Computational methods in mass spectrometry-based structural proteomics for studying protein structure, dynamics, and interactions

Seungjin Na, Eunok Paek*
Dept. of Computer Science, Hanyang University, Seoul 04763, Republic of Korea

ABSTRACT

Mass spectrometry (MS) has made enormous contributions to comprehensive protein identification and quantification in proteomics. MS is also gaining momentum for structural biology in a variety of ways, complementing conventional structural biology techniques. Here, we will review how MS-based techniques, such as hydrogen/deuterium exchange, covalent labeling, and chemical cross-linking, enable the characterization of protein structure, dynamics, and interactions, especially from a perspective of their data analyses. Structural information encoded by chemical probes in intact proteins is decoded by interpreting MS data at a peptide level, i.e., revealing conformational and dynamic changes in local regions of proteins. The structural MS data are not amenable to data analyses in traditional proteomics workflow, requiring dedicated software for each type of data. We first provide basic principles of data interpretation, including isotopic distribution and peptide sequencing. We then focus particularly on computational methods for structural MS data analyses and discuss outstanding challenges in a proteome-wide large scale analysis.

© 2020 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents

1. Introduction ..................................................................................................... 1392
2. Basics of mass spectrometry ........................................................................ 1392
3. Hydrogen/deuterium exchange .................................................................... 1392
   3.1. Fundamentals of data interpretation ...................................................... 1393
   3.2. The determination of deuterium incorporation ....................................... 1394
   3.3. Deconvolution of bimodal behavior ...................................................... 1396
4. Covalent labeling ........................................................................................... 1396
   4.1. Fundamentals of data interpretation ...................................................... 1396
   4.2. MS/MS-based identification of modified residues ................................. 1396
5. Chemical cross-linking .................................................................................. 1397
   5.1. Fundamentals of data interpretation ...................................................... 1397
   5.2. Computational methods for cross-linked peptides ................................. 1397
   5.3. Disulfide bond ..................................................................................... 1399
   5.4. Confidence and false positives ............................................................. 1399
6. Conclusions .................................................................................................. 1400
Declaration of Competing Interest .................................................................... 1400
Acknowledgments ............................................................................................ 1400
References ....................................................................................................... 1400

* Corresponding author at: Dept. of Computer Science, Hanyang University, 222 Wangsimni-ro, Seongdong-gu, Seoul 04763, Republic of Korea. Tel: +82-2-2220-2377; fax: +82-2-2299-2377.
E-mail address: eunokpaek@hanyang.ac.kr (E. Paek).

https://doi.org/10.1016/j.csbj.2020.06.002
2001-0370/© 2020 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
1. Introduction

Protein function is regulated by changes in its three-dimensional structure and enzymatic activity in response to protein modifications, ligand binding, or protein–protein interactions. Elucidating the three-dimensional structures of proteins, therefore, is an important step toward understanding their role in molecular functional biology. For decades, the structural models of proteins have been studied at atomic resolution using X-ray crystallography [1], nuclear magnetic resonance (NMR) spectroscopy [2], and (cryo-)electron microscopy (EM) [3]. However, these approaches are not applicable to all proteins and protein complexes owing to their limited amount, low solubility, large size, or unavailable crystals. To bridge the gap and study dynamic protein complexes, there are needs for complementary techniques taking advantages of 1) analyzing proteins in native solution condition; 2) dealing with limited sample amounts; 3) analyzing protein complexes or interactions; and 4) providing information for active dynamic structural changes of a protein under various biological conditions. In that respect, mass spectrometry (MS) has become increasingly popular in the field of structural biology [4]. Over the past few decades, MS has made substantial contributions to proteomics, from the identification and quantification of individual proteins, to high throughput analyses of whole proteome and the studies of protein localization and post-translational modifications [5–7]. Recent high-performance liquid chromatography (HPLC) and high MS acquisition speed enabled the detection of tens of thousands of peptides in an hour [8], the detection of over 10,000 proteins from human cell lines [9], and the exploration of the human interactome consisting of ~56,000 interactions [10]. MS approaches have also been applied to various structural biology studies ranging from epitope mapping and protein–ligand interactions to probing structures of membrane proteins [11–18]. MS coupled with chemistry could infer global, local and site-specific structural information with high sensitivity and accuracy even for low concentration proteins.

The unique strength of MS is its ability to detect changes in each individual amino acid level as well as in protein and peptide levels. MS-based structural proteomics can be divided into two main streams, 1) peptide-centric and 2) protein-centric, depending on what analyze MS actually analyzes to obtain the structural information [19–21]. In peptide-centric methods, more popular than protein-centric methods, the concept is that 1) chemical reagents introduce modifications into native proteins or protein complexes in solution, 2) such modifications encode some structural information about which local regions are exposed/buried and which residues are in close spatial proximity (mediating protein–protein interactions) in native proteins or complexes, 3) MS-based proteomics identifies and localizes the modifications at the peptide level to infer the structural information. In a protein-centric method, called native MS, MS directly analyzes intact biomolecules that include all subunits and cofactors that make up the functional complex. It is possible only when the process of native MS can maintain weak noncovalent interactions between protein subunits and associated biomolecules such as DNA, cofactors, and ligands. The mass and charge state distribution detected by native MS allows for the determination of protein complex stoichiometry, cofactor content, and conformation. Native MS has also been coupled with other hybrid techniques such as ion mobility spectrometry, hydrogen/deuterium exchange, and top-down proteomics. Ion mobility spectrometry enables an additional separation by the overall shape of protein complexes [22], hydrogen/deuterium exchange reveals the conformational dynamics of intact proteins [23], and top-down proteomics, sequencing intact proteins, provides an additional level of information about specific proteoforms arising from sequence variations and post-translational modifications [24]. Although MS-based methods provide opportunities to probe the protein structure, dynamics, and interactions in native environments, the individual pieces of data from different methods typically do not provide sufficient information to derive a structural model of a protein or complex by itself. The structural features and constraints monitored by MS-based methods can be used complementary to conventional techniques such as X-ray crystallography, EM, and NMR spectroscopy. The improved accuracy and completeness of a model will be achieved by simultaneous use of all such information from multiple sources of data [25–27].

This review will focus on computational methods and discuss challenges in three MS-based peptide-centric methods, 1) hydrogen/deuterium exchange - exchanging peptide amide hydrogen atoms with deuterium atoms in solution, 2) covalent labeling - introducing irreversible modifications to amino acid residues, and 3) chemical cross-linking - covalently linking two spatially proximal amino acid residues. The three approaches generate different type of MS data, each requiring dedicated software to decode the structural information. The MS-based identification of disulfide bonds involved in protein tertiary structures will be also discussed under chemical cross-linking category owing to their similarity.

