7 after (lane 1) and before (lane 2) reaction with DSS.

one portion, the solution was briefly mixed on a vortex mixer and then continuously mixed at room temperature by clamping the tube above the vortex mixer. After 0, 0.5, 1, and 4 h, an aliquot of the reaction mixtures (9 μL) was analyzed by SDS-PAGE on a 15% gel followed by silver staining (see Figure 2, panel C). Reactions were allowed to proceed overnight to ensure that all NHS ester had reacted with either amino groups of the protein complex or water. Thereafter, respective light and heavy cross-linker samples were combined (I + II, III + IV) and subjected to a protease digestion protocol.

The two samples, each containing 100 μg of protein, were dried in a Speed Vac (Savant Instruments) and the residue dissolved in 8 M urea, 50 mM ammonium bicarbonate and incubated for 30 min at room temperature. For some samples (see Results), cysteines were reduced and alkylated by adding 5 μL of 45 mM DTT and incubation for 30 min at 56 °C. This was followed by addition of 5 μL of 100 mM iodoacetamide and incubation for a further 30 min at room temperature in the dark. The solutions were diluted 4-fold to give 2 M urea by addition of 50 mM ammonium bicarbonate (pH 8) and treated

with agarose-immobilized, TPCK-treated trypsin (Pierce Chemicals Inc. Cat. No. 20230, ≥200 TAME units/mL) as follows: trypsin slurry (400 μL) as supplied by the manufacturer was centrifuged, and the gel was resuspended in 400 μL of 50 mM ammonium bicarbonate. This washing step was repeated three times. The washed pellet was resuspended in 100 μL of 50 mM ammonium bicarbonate, and 20 μL of slurry was added to each cross-linked protein sample. The samples were shaken on an orbital mixing platform at 37 °C overnight.

After centrifugation, the supernatant was transferred to a new tube, the yield in tryptic peptides estimated to be 100 μg (100%), and endoproteinase Asp-N (0.27 μg, CalBiochem Inc.) was added. Samples were incubated overnight at 37 °C, then concentrated to dryness in a Speed-Vac, and finally residues were dissolved in 60 μL water/0.2% trifluoroacetic acid in preparation for LC/MS.

Micro-Liquid Chromatography/MALDI Mass Spectrometry Analysis. Six μL aliquots of each of the [^{16}0]water samples (see above) (1/10 of the original amount of protein, ca. 10 μg) were injected into an Eksigent Express-100 nanoliquid chromatography system (Eksigent, Livermore, CA) equipped with an Endurance Autosampler (Spark Holland BV, NL, obtained from Eksigent). Binary solvent mixtures of solvent A (0.1% trifluoroacetic acid in water) and solvent B (0.1% trifluoroacetic acid in acetonitrile) were used. Samples were loaded onto a 3.5 μm Zorbax3000SB−C18 column (Agilent, 0.3 × 50 mm) using an initial washing step of 5% B for 8 min, followed by a linear gradient from 5% B to 50% B over 22 min, followed by a linear gradient from 50% B to 90% B over 3 min, all at a flow rate of 4 μL/min. The micro-LC eluant was mixed in a 10:4 ratio with a solution of α-cyano-4-hydroxycinnamic acid matrix solution (6.64 mM in 36% methanol, 56% acetonitrile, 8% deionized water, Agilent Technologies, Palo Alto, CA, containing the two peptides ACTH (Proteomass ACTH, Fragment 18−39, MALDI-MS Standard from Sigma) and angiotensin II (Mass Spec Standard from Sigma), both at 100 fM concentration for internal mass calibration) in a mixing "T" before spotting onto a 192-well MALDI plate at 8 s intervals with a Probot Micro fraction collector (LC Packings).

Alternatively, a 4 μL aliquot (1/15 of the original amount of protein, ca. 6.7 μg) of each of the two protease digests (see above) was injected using an Ultimate HPLC system coupled to a Famos Micro Autosampler (LC Packings/Dionex Inc.) onto a reverse-phase column (house-packed C18 column, 150 μm I. D. × 12.5 cm, 100 Å Magic C18AQ, Michrom BioResourses Inc.). The column was developed using a gradient of 5% B to 70% B over 70 min at a flow rate of 1.5 μL/min. The micro-LC eluant was mixed with an equal volume of α-cyano-4-hydroxycinnamic acid matrix solution (6.64 mM in 36% methanol, 56% acetonitrile, 8% deionized water, Agilent Technologies, Palo Alto, CA) in a mixing "T" before spotting onto a 192-well MALDI plate at 22 s intervals with an Probot Micro fraction collector (LC Packings).

The samples were analyzed using a MALDI-TOF/TOF tandem mass spectrometer (ABI 4700 Proteomics Analyzer, Applied Biosystems, Inc.). Both mass spectrometry and tandem mass spectrometry data were acquired with a Nd:YAG laser with a 200 Hz sampling rate. For mass spectra, 1000 laser shots per spot were used to ensure appropriate ion statistics for quantification. Tandem mass spectrometry mode was operated with 1 keV collision energy. The collision-induced dissociation was performed using air as the collision gas. Typically 2000 laser shots were used for tandem mass spectrometry data acquisition. Both mass and tandem mass spectrometry data were acquired using the instrument default calibration, or internal calibration based on the masses of the two standard peptides ACTH fragment 18−39, and angiotensin II (see above).

Data Analysis. The programs iXLINK, doXLINK, and XLink-Viewer along with a detailed instruction manual for analyzing the MALDI mass and tandem mass spectrometry data are available in CD format from the authors upon request. These packages are also available under an open source software license from

http://www.systemsbiology.org/Resources_and_Development/Downloadable_Software.

Results

Description of the Cross-Linking Analysis. We first outline our new method in general terms and then show the experimental results from applying the protocol to a heterodimeric protein sample.

As shown in panel A of Figure 2, a sample of protein in [^{16}O]-H_2O buffer is split into two tubes. One tube is treated with the isotopically light cross-linker (disuccinimidyl suberate, [d_0]DSS or disuccinimidyl glutarate, [d_0]DSG) and the other tube is treated in an identical way with the deuterated reagent ([d_{12}]-DSS or [d_6]DSG). After incubation, the two samples are combined, and the sample protein is digested with one or more proteases. Note it is also possible to combine the light and heavy cross-linker reagent prior to addition to a single tube of protein sample; however, this should not be done if peptides containing more than one attached cross-linker are to be analyzed (which was not done in this study). Consequently, the presence of both light and heavy cross-linking reagent will cause any cross-linker-modified peptide to appear as a doublet of equally intense peaks in the MALDI mass spectrum due to the presence of both light and heavy cross-linking reagent. An identical sample of protein is also prepared in [^{16}O]H_2O/[^{18}O]-H_2O (50/50) buffer, split into two samples, and treated with light and heavy cross-linker and combined, as above. This step distinguishes cross-links from monolinks in the mass spectrum since only the latter will contain solvent-derived oxygen, which is incorporated by the hydrolysis of the cross-linking reagent's reactive ester at the end not attached to the protein. Thus, any mass spectrometry peak arising from a monolink modified peptide will show an additional 2 Da splitting, whereas a cross-link species will not show this additional splitting. Using [^{18}O]-water in this way was first mentioned by Collins et al.,[15] but the method was not put into practice until now.

The proteolytic digest from the [^{16}O]water sample and [^{16}O]/[^{18}O]water sample are independently fractionated by reverse phase chromatography and spotted onto two standard MALDI sample plates. The MALDI mass spectrum is obtained from each spot. Chromatographic fractionation is required because the number of MALDI peaks obtained from the unfractionated proteolytic digest is sufficiently large that significant overlap of peaks makes visualization of the isotopic signatures problematic, especially since cross-link species are rare compared to the much more abundant monolink and nonmodified species.

