Bioinformatics tools for the structural elucidation of multi-subunit protein complexes by mass spectrometric analysis of protein–protein cross-links

Shannon L. N. Mayne and Hugh-George Patterton

Submitted: 17th September 2010; Received (in revised form): 7th December 2010

Abstract

Multi-subunit protein complexes are involved in many essential biochemical processes including signal transduction, protein synthesis, RNA synthesis, DNA replication and protein degradation. An accurate description of the relative structural arrangement of the constituent subunits in such complexes is crucial for an understanding of the molecular mechanism of the complex as a whole. Many complexes, however, lie in the mega-Dalton range, and are not amenable to X-ray crystallographic or nuclear magnetic resonance analysis. Techniques that are suited to structural studies of such large complexes, such as cryo-electron microscopy, do not provide the resolution required for a mechanistic insight. Mass spectrometry (MS) has increasingly been applied to identify the residues that are involved in chemical cross-links in compound protein assemblies, and have provided valuable insight into the molecular arrangement, orientation and contact surfaces of subunits within such large complexes. This approach is known as MS3D, and involves the MS analysis of cross-linked di-peptides following the enzymatic cleavage of a chemically cross-linked complex. A major challenge of this approach is the identification of the cross-linked di-peptides in a composite mixture of peptides, as well as the identification of the residues involved in the cross-link. These analyses require bioinformatics tools with capabilities beyond that of general, MS-based proteomics analysis software. Many MS3D software tools have appeared, often designed for very specific experimental methods. Here, we provide a review of all major MS3D bioinformatics programmes, reviewing their applicability to different workflows, specific experimental requirements and the computational approach taken by each.

Keywords: MS3D; protein complex; structural biology; interactomics; modelling

INTRODUCTION

Supra-molecular protein complexes are involved in numerous fundamental biochemical processes including catalysis, protein secretion, nuclear transport, protein degradation, protein folding, gene regulation, RNA synthesis, protein synthesis, signal transduction, chromosome segregation and in DNA replication and repair. A mechanistic understanding of these composite protein assemblies requires an insight into the molecular arrangement and interactions of the subunits. A structural insight into the arrangement of the components in such complexes is, however, still limited. Traditional methods of protein structure determination such as X-ray crystallography (XRC) and nuclear magnetic resonance (NMR) have technical limitations, restricting the size of the protein complex that can be crystallized or the resolution at which large structures can be interpreted. Cryo-electron microscopy holds great promise for the structural elucidation of...
mega-Dalton protein complexes, but the resolution is currently insufficient for a detailed structural analysis [1].

Recently the application of mass spectrometry (MS) to identify the positions of chemical cross-links between the protein subunits of complexes has significantly advanced our understanding of the arrangement and interaction surfaces involved in mega-Dalton protein complexes [2]. If the structures of the components are known, the location of the cross-links allows one to very precisely place each subunit in the correct orientation within the complex. The use of a range of cross-linking reagents, each with a specific atomic reach, has allowed the further refinement of models of quaternary protein structures. This approach, termed ‘MS3D’, has developed into a powerful technique for the structure elucidation of multi-subunit protein complexes.

Many software packages are available for standard protein identification or de novo sequencing analyses by MS [3]. Software for the analysis of MS and MS/MS spectra to identify cross-linked peptides and the sequence positions of the residues involved in the cross-link has been developed for very specific experimental methodologies. In this review, we provide an overview of software currently available for MS3D analyses, noting the application niches of programmes as well as giving a detailed comparison of their functionalities. This gives an overview of the current software landscape as well as facilitating selection of the most suitable tool for a particular MS3D application. For the basic mass spectrometric concepts of this technique, we refer the reader to a number of recent reviews [4–6].

**MS3D data analysis**

In a typical MS3D experiment a protein complex is isolated or reconstituted and then treated with a cross-linking reagent. A wide variety of cross-linking reagents are commercially available, with homo- and bi-functional reactive groups targeting a specific or a narrow range of residue side-chains [7]. Some of these reagents allow subsequent cleavage or affinity purification [8]. Following cross-linking, the complex is digested with a sequence-specific protease such as trypsin [9] or endo gluC [10]. This yields a mixture of (possibly multiply) cross-linked, singly linked (‘dead-end’), intra-linked and uncross-linked peptides (Figure 1).

Cross-linked peptides are typically identified by MS, often followed by confirmation with MS/MS. The residues involved in the cross-link in the identified peptides are usually pinpointed by MS/MS. Software used for the MS-based structural elucidation of cross-linked protein complexes normally perform four steps: (i) detection of the cross-linked peptides; (ii) identification of the cross-linked peptides; (iii) identification of cross-linked residues in the di-peptide; and (iv) interpretation of cross-linked data in terms of spatial proximities and subsequent refinement of the structural model (Figure 2). We discuss each of these four steps individually, mentioning the different experimental routes that have been reported, and indicate the applicability of programmes to each of the various approaches. A summary of the abilities of each programme is presented in Table 1. No single software programme is currently available that can perform all four tasks. In particular, the interpretation of spatial constraints and model refinement still require significant manual input. The reader should also take note of the very similar naming of X!Link, X-Link, X-Links, XLINK, Links/MS2Links and SearchXLinks. These are all distinct programmes.