2. Basics of mass spectrometry

MS has become a powerful tool for analyzing complex protein mixtures [28]. Protein mixtures are digested by residue-specific enzymes, and the digested peptides are separated by liquid chromatography (LC) and high MS acquisition speed enabled the detection of tens of thousands of peptides in an hour [8], the detection of over 10,000 proteins from human cell lines [9], and the exploration of the human interactome consisting of ~56,000 interactions [10]. MS approaches have also been applied to various structural biology studies ranging from epitope mapping and protein–ligand interactions to probing structures of membrane proteins [11–18]. MS coupled with chemistry could infer global, local and site-specific structural information with high sensitivity and accuracy even for low concentration proteins.

The unique strength of MS is its ability to detect changes in each individual amino acid level as well as in protein and peptide levels. MS-based structural proteomics can be divided into two main streams, 1) peptide-centric and 2) protein-centric, depending on what analyze MS actually analyzes to obtain the structural information [19–21]. In peptide-centric methods, more popular than protein-centric methods, the concept is that 1) chemical reagents introduce modifications into native proteins or protein complexes in solution, 2) such modifications encode some structural information about which local regions are exposed/buried and which residues are in close spatial proximity (mediating protein–protein interactions) in native proteins or complexes, 3) MS-based proteomics identifies and localizes the modifications at the peptide level to infer the structural information. In a protein-centric method, called native MS, MS directly analyzes intact biomolecules that include all subunits and cofactors that make up the functional complex. It is possible only when the process of native MS can maintain weak noncovalent interactions between protein subunits and associated biomolecules such as DNA, cofactors, and ligands. The mass and charge state distribution detected by native MS allows for the determination of protein complex stoichiometry, cofactor content, and conformation. Native MS has also been coupled with other hybrid techniques such as ion mobility spectrometry, hydrogen/deuterium exchange, and top-down proteomics. Ion mobility spectrometry enables an additional separation by the overall shape of protein complexes [22], hydrogen/deuterium exchange reveals the conformational dynamics of intact proteins [23], and top-down proteomics, sequencing intact proteins, provides an additional level of information about specific proteoforms arising from sequence variations and post-translational modifications [24]. Although MS-based methods provide opportunities to probe the protein structure, dynamics, and interactions in native environments, the individual pieces of data from different methods typically do not provide sufficient information to derive a structural model of a protein or complex by itself. The structural features and constraints monitored by MS-based methods can be used complementary to conventional techniques such as X-ray crystallography, EM, and NMR spectroscopy. The improved accuracy and completeness of a model will be achieved by simultaneous use of all such information from multiple sources of data [25–27].

This review will focus on computational methods and discuss challenges in three MS-based peptide-centric methods, 1) hydrogen/deuterium exchange - exchanging peptide amide hydrogen atoms with deuterium atoms in solution, 2) covalent labeling - introducing irreversible modifications to amino acid residues, and 3) chemical cross-linking - covalently linking two spatially proximal amino acid residues. The three approaches generate different type of MS data, each requiring dedicated software to decode the structural information. The MS-based identification of disulfide bonds involved in protein tertiary structures will be also discussed under chemical cross-linking category owing to their similarity.

2. Basics of mass spectrometry

MS has become a powerful tool for analyzing complex protein mixtures [28]. Protein mixtures are digested by residue-specific enzymes, and the digested peptides are separated by liquid chromatography (LC), thereby reducing the complexity of the peptide mixture prior to MS. The separated peptides are then subjected to electrospray ionization (ESI) and introduced into a mass spectrometer (Fig. 1a). The instrument generates a mass spectrum (MS1), in which a peptide ion is detected and recorded as a peak with the intensity at its mass-to-charge ratio (m/z) value measured by a mass analyzer. The measurement accuracy is achieved at a level of a few parts per million (ppm) using high-resolution mass analyzers such as Orbitrap and Fourier transform ion cyclotron resonance (FTICR).

Tandem mass spectrometry (MS/MS or MS2) is used to obtain peptide sequence information (Fig. 1b). The isolated peptide, a precursor ion, is split into fragment ions by low-energy collision-induced dissociation (CID) and their peak (a pair of m/z and intensity) information is recorded in MS/MS spectra. Series of these fragment ions detected at MS/MS readily reveal the amino acid sequence of the precursor ion. To interpret an experimentally obtained MS/MS spectrum as a peptide, the most popular approach is to compare the experimental MS/MS spectrum with theoretical MS/MS spectra, generated from candidate peptides stored in a protein sequence database, using database search software [29–31] (Fig. 1c). The software retrieves from the database, candidate peptides whose masses are within a specified mass tolerance of a precursor ion mass. The validity of peptide-spectrum matches can be assessed by target-decoy strategy [32]. For a comprehensive review of the peptide and protein identification, see refs [33,34].

3. Hydrogen/deuterium exchange

Surface labeling is based on the concept that solvent-exposed regions in proteins will react quickly with labeling reagents and therefore will be labeled/modified, while buried regions will be labeled/modified slowly or not at all [35]. Protein conformational changes or protein–ligand binding can affect the degree of solvent exposure for certain protein regions, and the changes in labeled/modified degree by labeling reagents indicate which regions are
undergoing a structural change or forming an interface with an interacting partner. The best-known and most widely used strategy for surface labeling is hydrogen–deuterium exchange (HDX) that monitors the exchange of backbone amide hydrogens with deuteriums [36–38]. The exchange rates are sensitive to changes in hydrogen bonding, secondary structure, solvent accessibility and dynamics. The general HDX-MS workflow is depicted in Fig. 2a. The target protein(s) are incubated with D2O to exchange accessible hydrogen atoms with deuterium atoms. The exchange reaction is quenched at different time points to plot deuteration rate as a function of exchange time (from seconds to days). The deuterated proteins are subjected to proteolytic digestion followed by LC-MS. MS measures mass increases of peptides by deuterium incorporation. The amount of deuterium incorporation is usually determined using only peptide masses without MS/MS fragmentation due to H/D scrambling under conventional collisional activation (intramolecular H/D rearrangement). Electron-mediated fragmentation techniques such as electron capture and transfer dissociation (ECD/ETD) can be employed to avoid such scrambling and measure the exchange at the individual amino acid level [39]. In fact, the digestion and LC-MS workflow leads to back exchange [36]. To minimize the back exchange effect, in most applications of HDX, the deuterated proteins are digested using pepsin at low temperature and low pH (with a minimum at ~2.5) and the peptide mixtures are separated through chromatography columns cooled to temperatures close to 0 °C. Back exchange may be corrected on the basis of deuterium incorporation in a completely deuterated sample [37]. Most HDX analyses, however, measure relative rather than absolute deuterium incorporation and studies have shown that back exchange correction does not affect relative measurement.

HDX is a non-covalent labeling approach, while covalent labeling measures the solvent accessibility of side chains by introducing modifications to side chains. Covalent labeling is discussed in Section 4.