Given that the splitting features of the mass spectrum due to isotopic substitution are enforced by chemistry, cross-linked, monolinked and nonmodified peptides can be distinguished by computational analysis of mass spectrometry data files. To do this, we developed the iXLINK program, and the process is illustrated in Figure 3.

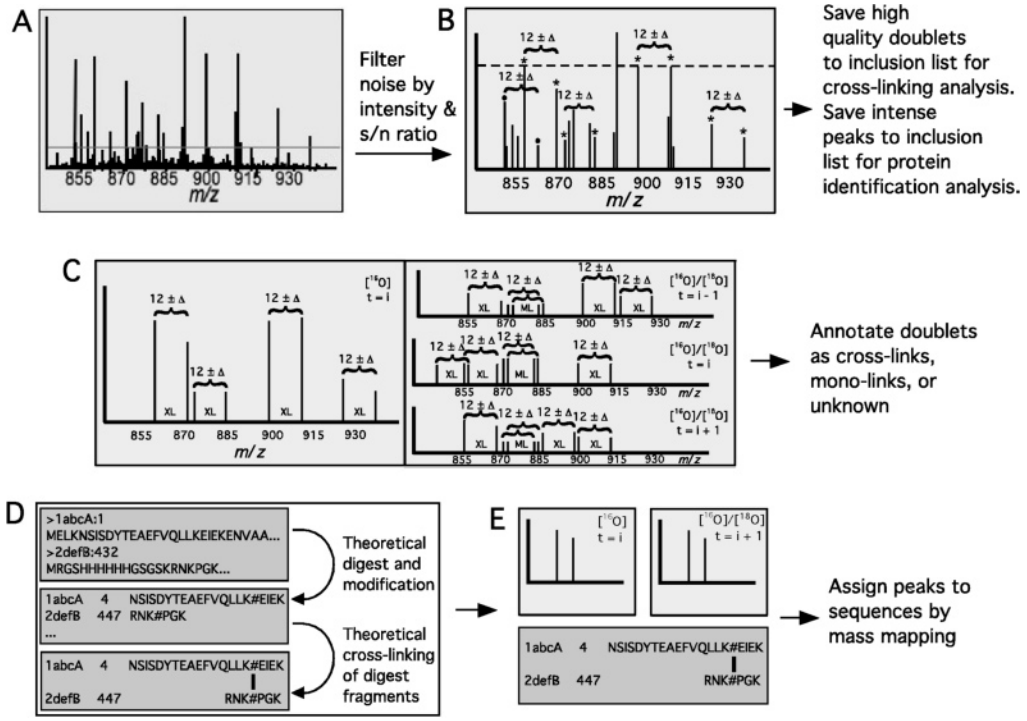

Figure 3. Schematic illustration of extraction of MALDI mass spectrometry data by the program iXLINK. Those peaks above the threshold line in panel A are analyzed for mass doublets, which are caused by modification of the peptide with the light and heavy forms of the cross-linker (peaks marked by asterisks in panel B are those with an intensity ratio between paired peaks within the specified value, 0.3−3.0 in this case; those marked by dots are outside this range). Additional high-intensity peaks (above the dashed line in panel B) are saved to an inclusion list for protein identification if necessary. The left part of panel C shows a portion of the MALDI mass spectrum from a fraction eluting at $t = i$ from the column in an experiment using [^{16}O]water, and the right part is from the reaction using [^{16}O]/[^{18}O]water eluting within a window of retention time close to that for the left figure (in this case in 3 adjacent fractions at $t = i - 1$, i and $i + 1$). Likely cross-linked and monolinked peptide peaks are marked by XL and ML, respectively, based on whether additional splitting is observed due to incorporation of isotopic solvent. Panel D shows the steps performed by iXLINK starting from protein sequences for the two subunits of DNase/Im7 leading to the monolinks database and finally to the cross-link database. The cross-linker-modified lysine is denoted K#, and the link between cross-linked lysines is shown by a vertical bar. Panel E shows that the mass of a cross-linked peptide pair is matched to the observed MALDI mass spectrum (note that the observed mass may be the same as those of multiple computer generated masses, within the user specified mass tolerance window).

Each of the mass spectrometry spectra files from the [^{16}O]-water and [^{16}O]/[^{18}O]water runs are first screened for high-intensity peaks and for mass doublets that are separated by 12 ± Δ Da (12 is the mass difference between the light and heavy DSS reagents (Figure 2, panel B) and Δ is the tolerance in this mass difference; this parameter was set to 0.1 in our analysis. Each of the peaks in the doublet is required to have intensity and signal-to-noise values above specified thresholds (shown as a horizontal line in panel A of Figure 3). In addition, the two mass peaks that comprise a doublet should have similar intensities since equal amounts of heavy and light cross-linker were used. Since the presence of deuterium in the heavy cross-linker can cause a slight shift in the chromatographic retention time compared to the protium cross-linker, the doublet peak intensity ratio may be allowed to lie in a fairly liberal range (i.e., 1/3−3). As shown in panel B of Figure 3, four doublets (marked with asterisks) satisfy these criteria, and one doublet (marked by dots) does not. Doublets that meet these criteria are considered to be monolinked or cross-linked peptides and are saved to an inclusion list. MALDI peaks for which there is no corresponding isotopically shifted peak are considered to arise from nonmodified peptides.

As shown in panel C, Figure 3, cross-links and monolinks can be distinguished by the effect of the solvent isotope. Doublets that do not show the additional splitting of 2 Da (due to the presence of a mixture of [^{16}O] and [^{18}O] in the terminal carboxyl of the cross-linking reagent) are annotated as cross-links (designated XL in panel C of Figure 3), whereas those that show the 2 Da splitting are annotated as monolinks (designated ML). Since two separate LC/MALDI runs are used, iXLINK analyzes adjacent mass spectrometry files over a user specified retention time window. To avoid redundancy that may arise when a doublet appears in multiple elution fractions, we identify the most intense doublet within our retention time window from each experiment series and discard the rest.

The iXLINK program also creates a database of theoretical monolinks and cross-links for the protein(s) under investigation. The program first generates a list of monolinks by performing an in-silico protease digest of all proteins present in the cross-linking reaction mixture (panel D, Figure 3). iXLINK allows the user to specify the peptide cleavage sites and the maximum number of protease miss-cleavages. The program also allows for other covalent modifications, including alkylation of cysteine and [multiple] oxidation of methionine residues (these modifications frequently occur artifactually and are created as additional database entries). Unnatural amino acids can be defined and included in the protein sequence as well as any chemical modifications. Next, the amino groups (i.e., the lysine side chain and the protein N-terminus) of the protease fragments are in-silico modified by the cross-linking

reagents. Once the database of monolinks has been created, a database of cross-links and looplinks is created. Looplinks are defined as a single proteolytically generated peptide that contains a cross-linker amide linked at both ends. Cross-links are computed for all combinations of peptides in the theoretical monolink list minus the mass of one hydrolyzed cross-linker molecule. iXLINK then compares the observed mass of each of the classified monolinks and cross-links to the theoretical mass list. Because the database of theoretical molecular species is large (25 975 entries for the heterodimeric protein described below when digested with trypsin followed by Asp-N and allowing up to 6 missed cleavages per peptide, 1 allowed cross-linker modification per peptide or cross-linked peptide pair, and singly and doubly oxidized methionines), the mass of observable monolinks and cross-links often matches to more than one entry in the theoretical database within the expected mass window—according to the observed mass accuracy of the instrument. It is thus necessary to use tandem mass spectrometry to make conclusive peptide assignments. Even in cases where a unique match is found (typically for small peptides), tandem mass spectrometry data is useful by providing further confirmation of the peptide assignments. iXLINK creates a precursor mass inclusion list of all species, along with their MALDI plate spot location, to be submitted to tandem mass spectrometry. iXLINK also creates a list of all high abundant peptides that do not contain a covalently attached cross-linker. This list may be used along with standard protein identification software (for example Mascot, Matrix Science, London, UK, or SEQUEST, Thermo Electron Corporation, San Jose, CA), in cases where the identities of the proteins are unknown. Typically the MALDI sample plate from the [^{16}O]water run is used for cross-linking analysis, and the plate from the [^{16}O]/[^{18}O]water run can be saved for subsequent tandem mass spectrometry for protein identification analysis. Those precursor ions that are annotated by iXLINK as candidate cross-links are given highest priority for precursor ion selection for tandem mass spectrometry since the identification of cross-linked peptides provides more valuable structural information than does the identification of nonmodified peptides or peptides modified by monolinks. In cases where the identity of the cross-linked proteins are not known, priority can be first given to tandem mass spectrometry analysis of nonmodified peptides so that the protein(s) in the sample can be identified, a necessary prerequisite for iXLINK to create a database of theoretical monolinks and cross-links.

The method described above is designed to find peptides that contain at most one covalently attached cross-linker-derived species. More complex species may form, i.e., two peptides joined by a cross-link and also bearing one or more monolinks on either or both of the two peptides, but these are not detected with our computational search method. A complete and systematic nomenclature for cross-linker reagent-modified peptides has been published.[5] In this present study, we prefer the designations "monolink" (Type 0[5]), "looplink" (Type 1[5]), and "cross-link" (Type 2[5]) because these are the species that our method identifies.

The program doXLINK is applied to assign the fragment ion spectra to the cross-linked peptide sequences. doXLINK generates a list of mass fragments for each iXLINK candidate and attempts to match the data to the experimental fragments. Tandem mass spectra are first filtered to retain peaks above a user specified signal-to-noise threshold. doXLINK compares the tandem mass spectra from the precursor peptide ion modified with the light cross-linker to that from the heavy cross-linker modified peptide and annotates mass fragment peaks that differ by the cross-linker isotopic mass difference as cross-linker-containing fragment ions. For each comparison of experimental tandem mass spectrometry data to that predicted from the candidate molecular species, doXLINK assigns a matching score that is based on the ProbID score function.[18] In addition, the scoring includes a component based on whether the observed isotopically shifted mass fragments (i.e., those containing the cross-linker) are predicted to be mass shifted. It may be noted that software is available that can accomplish some of the tasks carried out by iXLINK and doXLINK (i.e., ASAP and MS2Assign by Young et al. http://roswell.ca.sandia.gov/~mmyoung/,[5] http://prospector.ucsf.edu/ucsfhtml4.0/msbridge.htm,[19] SearchXLinks by Wefing et al. http://www.searchxlinks.de/cgi-bin/home.pl,[20] CLPM by Tang et al.,[21] or VIRTUALMSLAB by de Koning),[22] but none of these packages provides the integrative task achieved by the combination of iXLINK and doXLINK that is required for the cross-linking analysis of isotope-labeled peptides using MALDI−MS and −MS/MS carried out on multisubunit proteins.

Application of the Protein Complex Cross-Linking Platform. We have chosen to test our platform by analyzing the heterodimeric complex of Colicin E7 DNAse (MW 16.2 kDa) with the Im7 immunity protein (MW 9.9 kDa); this complex was chosen because it can be readily prepared by bacterial expression, and an X-ray structure is available.[17] The latter enables us to evaluate the validity of our identified cross-links by comparison of interresidue distances from the 3D structure of the protein. To test the general applicability of our method, we also applied our experimental and analytical workflow to the five proteins beta-lactoglobulin, cytochrome C, lysozyme, myoglobin, and ribonuclease A; these proteins are all of which are commercially available and have characteized 3D structures (X-ray and/or NMR) in the literature, to the same experimental and analytical workflow.

Gel electrophoretic analysis of the cross-linking reaction of DNase/Im7 treated with DSS for 30 min shows that approximately 30% of the heterodimer contained intersubunit cross-links (Figure 2, panel C). There were no significant changes in the gel band intensity ratios when reaction aliquots incubated for more than 30 min were analyzed (not shown). A similar gel profile was observed when the shorter cross-linker DSG was used (not shown). After cross-linking, the protein was denatured in 8 M urea. Reduction and alkylation of cysteines were used for the five monomeric proteins, but not for the DNase/Im7 heterodimer which lacks cysteines. After sample dilution to reduce the denaturant to 2 M, the protein was digested with trypsin followed by Asp-N. The use of two proteases was explored since the typical size of cross-linked tryptic peptides is on average ~4-fold larger than that of peptides generated from non-cross-linked proteins. It is well-known that the identification of larger peptides by MALDI tandem mass spectrometry is problematic in general.

For the task of differentiating between monolinks and cross-links, we have three methods. The first is based on the use of [^{16}O]/[^{18}O] isotope labeling (see Figure 4), the second is based

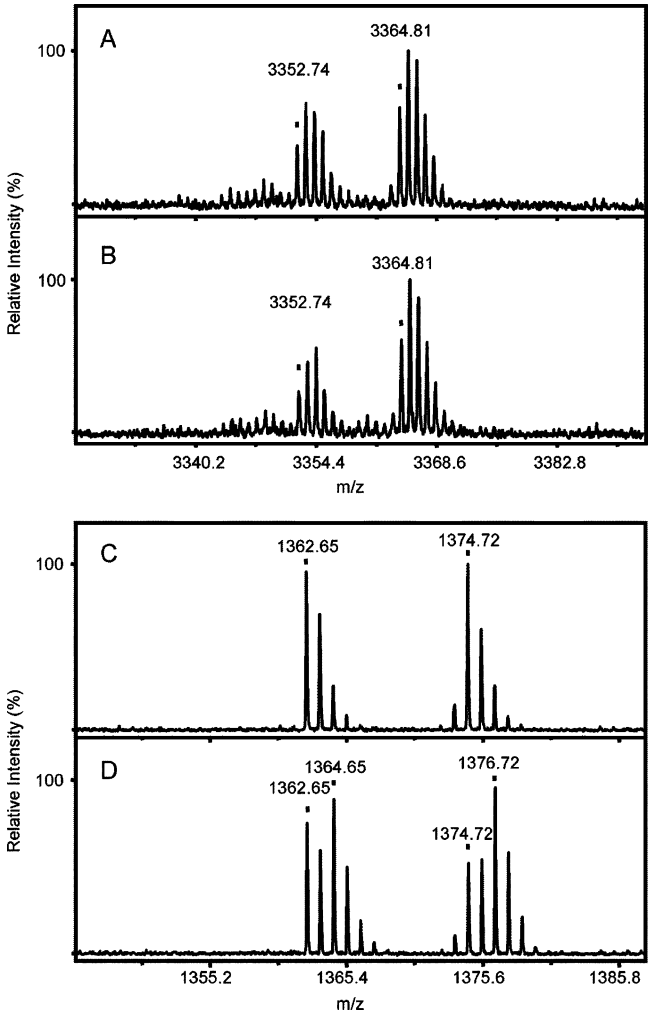

Figure 4. Typical MALDI mass spectra of peptides containing covalently attached DSS cross-linker. See text for a discussion of panels A−D. The natural abundance isotope pattern for the peptide containing the heavy cross-linker is similar to the pattern seen for the corresponding peptide attached to the light cross-linker except for the presence of a small peak at $(M+H^+) - 1$ due to the presence of residual protium in the heavy reagent.

in the use of reporter ions from fragment ion spectra (see Figures 5 and 6), and the third is based on MS/MS analysis.