3.1. Fundamentals of data interpretation

HDX data analysis generates a list of peptides and quantifies the deuterium incorporated to each peptide at each labeling time point. The peptide list is usually compiled from database search-driven identification of a separated MS1 spectra. To increase the protein sequence coverage, additional peptides from in silico digestion (using protease specificity) can be considered if they can be assigned to peaks in MS1 spectra. It is assumed that the chromatographic properties of deuterated peptides are identical to those of natural peptides and thus the isotopic distributions of deuterated peptides can be annotated using the retention time of natural peptides. Since there are small differences in retention time between the two, the deuterated isotopic
A topic distribution is the union of isotopic distributions of different peptides with different deuteration levels. Thus, the deuterated isotopic distribution can result in multiple species of deuterated incorporation to a peptide. The incorporation of these peptides can be measured by monitoring their isotopic distributions by LC-MS (shown at the bottom). The high precision and sensitivity of mass instruments enable the detection of such subtle changes. Covalent labeling MS is similar to that of HDX-MS. The difference is that covalent labeling introduces irreversible modifications (marked by stars in this figure) to soluble accessible regions. The protected region of protein A was rarely modified when interacting with protein B. The covalent modifications are highly stable during sample handling and thus modified sites as well as modification types on the same peptides can be distinguished by MS/MS as shown at the bottom. Chemical cross-linking MS covalently connects two residues in close spatial proximity between two proteins or within a protein. The constraints obtained by the chemical cross-links narrow down the possible location of the binding interface. Since the link of two peptides is covalent, the analysis is amenable to MS/MS. In the MS/MS spectra, fragment ions from two peptides are present at the same time. In this figure, curvy lines represent peptides. Lines or parts in lines shaded in red mean that their sequences belong to the red region in protein A. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Deuterium incorporation into a peptide can be analyzed mainly in two different ways, 'centroid' or 'theoretical fitting' methods [40]. The centroid methods simply calculate the difference between the weighted average mass of the deuterated isotopic distribution and its natural (non-deuterated) isotopic distribution. Theoretical fitting methods determine the number of incorporated deuterium by fitting theoretical deuterated isotopic distributions with the observed isotopic distribution.

### 3.2. The determination of deuterium incorporation

Many software tools are available for automated data analysis and have made HDX-MS more widely used [41–56]. Theoretical fitting methods are computationally more expensive, but are often more accurate than centroid methods. As deuterium is gradually incorporated, the number of incorporated deuterium is not deter-
mined as a single value, but as a distribution of incorporated deuterium numbers. More specifically, theoretical fit method is to solve linear equations, where the (partially) deuterated isotopic distribution is defined as convolution of its natural isotopic distribution and the deuterium incorporation distribution of the peptide.

In many studies, the deuterium incorporation distribution of a peptide was assumed a statistical model such as the binomial distribution \([49–51]\). The methods involve three relevant distributions as shown in Fig. 3: two observed isotopic distributions- 1) natural isotopic distribution, ‘Dnat’; 2) deuterated isotopic distribution, ‘Ddeu’, and one distribution of deuterium incorporation to Ddeu, ‘Dlev’. Dlev is expected to conform to the binomial distribution \(B(n, d/n)\), where \(n\) is the number of exchangeable hydrogens in a peptide and \(d\), of interest in HDX-MS analysis, is the average deuterium incorporation. The convolution of Dnat and Dlev results in Ddeu, which spans a wider \(m/z\) range than Dnat since Dlev is not a single value but a distribution. The best fit to all of the linear equations was achieved using least-squares method (LSQ) \([52]\) or maximum entropy method (MEM) \([53]\). The LSQ-based method is simple but not desirable when the number of exchangeable hydrogens is big or signal-to-noise ratio is poor. The MEM-based method was more robust to the noise than LSQ and could result in a smoothed deuteration distribution at the expense of longer computation time. HX-Express2 \([54]\) employed Gold’s iterative deconvolution algorithm that assigns positive distribution coefficients using a gradient descent and favors smooth solutions similar to maximum entropy regularization. DEX \([55]\) applied a Fourier deconvolution method and the deconvolution step resulted in the noise reduction, because the noise is averaged by Fourier transforms. These approaches are interested in the deuterium incorporation distribution itself for a peptide (allowing variations in extent of deuterium incorporation) since deuterated isotopic distributions frequently overlap with other isotopic distributions due to their wide span. HX-Express2 \([50]\) implemented asymmetric least squares regression so that an error contribution gets greater for peaks of lower intensity than the calculated distributions, thereby favoring the calculated distribution to fit to uncontaminated peaks. Contrary to minimizing errors, deMix \([51]\) proposed a measure, referred as to ‘Matched Peak Count’, which counts matching peaks between the two distributions, whose intensities are comparable within an intensity-proportional tolerance. In particular, the measure was very robust against high-intensity interfering peaks by minimizing their effect.

As an alternative, the observed deuterated isotopic distribution could directly be deconvoluted using the natural isotopic distribution, where Ddeu is a linear combination of mass-shifted natural isotope distributions (i.e., there is no assumption about Dlev in Fig. 3). The best fit to all of the linear equations was achieved using least-squares method (LSQ) \([52]\) or maximum entropy method (MEM) \([53]\). The LSQ-based method is simple but not desirable when the number of exchangeable hydrogens is big or signal-to-noise ratio is poor. The MEM-based method was more robust to the noise than LSQ and could result in a smoothed deuteration distribution at the expense of longer computation time. HX-Express2 \([54]\) employed Gold’s iterative deconvolution algorithm that assigns positive distribution coefficients using a gradient descent and favors smooth solutions similar to maximum entropy regularization. DEX \([55]\) applied a Fourier deconvolution method and the deconvolution step resulted in the noise reduction, because the noise is averaged by Fourier transforms. These approaches are interested in the deuterium incorporation distribution itself for a peptide (allowing variations in extent of deuterium incorporation)
as well as the average deuterium incorporation, while the binomial fitting-based approaches relatively emphasize the average deuterium incorporation for a peptide.

3.3. Deconvolution of bimodal behavior

One of the difficulties in determining deuterium incorporation is that often a deuterated isotopic distribution is observed as a bimodal form, not unimodal, which arises from EX1 behavior or heterogeneous conformational populations [57]. The kinetics for HDX has two limiting regimes, referred to as EX2 and EX1 [38,58]. In EX1 condition, as a portion of the proteins unfold, the deuterium incorporation takes place on a number of residues simultaneously before returning to their ‘closed’ state. As a result, there are two populations represented as a bimodal form: 1) a lower mass population that has never undergone the slow unfolding transition and 2) a higher mass population that has undergone the transition at least once. With increasing D2O labeling time, the higher mass population would gradually grow while the lower mass population shrink. In EX2 condition (much more common under native conditions), the rate constant of the closing reaction is faster than the intrinsic rate of exchange and a progressive mass shift in single deuterated distribution is observed with increasing D2O labeling time. Most interestingly, the coexistence of two protein conformations may lead to simultaneous, progressive mass shifts in two distributions with increasing D2O labeling time. Some software tools such as ExMS [48], ExMS2 [56], HX-Express v2 [50], deMix [51], and Hexicon 2 [54] have been proposed to analyze a bimodal exchange pattern.

deMix [51] first performs an unimodal analysis via binomial fitting for an observed deuterated distribution and decides whether to perform bimodal analysis without any human intervention based on the unimodal analysis results. In the bimodal analysis, Dlev in Fig. 3 was assumed as a mixture of two binomial distributions. deMix used ‘Matched Peak Count’ for distribution fitting (as mentioned in the previous section), which not only was robust over random noises in comparing two distributions but also had strength in fast dissection of bimodal deuterated distributions by efficaciously fitting non-overlapped area. ExMS2 [56] provides various fitting options in multimodal analysis, where the number (up to 3) and shape (binomial, Gaussian, or reference) of component distributions can be set.