Figure 4 shows typical MALDI mass spectra obtained by cross-linking analysis. Panel A shows a pair of peptide peaks that differ in mass by 12 Da; thus this pair likely represents the same peptide containing the light or heavy DSS cross-linking reagent. The fact that the same mass spectrum pattern is seen for both the [^{16}O]water and [^{16}O]/[^{18}O]water experiments (compare panels A and B) suggests that these signals arise from two peptides covalently cross-linked together. In contrast, panels C and D of Figure 4 show a doublet with 12 Da splitting that shows additional 2 Da splitting when the cross-linking reaction is carried out in a mixture of light and heavy solvent. These signals likely arise from a single peptide containing a monolink. As noted above, the iXLINK program processes the MALDI mass spectrometry datasets to find these peptides automatically and to classify them into their appropriate groups.

Figure 5 shows typical tandem mass spectra for monolink, looplink, and cross-link species.

Previous studies[5,23] of peptides modified with DSS showed that monolinks give rise to a signature 240 Da fragment, which is due to the fragmentation reaction sequence shown in Figure 6.

In the 12 cross-linking experiments carried out with Colicin E7 DNAse/Im7, beta-lactoglobulin, cytochrome *c*, lysozyme, myoglobin, and ribonuclease A using the two cross-linking reagents DSS and DSG respectively (the heterodimer sample involved one LC−MS rerun, each), there were a total of 645 tandem mass spectra assigned to cross-linker-modified peptide species (monolinks, looplinks, and cross-links) analyzed with doXLINK and XLinkViewer, including redundant precursor ions, i.e., species that were found in multiple MALDI plate spots. In the current study we also found that cross-linked peptide species undergo an analogous fragmentation leading to a signature ion at 222 Da for the peptide modified with the light cross-linker (Figure 6) and at 234 Da for the corresponding heavy cross-linker-modified species (in the case of DSS, 180/186 Da for DSG cross-links). The monolink signature fragment ion pair (240/252 Da in the case of DSS, 198/204 Da in the case of DSG) was found in 53% of the spectra assigned to monolinks. In only 3.3% of all spectra assigned to monolinks, a cross-link reporter ion was observed. This emphasizes that this signature ion provides some value for diagnosing monolink assignments. The cross-link reporter ion is notably less valuable as it appeared in only 6% of the tandem mass spectra assigned to cross-links or looplinks, but the monolink reporter fragment was observed in 22% of these spectra. Given the lack of specificity of these ions, doXLINK screens all tandem MS spectra for monolink and cross-link signature fragments, and only uses the monolink reporter ion information as a component of the scoring algorithm. doXLINK reports all observed reporter ions in its results to facilitate/support manual peptide assignment with the XLinkViewer program.

In general, we find that tandem mass spectra of nonmodified, monolink- and looplink-modified peptides contain more fragment ions compared to those obtained from cross-linked peptides (compare Figure 5 panels A and B to panel C). This has been noted previously, for example see Petrochenko et al.[10] This may be because the charge on cross-linked species is distributed over a larger number of proton acceptors. For example, for a (+1) ion derived from a cross-linked species, each of the two amino termini and each of the two C-termini (lysine or arginine) will have on average a charge of ∼+0.25, whereas the termini of a nonmodified, monolink or looplink peptide will have a charge of ∼+0.5. It is generally thought that protonation of the peptide termini facilitates fragmentation of peptide bonds.[24] Petrochenko and co-workers developed an elegant strategy of using cleavable cross-linkers so that cross-linked peptides could be chemically converted to two monolinked peptides prior to tandem mass spectrometry.[10] We chose to use chemically simpler, noncleavable cross-linking reagents, which are relatively short in length and thus provide more rigid distance constraint information. As we show in this study, the limited amount of MALDI fragmentation data derived from cross-linked peptides is sufficient to identify the cross-linked peptides when the sequences of the cross-linked proteins are known a priori. These points are discussed further in the Discussion section. When the identity and thus the sequence of the cross-linked proteins is not known, mass and tandem mass spectrometry data from the much more plentiful nonmodified peptides rather than from cross-linked or monolinked

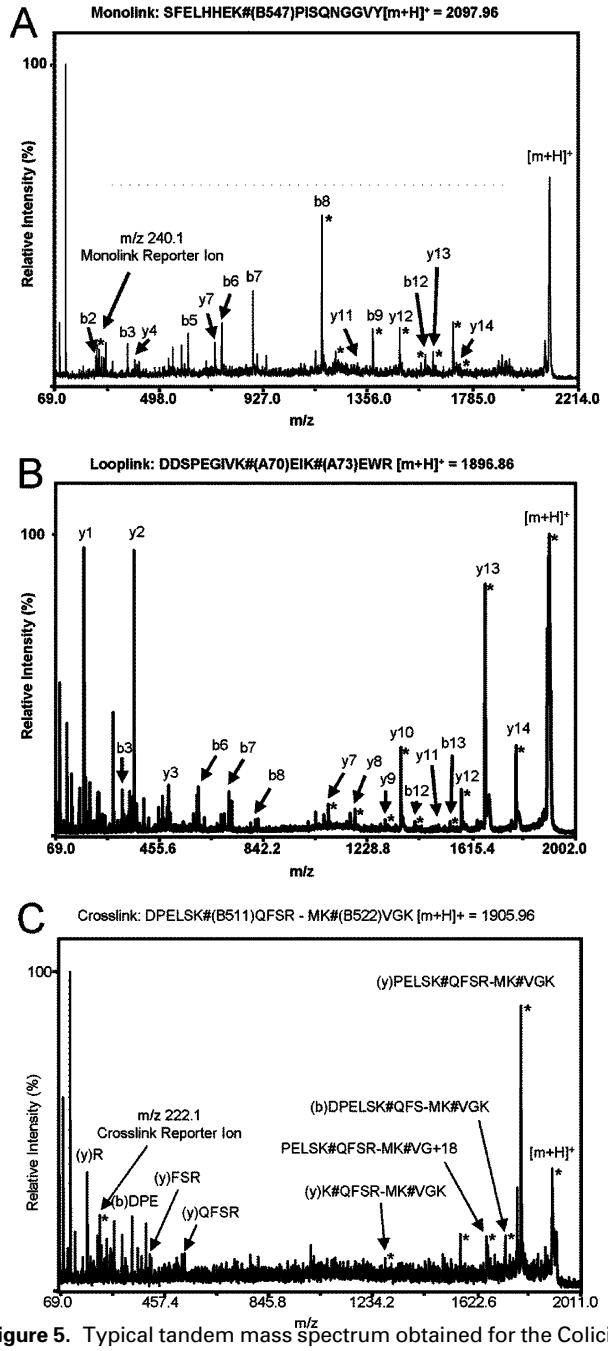

Figure 5. Typical tandem mass spectrum obtained for the Colicin E7 DNAse/Im7 heterodimer. Panel A shows the tandem mass spectrum of a monolink peptide SFELHHEK#PISQNGGVY with an internal lysine residue (K#) modified with [d_0]DSS. It shows an extensive series of y- and b-ion fragments. Peaks marked with an asterisk were found to be shifted by 12 Da in the spectrum of the same monolink peptide but containing the heavy cross-linker (parent ion with $[M+H]^+$ 2109.98 in the same fraction). In addition, a peak at m/z 240.2 (252.3 in the [d_{12}]DSS-derived spectrum) was observed and is assigned to the diagnostic reporter ions for monolinks (see text). Panel B shows the tandem mass spectrum of the looplink peptide DDSPEGIVK#EIK#EWR with two internal lysines (K#) bridged by [d_0]DSG. Panel C shows the tandem mass spectrum of the [d_0]DSS-derived cross-link containing peptide DPELSK#QFSR-MK#VGK. Five peaks could be assigned to cross-linker-containing fragment-ions (also seen in the corresponding spectrum of the [d_{12}]DSS-derived cross-link species. In addition, a peak at m/z 222.1 (234.2 in the [d_{12}]DSS-derived spectrum) was observed, which is assigned to the diagnostic reporter ion for cross-links (see text).

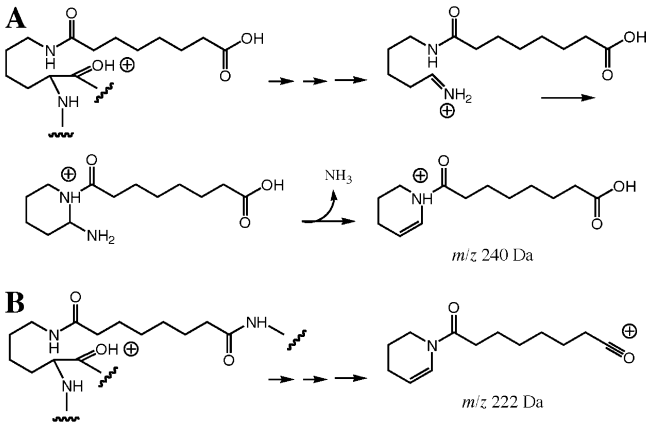

Figure 6. Proposed reaction pathway for the generation of the m/z 240 Da signature ion for monolinks (panel A) and the m/z 222 Da signature ion for cross-links (panel B).

peptides would be used for protein identification prior to the processing of the mass spectrometry data by iXLINK.

Table 1 summarizes the number of monolink and cross-link species, along with the indicated statistics, identified by iXLINK, doXLINK, and XLinkViewer analysis of LC−MS data derived from cross-linking of β-lactoglobulin, cytochrome *c*, lysozyme, myoglobin, and ribonuclease A and the Colicin E7 DNAse/Im7 heterodimer with the cross-linking reagents DSG or DSS, followed by digestion with trypsin and endoproteinase Asp-N. Table 2 gives a list of all monolinks identifed in the Colicin E7 DNAse/Im7 heterodimer found using DSG or DSS cross-linking reagents. With DSG, 51% of all possible monolinks were identified (assuming every protein amino group could become modified). Using DSS, the corresponding number is 48%. Together, 66% of all protein amino groups were identified as being part of a monolinked species. All of the results described in Tables 1 and 2 and Figure 7 below (and analogous figures given as Supporting Information) were carried out without the use of [^{18}O]water labeling. The results using [^{18}O]water labeling to distinguish monolinks from cross-links are discussed at the end of the Results section.

Figure 7 shows the location of the cross-linked lysines and/or N-termini with respect to the X-ray structure of the Colicin E7 DNAse/Im7 heterodimer. Analogous diagrams for the other five proteins are given in the Supporting Information.

Not shown in Figure 7 are cross-links identified that involve the N-terminal 16 amino acid segment of the B subunit of the Colicin E7 DNAse/Im7 heterodimer including the 6His affinity tag since this segment was not observed in the X-ray structure. These cross-links are: A73/B−N-terminus; B−N-terminus/B449; B446/B497; B−N-terminus/B537.

As seen in Figure 7, the inter-β-C atom distances of all identified cross-linked residues are found in the X-ray structure to be ≤17 Å for DSG and ≤22 Å for DSS. The calculated distance between α-C of cross-linked residues assuming a fully extended conformation of the cross-linker and protein amino acid residues is 20 and 24 Å for DSG and DSS, respectively.[25,26]

The DSS experiment yielded the largest number of identified cross-links, 20 molecular species, 15 of which involve unique residue pairs (Table 1). The DSG experiment yielded 17 molecular cross-linked species, 11 of which involve unique residue pairs (Table 1). The total number of molecular species when results from both experiments are combined is 33, 23 of which are structurally unique (Table 1). This illustrates the

Table 1. Number of Monolinks and Cross-Links Identified for the Cross-Linking Analysis of the Proteins Listed below with the Cross-Linking Reagents DSG and DSS

	mass pairs[a]		identified monolinks[b]			identified cross-links[c]			overlap DSG/DSS residues[d]		cross-linked distance[e]	
protein − cross-linker	total	identified	total	peptides	residues	total	peptides	residues	(monolinks)	(cross-links)	range [Å]	avg [Å]
lysozyme-DSG	55	7	1	1	1	2	1	1	1(3)	1(3)	19.0	19.0
									33.3%	33.3%		
lysozyme-DSS	294	22	3	3	3	2	2	2			19.0−20.8	19.9
ribonuclease A-DSG	79	14	8	6	6	6	5	4	2(5)	2(5)	5.7−22.1	12.6
									20%	20%		
ribonuclease A-DSS	32	5	2	1	1	3	3	2			13.5−24.1	20.6
β-Lactoglobulin-DSG	62	17	7	6	6	10	6	4	3(9)	2(6)	5.4−14.3	7.9
									33.3%	33.3%		
β-Lactoglobulin-DSS	93	29	15	10	6	14	7	4			5.4−15.9	7.4
cytochrome C-DSG	119	26	1	1	1	25	8	8	1(2)	5(12)	4.2−30.7	17.7
									50%	41.7%		
cytochrome C-DSS	72	16	3	2	2	13	9	9			4.2−27.9	13.6
myoglobin-DSG	284	57	33	15	12	24	7	6	4(14)	4(10)	5.3−28.4	11.1
									28.5%	40.0%		
myoglobin-DSS	325	42	27	17	10	15	11	8			7.2−22.8	12.0
colicin E7 DNAse/Im7-DSG[F]	299	132	68	23	16	64	17	11	11(19)	3(23)	5.3−17.6	12.2
									57.9%	13%		
colicin E7 DNAse/Im7-DSS[f]	303	76	39	24	14	24	20	15			11.3 − 22.1	16.2

[a] Mass pairs designates all light cross-linker-modified and heavy cross-linker-modified precursor ion pairs that match to at least 1 species in the iXLINK-generated peptide database. [b] Peptides designates the number of unique peptide species that were identified, and residues designates the number of unique lysine or N-termini residues that were identified. For example, if two peptides were identified that contain the same amino acid sequence but one with a Met and the other with an oxidized Met, these are counted as two peptides with 1 residue modified. The same applies to protease mis-cleavages. *Total* designates the total number of identified monolinks including redundancies. For example, identification of the same molecular species more than once because it appears in multiple MALDI plate spots is counted multiple times. [c] Total, peptides and residues is defined as for monolinks except that each that identified cross-link species have two modified residues and these are counted as one. [d] See footnotes *C* and *D* for definition of *residues*. The number of residues that were found in both the analyses with DSG and DSS is given, and the number in parenthesis is the total number of residues identified. [e] Distance between cross-linked β-C of lysine or N-terminal residues calculated from the experimental structures in the Brookhaven Database (2LYM, 3RSP, 2AKQ, 1AKK, 1MBO, 1UJZ). Both the range and average distances are given for the set of identified cross-linked residues. The maximal cross-linking distance expected for DSG/DSG is 20/24 Å,[25,26] respectively. [f] data combined from two LC−MS runs

Table 2. Monolinks Identified in the Colicin E7 DNAse/Im7 Heterodimer Using DSG and DSS

residue no.[a]	identified with DSG[b]	identified with DSS[b]
A N-terminus	X	X
A4		
A20	X	
A24		
A43	X	
A70		X
A73	X	X
A81	X	X
A85	X	X
B N-terminus		
B446		
B449		
B452		
B456		
B458		
B463	X	
B470		X
B483	X	X
B487	X	
B490	X	
B497		
B498		
B505	X	X
B511	X	X
B522		X
B525		X
B537	X	X
B547	X	X
B567	X	X

[a] Listed are the N-terminus of the A and B subunits and all lysine residues. [b] X designates that a monolink to the listed residue was identified.

power of the method for finding a large number of cross-links among the much larger collection of nonmodified and monolink peptides. Previously reported methods for finding cross-links typically lead to the identification of only a few cross-links per protein species.