4. Covalent labeling

Covalent labeling is a powerful tool for monitoring the solvent accessibility of side chains in proteins or protein complexes by introducing irreversible modifications to reactive side chains [59–64]. This is conceptually similar to HDX-MS analysis except that the labeling is restricted to one or a few amino acids. The most common targets have been primary amines (Lys and N-termini), carboxyl groups (Asp/Glu and C-termini), or thiols (Cys) [59]. The general covalent labeling MS workflow is depicted in Fig. 2b. In brief, subsequent labeling of proteins, the modified proteins are subjected to proteolytic digestion followed by LC-MS/MS. Since the modifications by covalent labeling reagents are highly stable and irreversible during sample handling, the downstream analyses to calculate the extent of modification are relatively flexible and MS/MS is usually applied to determine modification types as well as modified sites in peptides. On top of the restricted labeling to a few amino acids, diethyl pyrocarbonate (DEPC) reagent can modify nucleophilic side chains (Cys, His, Lys, Thr, Tyr and Ser) and N-termini via nucleophilic substitution reactions [60]. Oxidative labeling with hydroxyl radicals introduces simultaneous modifications into many different amino acids and has been widely used in many studies [61–64].

The strength of covalent labeling is that the samples can be analyzed using conventional proteomics workflows, including sample handling techniques, LC-MS/MS, and data analysis. In HDX, optimized sample handling techniques are needed to minimize back exchange and scrambling due to reversible and labile nature of deuterium incorporation. In chemical cross-linking in Section 5, two or more peptides are connected and thus the data interpretation requires dedicated software.

4.1. Fundamentals of data interpretation

A list of unmodified and expectedly modified peptides from in silico digestion of target proteins are assigned to peptide peaks detected by MS (peptide mass fingerprinting). For confirmed peptides by MS, the peptide modification extents are calculated based on Extracted Ion Chromatograms (XIC) [65,66]. The fraction of unmodified or modified peptide is derived from the ratio of XIC areas under the unmodified/modified species to total XIC areas of all the species (including unmodified and modified). Plotting the extent of modification relative to labeling reaction time can tell which protein regions are the most solvent-exposed. Peptide mass fingerprinting relying on accurate mass is sometimes insufficient to determine modified peptides. MS/MS peptide sequencing is necessary in cases of labeling many amino acids or studying unspecific proteins (i.e., a large search space). MS/MS enables the identification of specific amino acids undergoing modifications; however many modifiable amino acids give rise to a combinatorial issue in identifying modified sites using MS/MS (we discuss this issue in the next section).

4.2. MS/MS-based identification of modified residues

MS/MS allows site resolution for modifications, which can be detected by modification-related diagnostic mass shifts of frag-

![Fig. 4. MS/MS spectra of unmodified (top) and modified (bottom) forms of a peptide PEPTDLEK are shown. In the modified spectrum, modification-related mass shifts (f10, f12 and f13 relative to e5, e6 and e7, respectively) are detected, starting with the modified site (i.e., the fifth residue D). Even with the same modification, peptides modified at different sites generate different spectra, but very similar. Theoretically, differently modified spectra can always be distinguished and thus database search tools can work for modified peptides. Experimental spectra typically contain lots of noise (shown as black peaks) that were not aligned in the bottom spectrum and matching such noise peaks might lead to incorrect assignments.](image-url)
ment ions in MS/MS spectra as shown in Fig. 4. MS/MS has been the most powerful tool to identify numerous types of modifications on peptides and a large number of software tools have been developed to search for modified peptides [67]. For residue-specific labeling [59], standard database search tools [29–31] could detect the modified sites by allowing the corresponding modification on targeted residues during search. However, if the labeling reagent can target several residues or unspecific residues [60], dedicated software to modification searches might be needed. Oxidative labeling studies with hydroxy aldehydes have reported that in practice every amino acid with the exception of Gly could be modified in oxidative labeling and various mass shifts such as +16, +14, and +32 Da were also observed [68]. In general, modification searches are performed by matching all possible modified forms of each peptide from a protein sequence database with MS/MS spectra. Given a list of possible modifications, the number of possible modified forms for a peptide of length $n$ is expressed as $\prod_{i=1}^{n} (V_i + 1)$, where $V_i$ is the number of possible modifications at $i$-th residue of the peptide. For a large scale analysis, the modification search speed and sensitivity are overwhelmed by the huge search space that follows from enumerating all possible modified forms of peptides derived from a large database. What matters more than the much longer search time, is the fact that more false positives require a more stringent threshold to maintain an acceptable error rate, resulting in the loss of true identifications [69].

A number of computational methods and strategies have been developed to identify modified peptides while taking into account more modification types [70–94]. Recent developments for modification searches have mainly focused on open search, often referred to as ‘blind’ search, a strategy that searches MS/MS spectra for modifications of arbitrary masses [67]. The open searches take into account all peptides in a database as candidates of MS/MS spectra and a modification mass is calculated as the mass difference between an MS/MS precursor ion and a candidate peptide. MSSHagger [91] could identify modified peptides in ultrafast fashion by pre-indexing the masses of all possible peptides from a database and even their fragment ions. However, most open searches usually allowed for only one unexpected modification per peptide because their accuracy and performance tend to degrade rapidly with increasing numbers of modified sites per peptide.

For oxidative labeling studies, the capacity to identify multiple modifications on a peptide is absolutely necessary to find the most solvent-exposed region since many amino acids in the region could be simultaneously modified by several type of modifications. Byonic [92] was applied to a study for probing in vivo structural dynamics of membrane proteins [95]. Byonic improved the search speed based on Lookup Peaks algorithm [93] that can fast extract candidate peptides from a database by utilizing likely b- and y-ion peaks in a spectrum. To narrow down the search space for multiple modifications, some approaches [74–77] take advantage of sequence tags which are partial sequences derived from fragment ions of a peptide in an MS/MS spectrum, e.g., a sequence tag 'LE' can be derived using fragment ions $f_{10}f_{12}f_{13}$ in the bottom spectrum of Fig. 4. On one hand, the sequence tags enable fast extraction of candidate peptides from a database. At the same time, they can localize modified regions within a peptide and thus make it possible to consider modifications only against shorter regions instead of the whole peptide’s extent. MODplus [94] further subdivides the regions into even shorter ones by simultaneously using multiple sequence tags. A dynamic programming algorithm, guided by the sequence tags, rapidly localizes modified regions in possibly modified peptides and then unexplained delta masses of the modified regions are interpreted with known modifications. In analyses of human proteome MS/MS datasets, MODplus worked well with the whole human database and ~1,000 modification types.

5. Chemical cross-linking

In cross-linking MS, cross-linking reagents covalently link two spatially proximal amino acid residues in native proteins or protein complexes, and MS/MS is applied to identify the linked amino acid residues. The two linked residues provide a powerful means to study the three-dimensional structure of proteins and protein complexes [96–98]. The general cross-linking MS workflow is depicted in Fig. 2c. In brief, proteins or protein assemblies under denaturing conditions are first reacted with bifunctional cross-linking reagents whose spacer-arm length confers a distance constraint on the two linked residues. The enzymatic digestion of the cross-linked proteins produces cross-linked peptides, followed by LC-MS/MS analysis. Subsequent informatics analysis of cross-linked MS data identifies cross-linked peptides and their linkage sites. Finally, distance constraints of the linkage sites are imposed on structure modeling. Owing to the usefulness of cross-linking MS, a large number of cross-linking reagents have been developed. Their underlying chemistry and reaction mechanisms are beyond the scope of this review. For more detailed discussion, see refs [99,100].

5.1. Fundamentals of data interpretation

Algorithms to identify cross-linked peptides would follow a strategy similar to conventional database searches for linear peptides, comparing an experimental MS/MS spectrum with theoretical spectra of candidate peptides in a database. The key difference is retrieving peptide-pairs as candidates, a computationally challenging aspect of interpreting MS/MS spectra of cross-linked peptides. A candidate peptide-pair can be determined if the sum of the masses of two peptides in a pair and the cross-linker spacer arm is matched with the precursor mass of an MS/MS spectrum within a mass tolerance. Assuming that a cross-linked peptide involves only two peptides, the number of all possible peptide-pairs as candidates is equal to the binomial coefficient $C_{n-2}$ approximately $n^2 / 2$, where $n$ is the number of peptides from a database. Since the search space grows quadratically as the number of peptides in a database increases, the database size is critical to search performance. The theoretical MS/MS spectrum of a candidate peptide-pair $(\alpha-\beta)$ can be generated as a multiplexed spectrum including all fragment ions from modified peptides, $\alpha^+\beta$ and $\beta^+\alpha$, of the two peptides, where each peptide is modified by the mass of the other peptide and the cross-linker on the reactive site (Fig. 5a). Then, theoretical spectra of all candidate peptide-pairs were compared with an experimental MS/MS spectrum and the best-scoring peptide-pair is assigned as a cross-linked peptide.

The molecule types generated from cross-linking MS are mono-linked (or dead end), loop-linked (a peptide contains a linker spanning two reactive sites), and cross-linked peptides [98]. Mono-linked and loop-linked types involve only single peptide and they can be identified as modified peptides by standard database searches allowing the cross-linker as a variable modification. Although mono-linked and loop-linked types also provide useful information on protein structure, this review will focus on cross-linked types in which identifying peptide-pairs gives rise to significant computational issues.

5.2. Computational methods for cross-linked peptides

In its early days, some efforts have been made to use the existing database search tools designed for linear peptide identification to identify cross-linked peptides. It involves the generation of a chimeric database, XL-DB, containing linear (concatenated together) sequences of all possible peptide pairs
Then, the database search was performed against the XL-DB while allowing the cross-linker as a variable modification. The method works reasonably well, but the scoring schemes for linear peptides were not optimal for cross-linked peptides. The search never considered all possible fragment ion types from cross-linked peptides \((a-b)\) simultaneously, since the linearized sequences, \(ab\) and \(ba\), were separately written in XL-DB and thus were matched to an experimental MS/MS spectrum one at a time during the search, i.e., utilizing only half fragment ions to identify cross-linked peptides. In a sense, the XL-DB approach eliminated the demand for dedicated software for identification of cross-linked peptides, but it faces the quadratic growth problem in the database size, with suboptimal fragment ion matching.

With the developments of open searches (details in MS/MS-based identification of modified residues in covalent labeling section), there also have been efforts to apply open search algorithms. Fig. 5.

**Fig. 5.** a) MS/MS spectrum of a cross-linked peptide \((a-b)\) can conceptually be regarded as multiplexed by the two spectra generated from the two modified peptides, \(\alpha^{\delta}\) and \(\beta^{\delta}\), where each peptide is modified by the mass of the other peptide (although actually two peptides in a cross-linked peptide are simultaneously fragmented). The real spectrum, however, becomes more complex than a simply multiplexed form due to additional ions (\(^{*}\) peaks) originating from double backbone cleavage. This characteristic makes the interpretation very complicated due to ambiguous peak assignments. b) Pseudo code for retrieving candidate peptide-pairs matched with a query mass (i.e., precursor ion mass, PM in the figure). Basically, the computational complexity is \(O(n^2)\). Using the sorted peptide list by their masses, the complexity gets down to \(O(n)\), where peptides that have been checked are never checked again, given a spectrum. For the schematic representation of the algorithm, let us assume a two-dimensional coordinate plane consisting of \(a\)-axis and \(b\)-axis representing peptide masses (right bottom in the figure). The triangles on each axis represent all (linear) peptides from a database at coordinates corresponding to their masses and then the light blue dots on the \(ab\) plane are all possible peptide-pairs \((105\) pairs for 14 peptides in the figure, calculated as \(\binom{14}{2}\)). Given a PM, candidate peptide-pairs are dots within the area satisfying an inequality \(PM-\varepsilon \leq \alpha's\ mass + \beta's\ mass + X \leq PM + \varepsilon\), where PM is precursor mass of a spectrum, \(\varepsilon\) is mass tolerance, and \(X\) is cross-linker mass. \(\alpha\) is assumed as lighter peptide.