Of the 15 unique residue pair cross-links identified using DSS, 3 of them were also found using DSG. Thus, despite the different lengths between the reactive esters in these reagents, they do not seek out a completely nonoverlapping set of residue pairs. Of the 15 and 11 unique residue pair cross-links found with DSS and DSG, respectively, 5 (A73/B−N-terminus not included in Figure 7) and 1 of them, respectively are intersubunit cross-links. As expected, most of the identified cross-links are within the same subunit, and for the shorter of the two cross-linking reagents, DSG, more looplinks and intramolecular cross-links were found. This means that, by applying a set of cross-linking reagents of different cross-linking spacer length such as DSS and DSG to the same protein sample, the yield for inter- vs intramolecular can be tuned according to the focus of the protein structure study. One might assume that cross-linking with DSS (11.4 Å spacer length) rather than DSG (7.7 Å) would lead to more cross-linked and less monolinked peptide species for a given protein sample because of the potentially larger number of residues in reach for DSS. However, we observed similar numbers of monolinks with both reagents, 14 with DSS and 15 with DSG. The number of observed crossllinks was 7 with DSG and 4 with DSS. These trends will be difficult to predict a priori. It seems that the best use of the cross-linking information is to rule out structure predictions where the two reactive amino acids are simply too far to be cross-linked together.

Five out of a total of 19 structurally unique monolinks were automatically identified for the Colicin E7 DNAse/Im7 heterodimer with iXLINK by the discernible mass splitting, caused

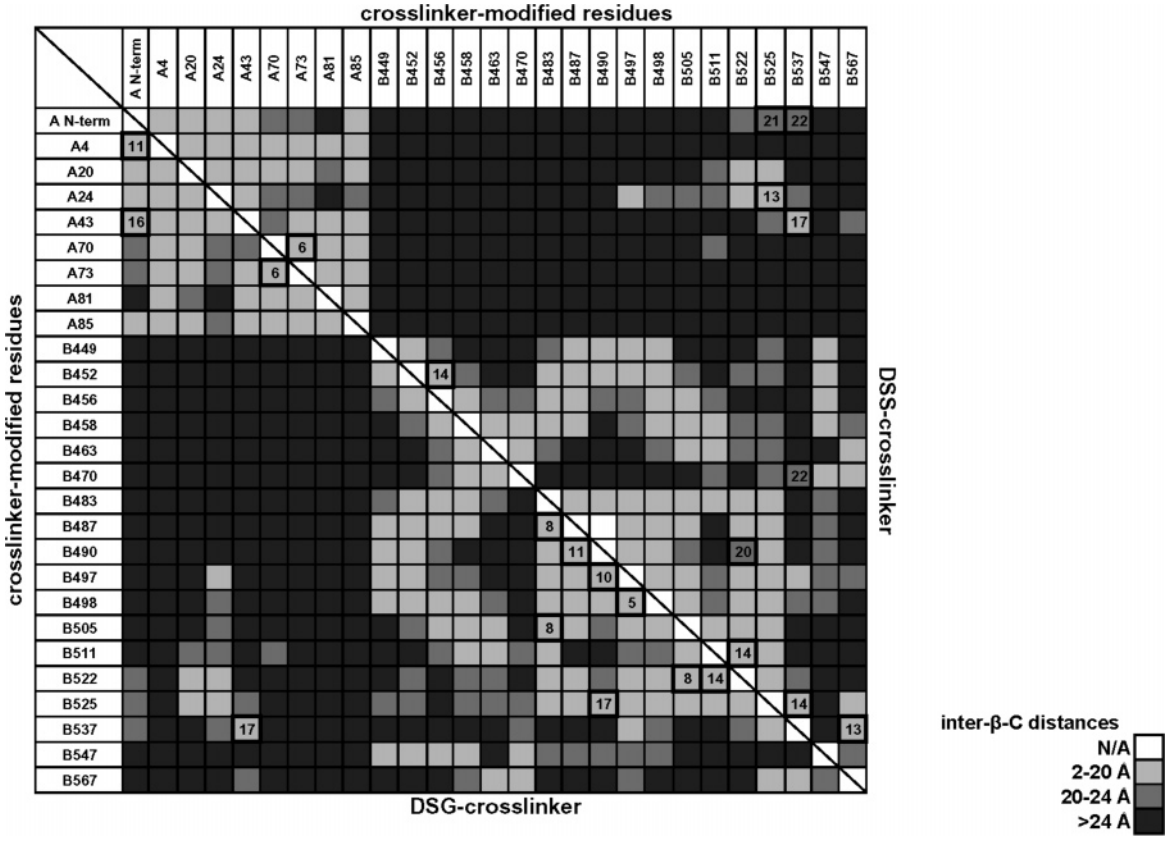

Figure 7. Results from cross-linking analysis of the Colicin E7 DNAse/Im7 heterodimer with DSG (below the diagonal line) and DSS (above the diagonal line) are shown in the context of the published X-ray structure of this complex by Kortemme et al. (1UJZ.pdb).[17] To the left and on the top of the chart is a list of all lysine residues and the N-termini (N-term), sorted according to their position in the protein sequence (A and B designate the two subunits). Each square symbolizes a unique pair of residues color coded by inter β-C distances according to the legend shown to the right side of the figure. Squares containing a number, the inter-β-C distance given in Å, are for those residue pairs that comprise the identified cross-linked species. As a consequence to the sorting, residues close in amino acid sequence are all clustered around the diagonal line (top left to bottom right).

by [^{18}O]-labeling during the cross-linking reaction with DSG and DSS. Because of the mass resolution limit of the MALDI spectrometer for high MW species, the 2 Da shift caused by [^{18}O] substitution was not reliably observed for species of MW > ~2000 Da. For these high MW species, assignment to monolinks vs cross-links was easily accomplished with doX-LINK analysis of the tandem mass spectrometry data. Given this resolution issue, we carried out the cross-linking analysis of the other proteins studied (beta-lactoglobulin, cytochrome *c*, lysozyme, myoglobin, ribonuclease A) without the use of the [^{18}O]solvent strategy.

Discussion

All 23 of the unique residue pair cross-links identified for the Colicin E7 DNAse/Im7 heterodimer are sufficiently close in 3D to allow cross-linking without significant structural distortion of the protein. Consequently the observed cross-linked residue pairs are reasonble constraints for 3D protein structure prediction. The type of graphical representation shown in Figure 7 shows that potential "informative" areas for cross-linking within a protein or protein complex—informative for 3D structure prediction—can be found further away from the diagonal line. This is because those areas would include residue pairs that are not necessarily close in sequence. Good examples for cross-links between residues far apart in sequence in this study are the intersubunit links A43/B537 or B490/B525. Other identified cross-links that could be critical for discrimination of computed protein structures are: A−N-terminus/A43; A−N-terminus/B525; and A−N-terminus/B537, A24/B525, A73/B−N-terminus, B470/B537, B−N-terminus/B537, B446/B497, B470/B537, and B490/B525.

The monolinks identified (Table 2) are useful to measure the reproducibility of the method (cross-linking experimental condition and mass spectrometry acquisition/analysis). All of the monolinks identified (Table 2) involve lysine residues that are on the surface of the Colicin E7 DNAse/Im7 heterodimer. Thus, monolink information is presumably also useful to distinguish surface lysines from buried ones and to distinguish lysine residues that are solvent exposed from those that are protected from reaction with the cross-linker by lying at the surface of the protein−protein contact interface in multisubunit protein complexes.