[Pseudo code]

\[
\begin{align*}
\alpha's\ index & \leftarrow 1; \quad \beta's\ index \leftarrow \text{size of peptide list}; \\
\text{while} (\alpha's\ index \leq \beta's\ index) & \\
\text{if} (\alpha's\ mass + \beta's\ mass + X > PM + \varepsilon) & \\
\beta's\ index & \leftarrow \beta's\ index - 1; \\
\text{else if} (\alpha's\ mass + \beta's\ mass + X < PM - \varepsilon) & \\
\alpha's\ index & \leftarrow \alpha's\ index + 1; \\
\text{else} & \\
ADD (\alpha - \beta)\ as\ candidate\ pair; & \\
\alpha's\ index & \leftarrow \alpha's\ index + 1; \\
\beta's\ index & \leftarrow \beta's\ index - 1;
\end{align*}
\]

For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.
to the identification of cross-linked peptides [103–107]. The rationale is that a cross-linked peptide may be considered as one peptide with a large variable mass modification (1,000 ~ 4,000 Da) corresponding to the mass of the other peptide, i.e., \( \alpha + \Delta m \) is a modified form of peptide \( \alpha \) with the modification mass of peptide \( \beta \) in the example in Fig. 5a. This open search-based approach, first, identifies a modified peptide \( \alpha + \Delta m \) (m is an arbitrary mass) that best matches an MS/MS spectrum and then, finds the second peptide \( \beta \) corresponding to the mass \( m \). If there were many second peptide candidates within a specified mass tolerance, additional validation is required. This approach, however, finds constituent peptides of a cross-linked peptide separately, and thus cannot interpret an MS/MS spectrum simultaneously with two constituent peptides, similar to XL-DB approach. Several software tools such as CXMS pipeline [103], Protein Prospector [104], Kojak [105], pLink [106], and MetaMorpheus [107] have adopted the open search strategy. CXMS pipeline that employs open search Popitam [108] reduced the search space of open search by considering only MS/MS spectra of high-charge state precursors unidentified from linear peptide search. As a recent automated approach, Kojak [105] finds top 250 modified peptides in the first pass via open search allowing modifications on linkable residues and then searches for cross-linked peptides via pairing the 250 peptides, in which two peptide masses plus the cross-linker mass is equal to the precursor ion mass. pLink [106] used similar strategy to Kojak, but in pairing peptides identified from open search, the heavier peptide in a pair is selected from top 50 peptides while top 250 peptides for the lighter peptide, assuming that the heavier (or longer) peptides would be identified more confidently from open search.

Various tools have been developed to take into account all possible fragment ions of cross-linked peptides at once during search [109–113]. The searches, however, have been limited to a database of only a few proteins or small purified protein complexes due to the huge search space. As an effort to reduce the search space, SIM-XL [112] created a dynamic database that consisted of all possible pairs containing at least one linear peptide identified with a dead-end via a preliminary search. Nevertheless, the underlying \( n \)-square problem originating from all peptide-peptide combinations could not be readily resolved. The expansion of cross-linking studies to a large database such as a complete human proteome has become feasible with the advent of isotopically coded cross-linkers and MS-cleavable cross-linkers [100,114,115]. For isotopically coded cross-linkers, cross-linked peptides by isotopically coded light and heavy linkers are mixed and analyzed together by MS. Isotopic pairs of chemically identical peptides could be recognized owing to the mass difference between light and heavy linkers. In MS/MS spectra of the two isotopic pairs, there are common fragment ions with identical masses, i.e., independent of cross-linkers. The software, xQuest [114], utilized such common ions to fast query candidate peptides to fragment-ion index generated from all peptides in the database. The queried peptides are compared with MS/MS spectra and only the combinations of the best matching peptides are evaluated as candidate cross-linked peptides. On one hand, MS-cleavable cross-linkers contain a preferential cleavage site (or sites) in the spacer arm and thus produces signature fragment ion peaks that indicate the masses of component peptides in a cross-linked peptide. XlinkKX [115] software was designed to take advantage of MS-cleavable cross-linkers. XlinkKX extracts the mass of each component peptide using signature peaks spaced by a unique mass difference (referred to as \( \Delta m \) principle) derived from the cross-linker. If at least one component peptide mass is observed with the \( \Delta m \) principle in an MS/MS spectrum, the spectrum is regarded as generated from a cross-linked peptide and all deduced peptide-pairs are submitted for further peptide sequence analysis. The extraction of component peptide masses reduced the search space from \( O(n^2) \) to \( O(2n) \) by eliminating the combination of all peptides in a database. While xQuest and XlinkKX were designed for specific cross-linkers, they also supported ‘enumeration mode’ that took into account all the peptide pair combinations against a small database. In other approaches, the signature peaks from cross-linkers were utilized for more sophisticated scoring of cross-linked peptides [116]. As a kind of MS-cleavable cross-linker, protein interaction reporter (PIR) containing two labile bonds in the cross-link can be specifically cleaved in situ to release a signature reporter ion and two intact peptide ions. X-links [117] supported identification of PIR-labeled products.

### 5.3. Disulfide bond

A disulfide bond is a post-translational modification that covalently links the sulfur atoms of two cysteine residues in close spatial proximity so that proteins are internally or externally cross-linked. Disulfide bonds are vital for the stabilization of final protein structures, and at the same time known to regulate mediation of various signaling pathways in a cell [118,119]. Thus the determination of their presence and location in proteins can provide invaluable insight into protein folds and functions. Similar to chemical cross-linking MS strategy, information on the disulfide bonds can be obtained by analyzing MS/MS spectra of intact disulfide-linked peptides, resulting from enzymatic digestion of native proteins under non-reducing condition [120,121]. The formation of nonnative disulfide bonds is prevented during sample preparation by controlling temperature, pH, and free cysteine [121]. The nature of disulfide-linked peptides is the same as chemically cross-linked peptides. Thus, algorithms for cross-linked peptides can be equally applied to the identification of disulfide-linked peptides by simply adjusting the linkable residues to Cysteine and the mass of cross-linker to \(-2.01565\) Da of didehydro.

Dedicated software such as DBond [122,123], SlinkS [124] and MassMatrix [125] has been introduced for the analysis of the MS/MS spectra of disulfide-linked peptides. The software tools were designed to recognize diagnostic fragment ion series from disulfide bonds that indicate the presence of disulfide bonds (not detected in linear peptides) and help assess the correctness of all the component peptides. SlinkS [124] utilized the abundance of fragment ions from disulfide bond-specific cleavages as well as peptide backbone fragments under electron transfer higher energy dissociation (ET/HD) condition and improved the identification performance of disulfide-linked peptides. The identification of disulfide-linked peptides requires the enumeration of all combinations of two peptides containing Cys. DBond [122] implemented a fast algorithm for enumeration of peptide combination by sorting all peptides in a database based on their masses (Fig. 5b). In the sorted list, all possible peptide-pairs can be obtained by retrieving the entire list once while keeping the sum of masses of lighter and heavier peptides closest to the query precursors mass. The algorithm could reduce the quadratic search space to a linear search space and DBond worked well with a whole human database without any filtration step. Disulfide-linked peptides with more than one disulfide bond can be analyzed by informatics methods, but their poor fragmentation would make the identification less confident. MassMatrix [125] supported MS/MS-based identification of disulfide-linked peptides with up to 2 disulfide bonds.