It is important to note that inter-heterodimer cross-links of the type shown in panel D of Figure 1 are not detected. This is presumably because at the employed protein concentration these cross-links are kinetically disfavored (they require two bi-molecular reactions, whereas intra-heterodimer cross-links require only one). Presumably, a second factor is that there are a very large number of possible inter-heterodimer cross-links; if they were to form randomly, any particular one would be present at very low abundance. These are just hypotheses;

we have no data other than the fact that we did not find inter-heterodimeric cross-links.

An additional point concerns the use of deuterium as the heavy isotope incorporated into the cross-linker reagent. It is known that this isotope often leads to a change in the chromatographic retention time on reverse-phase columns; the deuterated material often elutes slightly earlier than the light analogue from the column. This isotope effect on retention time will lead to a distortion of the ratio of light-to-heavy pairs for cross-linker-modified peptides away from the expected ratio of 1:1. However, this is not a significant problem because the chromatographic shift is typically small compared to the time used to collect each MALDI plate spot (>8 s in our study). Nevertheless, iXLINK allows the user to specify a range for the peak intensity ratio for the light and heavy components of the isotopic pairs of cross-linker-modified peptides. If we allowed this intensity ratio to be in the range of 0.33−3.0, all of the cross-linker-modified peptides listed in Table 1 were found, and 90% of these were found if the intensity ratio was allowed to lie in the range of 0.5−2.0. The alternative is to use heavy isotopes such as ^{13}C, ^{15}N, and ^{18}O. This would represent a simple improvement to our method.

doXLINK attempts to score candidate peptides including monolinked, looplinked and cross-linked peptides based on their tandem MS spectra. We have so far collected ca. 200 MALDI tandem MS spectra of cross-linked peptides. During our computer-based cross-linked peptide analysis it was apparent that the scoring scheme alone is often insufficient to make firm molecular species assignments. In other words, the confidence of the assignment of the experimental tandem MS spectra to a molecular species could not always be well represented by a single numerical score. In many cases, we needed to rely on manual inspection of the data using XLinkViewer to make high confidence peptide spectral assignments. Given below is a list of guidelines that we relied on for our analyses:

(1) Monolinks can be easily identified with our method. Tandem MS spectra of monolinks are relatively rich in b- and y-ions compared to those from cross-links. In addition to doXLINK's fragment ion matching, reporter ions known to be diagnostic for monolinks such as the ones shown in Figure 6 are used in many cases to identify monolinks with high confidence. Since monolinks are more easily identified than cross-links, we first assigned monolinks and then used the experimental masses and the calculated masses to readjust our mass accuracy tolerances, which can be applied to the mass window for the doXLINK peptide search. Reducing the magnitude of the mass accuracy tolerances as much as possible helps with the assignment of cross-link species since fewer candidate species are outputted by iXLINK.

(2) In our analysis of the Colicin E7 DNAse/Im7 heterodimer, for cross-link assignments, we insisted that at least three experimental b- or y-ions matched to theoretical fragments. These b- and y-ions were preferably evenly distributed throughout the whole tandem MS spectrum. XLinkViewer uses different colors to label b-and y-fragment ions.

(3) For high confidence assignments, we relied heavily on the mass shifts of fragment ions due to the heavy and light cross-linker. doXLINK takes into account the fact that a fragment ion in the tandem MS spectrum of the light cross-linker modified precursor ion is mass shifted compared to the corresponding ion in the tandem MS spectrum of the heavy precursor ion as long as the fragment ion contains the cross-linker. This analysis is incorporated into the doXLINK scoring function, and XLinkViewer displays this information using a coloring pattern.

(4) The use of trypsin and Asp-N together for digestion sometimes results in a subset of cross-linker-modified peptides that contain C-terminal residues other than Lys and Arg. Particularly poor CID fragmentation was observed in some cases for peptide species lacking a C-terminal Lys or Arg residue. These basic residues seem to be essential for good fragmentation in MALDI tandem MS. Therefore, the presence of a nontryptic C-termini of a peptide candidate could explain a low doXLINK matching score. Manual inspection of the data using XLinkViewer helps sort this out.

(5) The number of theoretical fragment ions expected for a cross-linked peptide species grows with the size of the species. Hence more matching fragment ion signals should be observed for larger molecules. The investigator should keep this in mind, especially when monolink and cross-link peptides are contained in the list of candidates that cannot be distinguished by making use of reporter ion signatures (see point 1 above).

The guidelines above should be taken into consideration in the manual validation process of cross-linking tandem MS data, and XLinkViewer allows the user to quickly gather all this information, compare multiple peptide candidates according to their theoretical fragment ion features, and either accept or reject certain doXLINK results. Additional information for our automatic and subjective tandem MS analysis is given in a detailed user manual, which can be found at http://www.systemsbiology.org/Resources_and_Development/Downloadable_Software. Recently a systematic study of dissociation patterns of cross-linked peptides was reported by Gaucher et al.[23] although the fragmentation of such peptides is still not fully understood.

The total amount of protein used for one cross-linking experiment is 100 μg, which is sufficient for more than 10 chromatographic/MALDI plate spotting runs. Each of these runs takes between 45 and 70 min. Collection of a MALDI−MS dataset for 1 plate with 192 spots takes approximately 25 min, and after ~1 h the subsequent tandem mass spectrometry acquisition can be started. For 800 tandem mass spectra (including light/heavy precursor mass pairs) data acquisition can take up to 7 h. Data export to create doXLINK input takes 30 min, and a preliminary doXLINK data analysis takes the experienced user ~1 h. Tandem mass spectrometry analysis (using subjective manual analysis) takes another ~2 h. Hence, we anticipate the total time for processing to be in the range of one to 2 days, only a small fraction of which is hands-on time.

In principle, our method can be carried out with any cross-linking reagent that can be synthesized in isotopically light and heavy forms. The use of bis-acylating agents allows, in principle, cross-links and monolinks to be distinguished using isotopic solvent; however, we expect the tandem mass spectrometry data to be sufficient in most cases to distinguish monolinks from cross-links. Furthermore, the splitting by ^{18}O of 2 Da is insufficient in the case of high MW species, as noted in the Results section. The new method is expected to work well with other types of cross-linking reagents including thiol-specific and photoactivatable reagents.

As noted in the Introduction, other methods to use mass spectrometry to identify protein cross-links rely on cross-linking reagents bearing an affinity handle so that cross-linker-modified peptides can be enriched by affinity capture. In our

method, we face the signal vs noise determination problem wherein we must distinguish peaks of labeled (cross-linked) peptide origin (the signal) from peaks of unlabeled peptide origin (the noise). However, when combined with LC, we fractionate our tryptic digest into several fractions, thus increasing the likelihood that cross-linker-modified peptide species will be found. It should also be mentioned that affinity-based enrichment methods can suffer from loss of peptides due to nonspecific absorption during affinity capture and release. A second point is that obtaining MALDI−MS/MS data is a relatively fast process without the need for manual intervention. Thus, it is not a major problem if MS/MS data is collected on some species that turn out to be peptides or non-peptide impurities that do not bear the cross-linker.

On the basis of preliminary data obtained with the newer spectrometer from Applied Biosystems, the 4800 MALDI-TOF/TOF analyzer, we found that MALDI MS/MS spectra are richer in fragmentation data, especially for higher m/z ratios, compared to spectra obtained with the older instrument, the 4700 Proteomics analyzer, which was used in this study. This promises to minimize the amount of manual inspection of MS/MS data in order to make high confidence peptide species identifications. The use of MS/MS data is expected to replace the $^{18}O/^{16}O$ experimental approach described in this paper for distinguishing monolinks from cross-links, especially for species of higher m/z ratios.

The method described in this study when applied to homo-multimeric proteins is problematic in that it is not possible to tell whether an observed cross-link is inter- vs intramolecular. One can envision a strategy for homodimers in which heavy and light isotope labeled subunits (from cell culture expression in the presence of heavy or light amino acids) are mixed together to give a binomial distribution of labeled proteins. If such a protein is analyzed with heavy and light cross-linkers (as carried out in the present study), then one obtains a different isotope splitting pattern depending on whether the cross-link is inter- vs intramolecular.

In the Colicin E7 DNAse/Im7 heterodimer study, our method was used to identify cross-linked peptides, including those that define the protein−protein interface. If applied to the analysis of protein complexes isolated by immunoprecipitation or after affinity tagging of one component of a protein complex, then the method is expected to resolve the average complex composition determined by traditional methods (for example, see Rigaut et al.[27]) into specific species of defined composition. The method, therefore, will provide structural information on wide utility for protein and proteome research.

Acknowledgment. This work was supported with federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health under Contract No. N01-HV-28179. Special thanks to Richard Bonneau, Nichole King, Patrick Pedrioli, and Brian Smart for technical help with this study.

Supporting Information Available: Figures 8−12 are results from cross-linking reactions of beta-lactoglobulin, cytochrome *c*, lysozyme, myoglobin, ribonuclease A with the cross-linking reagents DSS and DSG, analyzed with iXLINK, doXLINK, and XLinkViewer, the graphical representation of these results follows the figure legend of Figure 7. Tables 3−7 are the results of these same experiments, containing all residues identified as monolinks. This material is available free at http://pubs.acs.org.

References

(1) Bennett, K. L.; Matthiesen, T.; Roepstorff, P. Probing protein surface topology by chemical surface labeling, cross-linking, and mass spectrometry. *Methods Mol. Biol.* **2000**, *146*, 113−131.
(2) Brunner, J. New photolabeling and cross-linking methods. *Annu. Rev. Biochem.* **1993**, *62*, 483−514.
(3) Young, M. M.; Tang, N.; Hempel, J. C.; Oshiro, C. M.; Taylor, E. W.; Kuntz, I. D.; Gibson, B. W.; 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*, 5802−5806.
(4) Rappsilber, J.; Siniossolou, S.; Hurt, E. C.; Mann, M. A Generic Strategy To Analyze the Spatial Organization of Multi-Protein Complexes by Cross-Linking and Mass Spectrometry. *Anal. Chem.* **2000**, *72*, 267−275.
(5) Schilling, B.; Row, R. H.; Gibson, B. W.; Guo, X.; Young, M. M. MS2Assign, automated assignment and nomenclature of tandem mass spectra of chemically cross-linked peptides. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 834−850.
(6) Ye, X.; O'Neil, P. K.; Foster, A. N.; Gajda, M. J.; Kosinski, J.; Kurowski, M. A.; Bujnicki, J. M.; Friedman, A. M.; Bailey-Kellogg, C. Probabilistic cross-link analysis and experiment planning for high-throughput elucidation of protein structure. *Protein Sci.* **2004**, *13*, 3298−3313.
(7) Sinz, A. Chemical cross-linking and mass spectrometry for mapping three-dimensional structures of proteins and protein complexes. *J. Mass. Spectrom.* **2003**, *38*, 1225−1237.
(8) Geyer, H.; Geyer, R.; Pingoud, V. A novel strategy for the identification of protein-DNA contacts by photo-cross-linking and mass spectrometry. *Nucleic Acids Res.* **2004**, *32*, e132.
(9) Back, J. W.; de Jong, L.; Muijsers, A. O.; de Koster, C. G. Chemical Cross-linking and Mass Spectrometry for Protein Structural Modeling. *J. Mol. Biol.* **2003**, *331*, 303−313.
(10) Petrochenko, E. V.; Olkhovik, V. K.; Borchers, C. H. Isotopically Coded Cleavable Cross-linker for Studying Protein−Protein Interaction and Protein Complexes. *Mol. Cell. Proteomics* **2005**, *4*, 1167−79.
(11) Sinz, A. Chemical Cross-linking and Mass Spectrometry to Map Three-Dimensional Protein Structures and Protein−Protein Interactions. *Mass Spectrom. Rev.* **2006**, *25*, 663−682.
(12) Trester-Zedlitz, M.; Kamada, K.; Burley, S. K.; Fenyo, D.; Chait, B. T.; Muir, T. W. A modular cross-linking approach for exploring protein interactions. *J. Am. Chem. Soc.* **2003**, *125*, 2416−2425.
(13) Hurst, G. B.; Lankford, T. K.; Kennel, S. J. Mass spectrometric detection of affinity purified cross-linked peptides. *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 832−839.
(14) Müller, D. R.; Schindler, P.; Towbin, H.; Wirth, U.; Voshol, H.; Hoving, S.; Steinmetz, M. O. Isotope-Tagged Cross-Linking Reagents. A New Tool in Mass Spectrometric Protein Interaction Analysis. *Anal. Chem.* **2001**, *73*, 1927−1934.
(15) Collins, C. J.; Schilling, B.; Young, M.; Dollinger, G.; Guy, R. K. Isotopically labeled cross-linking reagents: resolution of mass degeneracy in the identification of cross-linked peptides. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 4023−4026.
(16) Kersey, P. J.; Duarte, J.; Williams, A.; Karavidopoulou, Y.; Birney, E.; Apweiler, R. The International Protein Index: An integrated database for proteomics experiments. *Proteomics* **2004**, *4* (7), 1985−1988.
(17) Kortemme, T.; Joachimiak, L. A.; Bullock, A. N.; Schuler, A. D.; Stoddard, B. L.; Baker, D. Computational redesign of protein−protein interaction specificity. *Nat. Struct. Mol. Biol.* **2004**, *11*, 371−379.
(18) Zhang, N.; Aebersold, R.; Schwikowski, B. A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* **2002**, *2*, 1406−1412.
(19) Clauser, K. R.; Baker, P. R.; Burlingame, A. L. Role of accurate mass measurement (± 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* **1999**, *71*, 2871−2882.
(20) Wefing, S.; Schnaible, V.; Hoffmann, D. SearchXLinks, http://www.searchxlinks.de/, center of advanced european studies and research (caesar), Bonn, Germany, **2001.**
(21) Tang, Y.; Chen, Y.; Lichti, C. F.; Hall, R. A.; Raney, K. D.; Jennings, S. F. CLPM: a cross-linked peptide mapping algorithm for mass

spectrometric analysis *BMC Bioinformatics* **2005** Jul 15; 6 Suppl 2: S9.
(22) de Koning, L. J.; Kasper, P. T.; Back, J. W.; Nessen, M. A.; Vanrobaeys, F.; Van Beeumen, J.; Gherardi, E.; de Koster, C. G.; de Jong, L. Computer-assisted mass spectrometric analysis of naturally occurring and artificially introduced cross-links in proteins and protein complexes *FEBS J.* **2006**, *273*, 281−91.
(23) Gaucher, S. P.; Hadi, M. Z.; Young, M. M. Influence of Cross-linker Identity and Position on Gas-Phase Dissociation of Lys-Lys Cross-linked Peptides. *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 395−405.
(24) Paizs, B.; Suhai, S. Fragmentation pathways of protonated peptides. *Mass Spectrom. Rev.* **2005**, *24*, 508−548.
(25) Green, N. S.; Reisler, E.; Houk, K. N. Quantitative evaluation of the lengths of homobifunctional protein cross-linking reagents used as molecular rulers. *Protein Sci.* **2001**, *10*, 1293−1304.
(26) Huang, B. X.; Dass, C.; Kim, H. Y. Probing conformational changes of human serum albumin due to unsaturated fatty acid binding by chemical cross-linking and mass spectrometry. *Biochem. J.* **2005**, *387*, 695−702.
(27) Rigaut, G.; Shevchenko, A.; Rutz, B.; Wilm, M.; Mann, M.; Seraphin, B. A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **1999**, *17*, 1030−1032.

PR060154Z