### 5.4. Confidence and false positives

The confidence of cross-linked peptide assignments involves the confidence of each component peptides. The assignment of cross-linked peptides (\( \alpha - \beta \)) would be confident if distinguishing fragment ions are evenly observed from both \( \alpha + \Delta m \) and \( \beta + \Delta m \) (i.e., both peptides are identified with very high confidence).
are ambiguous assignments where one peptide is identified with very high confidence but the other identification is poor (if the peptide is very short or if its fragment ions are rarely observed). The complete matching score of cross-linked peptides to MS/MS spectra is unable to distinguish such ambiguous assignments since their scores could be comparable to those in which both peptides are correctly identified. It should be noted that most false positive identifications in cross-linking MS arise from cases where one peptide is correctly identified but the second peptide is incorrect. In particular, the proportion of such cases increases if one peptide is very short [126]. To compensate for the problem, XLinkSearch [127] implemented a data-driven scoring scheme that independently estimates the probability of correctly identifying each individual peptide given knowledge of the correct or incorrect identification of the other peptide. Protein Prospector [128] reported the uneven fragmentation efficiency of the two component peptides and suggested to use metrics reflecting the quality of the spectral match to the less confident peptide, resulting in the most discriminatory power between correct and incorrect assignments. Machine-learning techniques were also applied to improve the classification. XProphet [129] applied linear discriminant analysis to optimally combine multiple scores so that the combined score maximized the separation of true positive and false positive hits.

False Discovery Rate (FDR) estimation using target-decoy strategy [32] is more complicated for cross-linked peptides than linear peptides. Because two peptides are involved, three labels exist, 1) target-target (TT), 2) target-decoy or decoy-target (TD), and 3) decoy-decoy (DD). False positives could be estimated using those with at least one decoy match. The most commonly used formula for false positives is $\#\text{TD} - \#\text{DD} [130]$. Some corrected formulas have been introduced as the types of cross-linkers (e.g., directional, non-directional and heterobifunctional) [131]. Finally, FDRs were calculated separately for mono-linked, loop-linked, and cross-linked peptides (that can be further separated into intraprotein and interprotein) because their a priori probabilities for matching as well as score distributions are different [129]. The separate FDR calculation showed an improvement in the validation when compared to global FDR calculation [128].

6. Conclusions

MS-based proteomics has been a versatile tool to study many aspects of proteins. Here, we reviewed state-of-the-art MS-based techniques and computational tools for structural analyses of proteins and protein complexes. The peptide-centric structural proteomics have limitations as well as particular advantages. It necessarily includes two procedures to introduce modifications into native proteins and digest the modified proteins into peptides. First of all, the protein digestion causes major limitations. If sequence coverage of a protein is observed to be low by recovered peptides, a significant amount of structural information is lost. If a peptide sequence is repeated within a protein or complex, it cannot be unambiguously assigned. For protein isoforms, assembly of recovered peptides can hardly determine exact proteoforms. Second, the excessive labeling may cause unwanted structural perturbations of native proteins. Third, the comparison between unmodified and modified samples assumes the same properties of unmodified and modified peptides in MS experiments. The detectability of the two, however, might be different because modifications might prevent enzymatic cleavage or change chromatographic retention time and ionization efficiency, making their quantitative comparison complicated. Finally, unexpected modifications that existed in native proteins or are introduced during MS experiments might hinder the identification of peptides of interest. For the success of MS-based techniques, the experimental conditions should be carefully controlled throughout the experiment and the choice of labeling or cross-linking reagents and processes should be carefully made by considering the distribution of amino acids in the protein sequence. A combination of different reagents or techniques may also be considered to increase the number of restraints if target amino acids for labeling are not sufficient in regions of interest.

MS-based techniques are extending the scope toward endogenous biomolecular or proteome-wide studies. It can be expected that the commensurate advances in data analysis tools will enable large-scale analyses, at the same time reduce false positives, and in the end increase the quality of constraints on protein or network modeling. The recent focus has been on integrative approaches that combine information from different types of experiments to compute structural models. MS-based techniques can work as complements to emerging in vivo and in situ technologies such as live-cell imaging and in-/on-cell NMR. Improvements would be necessary for integrative software to best combine a variety of structural data from different methods.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2017R1E1A101077412, 2019M3E5D3073568). S.N. was supported by BK21 Plus project.

References

[1] Ilari A, Savino C. Protein structure determination by x-ray crystallography. Methods Mol Biol 2008;452:63–87.
[2] Tugarinov V, Hwang PM, Kay LE. Nuclear magnetic resonance spectroscopy of high-molecular-weight proteins. Annu Rev Biochem 2004;73:107–46.
[3] Bai XC, McMullan G, Scheres SH. How cryo-EM is revolutionizing structural biology. Trends Biochem Sci 2015;40:49–57.
[4] Lössl P, van de Waterbeemd M, Heck AJ. The diverse and expanding role of mass spectrometry in structural and molecular biology. EMBO J 2016;35:2634–57.
[5] Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature 2003;422:198–207.
[6] Ong SE, Mann M. Mass spectrometry-based proteomics turns quantitative. Nat Chem Biol 2005;1:252–62.
[7] Mann M, Jensen ON. Proteomic analysis of post-translational modifications. Nat Biotechnol 2003;21:255–61.
[8] Hebert AS, Richards AI, Bailey DJ, Ullrich A, Coughlin EE, Westphall MS, et al. The one hour yeast proteome. Mol Cell Proteomics 2014;13:339–47.
[9] Geiger T, Wehner A, Schaab C, Cox J, Mann M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol Cell Proteomics 2012;11(M111):014050.
[10] Hurtlin EL, Bruckner RJ, Paulo JA, Cannon JR, Ting L, Baltzer K, et al. Architecture of the human interactome defines protein communities and disease networks. Nature 2017;545:505–9.
[11] Herzog F, Kahraman A, Boehringer D, Mak R, Bracher A, Walzthoeni T, et al. Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry. Science 2012;337:1348–52.
[12] Greber BJ, Boehringer D, Leibundgut M, Bieri P, Leitner A, Walzthoeni T, et al. The complete structure of the large subunit of the mammalian mitochondrial ribosome. Nature 2014;515:283–6.
[13] Staals RH, Zhu Y, Taylor DW, Kornfeld JF, Sharma K, Barendregt D, et al. RNA targeting by the type III-A CRISPR-Cas Ssm complex of Thermus thermophilus. Mol Cell 2014;56:518–36.
[14] Li J, Wei H, Krystek Jr SR, Bond D, Brender TM, Cohen D, et al. Mapping the energetic epitope of an Antibody/Interleukin-23 interaction with hydrogen/deuterium exchange, fast photochemical oxidation of proteins mass spectrometry, and alanine shake mutagenesis. Anal Chem 2017;89:2250–8.
[15] Limpiikrit P, Hale JE, Hazelbaker M, Huang Y, Jia Z, Yazdani M, et al. Covalent labeling and mass spectrometry reveal subtle higher order structural changes for antibody therapeutics. MAbs 2019;11:463–76.
[75] Dasari S, Chambers MC, Sileos RJ, Zimmerman LJ, Ham AJ, Taub DL, TagRecon: high-throughput mutation identification through sequence tagging. J Proteome Res 2010;9:1716–26.

[76] Na S, Jeong J, Park H, Lee KJ, Paek E. Unrestrictive identification of multiple post-translational modifications from tandem mass spectrometry using an error-tolerant algorithm based on an extended sequence tag approach. Mol Cell Proteomics 2012;11:152–63.

[77] Na S, Banderia N, Paek E. Fast multi-blind modification search through tandem mass spectrometry. Mol Cell Proteomics 2012;11(M11):011999.

[78] Savitski MM, Nielsen ML, Zuberbauer RA. ModiF Comb, a new proteomic tool for mining multiplex post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. Mol Cell Proteomics 2006;5:935–48.

[79] Banderia N, Tsur D, Frank A, Pevzner PA. Protein identification by spectral networks analysis. Proc Natl Acad Sci U S A 2007;104:6140–5.

[80] Falkner JA, Falkner JW, Yocum AK, Andrews PC. A spectral clustering approach to MS/MS identification of post-translational modifications. J Proteome Res 2008;7:4614–22.

[81] Na S, Payne SH, Banderia N. Multi-species identification of polymorphic peptide variants via propagation in spectral networks. Mol Cell Proteomics 2016;15:3501–12.

[82] David M, Ferlin G, Rogniaux H, Tessier D. SpecOMS: a full open modification search method performing all-to-all spectra comparisons within minutes. J Proteome Res 2017;16:3030–8.

[83] Ye D, Fu Y, Sun RX, Wang HF, Yuan ZF, Chi H, et al. Open MS/MS spectral library search to identify unannotated post-translational modifications and increase spectral identification rate. Bioinformatics 2010;26:399–406.

[84] Ma CW, Lam WA. Hunting for unexpected post-translational modifications by spectral library searching with tier-wise scoring. J Proteome Res 2014;13:2262–71.

[85] Bullo JC, Mirudhin YA, Tchekhovskoi DV, Markey SP, Heidbrink Thompson J, Park JK, Paek E. The Hybrid Shadow: a mass spectral library search method for discovery of modifications in proteomics. J Proteome Res 2017;16:1924–35.

[86] Chick JM, Kolippayan D, Nusinow DF, Zhai B, Rad R, Huang TL, et al. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. Nat Biotechnol 2015;33:743–9.

[87] Balbian RC, Hui X, Yu Y, Cai Y, Goldberg D. Lookup peaks: a hybrid of de novo sequencing and tolerant database search for protein identification by tandem mass spectrometry. Anal Chem 2017;89:3829–33.

[88] Chick JM, Kolippayan D, Nusinow DF, Zhai B, Rad R, Huang TL, et al. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. Nat Biotechnol 2015;33:743–9.

[89] Balbian RC, Hui X, Yu Y, Cai Y, Goldberg D. Lookup peaks: a hybrid of de novo sequencing and tolerant database search for protein identification by tandem mass spectrometry. Anal Chem 2017;89:3829–33.

[90] Singh P, Shaffer SA, Schier A, Holman C, Pfeiffer LR, Larson Freeman TJ, et al. Characterization of protein cross-links via mass spectrometry and an open-modification search strategy. Anal Chem 2008;80:8799–806.

[91] Chuah H, Baker PR, Bunting LM, Chalkley RJ. Finding chimeras: a bioinformatics strategy for identification of cross-linked peptides. Mol Cell Proteomics 2010;9:25–31.

[92] Hoopmann MR, Zelter A, Johnson RS, Riffle M, MacCoss MJ, Davis TN, et al. Kojak: efficient analysis of chemically cross-linked protein complexes. J Proteome Res 2015;14:2190–8.

[93] Yang B, Wu YJ, Zhu M, Fan SB, Lin J, Zhang K, et al. Identification of cross-linked peptides from complex samples. Nat Methods 2012;9:994–6.

[94] Lu L, Milikin RJ, Solonkew ST, Rolf S, Sald M, Shortreed MR, et al. Identification of MS-cleaveable and noncleavable chemically cross-linked peptides with MetaMorph, J Proteome Res 2018;17:2170–6.

[95] Hernandez P, Gras R, Frey J, Appel RD. Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. Proteomics 2003;3:870–8.

[96] Lee YJ, Lackner LL, Nunamji J, Phinnney BS. Shotgun cross-linking analysis for studying quaternary and tertiary protein structures. J Proteome Res 2007;6:3908–17.

[97] McElwain 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;9:2488–95.

[98] Du X, Choudhury SM, Manes NP, Wu S, Mayer MU, Adkins JN, et al. Xlink-identifier: an automated data analysis platform for confident identifications of chemically cross-linked peptides using tandem mass spectrometry. J Proteome Res 2012;11:3112–26.

[99] Lima DB, de Lima TB, Baibausa TS, Neves-Ferreira AGC, Barbosa VC, Gzzo FC, et al. SIM-XL: A powerful and user-friendly tool for peptide cross-linking analysis. J Proteomics 2015;129:51–5.

[100] Vilma S, Dropper F, Hulstaert N, Cerià M, Gevaert K, Economo A, et al. Xilinx: A new approach toward the identification of cross-linked peptides. Anal Chem 2016;88:9948–57.

[101] Rinner O, Seebacher J, Walzthoeni T, Mueller LN, Beck M, Schmidt A, et al. Identification of cross-linked peptides from large sequence databases. Nat Methods 2008;5:315–8.

[102] Liu F, Rijkers DT, Post H, Heck AJ. Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. Nat Methods 2015;12:1799–804.

[103] Götte M, Pettelkau J, Ritschke R, Bihlg CH, Schäfer M, Sinz A. Automated assignment of MS/MS cleavable cross-links in protein 3D-structure analysis J Am Soc Mass Spectrom 2015;26:83–97.

[104] Anderson GA, Tolic N, Tsur D, Frank A, Pevzner PA. Protein identification by spectral networks analysis. Proc Natl Acad Sci U S A 2007;104:6140–5.

[105] Bulleid NJ, Ellgaard L. Multiple ways to make disulfides. Trends Biochem Sci 2007;79:1393–400.

[106] Erickson MS, Groll M, Riezman H, Ludwig M. Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching. Mol Cell Proteomics 2007;6:2452–63.

[107] Erickson MS, Groll M, Riezman H, Ludwig M. Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching. Mol Cell Proteomics 2007;6:2452–63.

[108] Erickson MS, Groll M, Riezman H, Ludwig M. Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching. Mol Cell Proteomics 2007;6:2452–63.

[109] Erickson MS, Groll M, Riezman H, Ludwig M. Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching. Mol Cell Proteomics 2007;6:2452–63.

[110] Erickson MS, Groll M, Riezman H, Ludwig M. Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching. Mol Cell Proteomics 2007;6:2452–63.

[111] Erickson MS, Groll M, Riezman H, Ludwig M. Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching. Mol Cell Proteomics 2007;6:2452–63.

[112] Erickson MS, Groll M, Riezman H, Ludwig M. Structural analysis of multiprotein complexes by cross-linking, mass spectrometry, and database searching. Mol Cell Proteomics 2007;6:2452–63.