**Detection of cross-linked peptides**

The first activity in a MS3D experiment is the chemical cross-linking of the proteins in the biological complex [7, 11–13]. The yield of chemically cross-linked peptides under conditions that conserve the structural integrity of a biological complex is often very low [8]...
and the detection of the di-peptides in a complex mass spectrum can, therefore, be technically challenging. Four approaches have been reported to simplify the identification of the cross-linked peptide peaks in the mass spectrum (Step 1, Figure 2): comparison to a non-cross-linked control, using isotopically labelled di-peptides, identification of post-fragmentation reporter ions and the identification of peaks that match the theoretical mass of one of the possible peptide dimer combinations that can be formed from the known proteins in the complex.

**Non-cross-linked controls**

In the simplest approach, peaks that are present in the MS spectrum of a cross-linked sample and absent in an uncross-linked control sample are flagged as putative cross-linked di-peptides by CLPM [14], PROWL’s PeptideMap [15] and SearchXLinks [16]. In a variation of this technique, CrossSearch [17] identifies only inter-molecular cross-linked di-peptides by analysing peptides from the cross-linked complex as well as the two subunits cross-linked individually, which requires analysis of at least three cross-linked samples.

**Isotope labelling**

The isotopic labelling of cross-linked peptides is achieved by using a cross-linking reagent that was synthesized using compounds that contained heavy or light isotopes [18], the introduction of $^{18}$O from $\text{H}_2^{18}\text{O}$ during proteolytic hydrolysis of the proteins [19], or by the cross-linking of mixed isotope samples (MIX), typically prepared from cells that were cultured in media that contained $^{14}\text{N}$ or $^{15}\text{N}$ labelled amino acids [20].

Cross-linking with a mixture of heavy and light cross-linking reagent will produce peak pairs or doublets in the MS spectrum that are offset by the mass difference between the two isotopically labelled reagents, and with intensities that reflect the ratio of heavy:light reagent. Thus, cross-linked peptides can easily be found by scanning the mass spectrum for these peak pairs with a specific mass difference. GPMAW [21, 22], iXLINK [23] and xQUEST [24] can scan MS spectra to identify peak doublets and potential cross-linked di-peptides. In GPMAW and iXLINK, a custom mass difference in the peak pair of up to 8 Da can be selected. iXLINK can also identify single peptides that contain ‘dead-end’ links by detecting peak doublets that appear for each of the two isotopically labelled cross-linking reagents.
Table 1: The characteristics, experimental requirements and computational approaches of different MS3D bioinformatics tools

<table>
<thead>
<tr>
<th>Software</th>
<th>References</th>
<th>Sample preparation</th>
<th>Detect X-linked peptide</th>
<th>Seq of X-linked peptide</th>
<th>ID of X-linked residues</th>
<th>Distance limits</th>
<th>Struct. interpretation</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASAP and MS2Assign</td>
<td>[33]</td>
<td>None</td>
<td>None</td>
<td>MS and MS/MS</td>
<td>Yes</td>
<td>No</td>
<td>None</td>
<td>NP</td>
</tr>
<tr>
<td>CLPM</td>
<td>[14]</td>
<td>Non-X-linked control</td>
<td>Type ID MS</td>
<td>MS</td>
<td>No</td>
<td>No</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>CrossSearch</td>
<td>[17]</td>
<td>Single-protein X-linked control</td>
<td>Type ID MS</td>
<td>MS and MS/MS</td>
<td>Yes</td>
<td>No</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Crux</td>
<td>[37]</td>
<td>None</td>
<td>None</td>
<td>MS and MS/MS</td>
<td>Yes</td>
<td>No</td>
<td>None</td>
<td>P</td>
</tr>
<tr>
<td>Links and MS2Link</td>
<td>[40]</td>
<td>None</td>
<td>None</td>
<td>Top down MS/MS</td>
<td>Yes</td>
<td>No</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>MassMatrix</td>
<td>[47]</td>
<td>None</td>
<td>None</td>
<td>MS and MS/MS</td>
<td>Yes</td>
<td>No</td>
<td>None</td>
<td>P</td>
</tr>
<tr>
<td>MS2PRO</td>
<td>[41]</td>
<td>None</td>
<td>None</td>
<td>Top down MS/MS</td>
<td>Yes</td>
<td>No</td>
<td>None</td>
<td>NP</td>
</tr>
<tr>
<td>MS-Bridge</td>
<td>[34]</td>
<td>None</td>
<td>None</td>
<td>MS</td>
<td>No</td>
<td>No</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>MSX-3D</td>
<td>[35]</td>
<td>None</td>
<td>None</td>
<td>MS</td>
<td>No</td>
<td>Yes</td>
<td>Validation</td>
<td>None</td>
</tr>
<tr>
<td>PeptideMap (in PROWL)</td>
<td>[15]</td>
<td>Non-X-linked control</td>
<td>Type ID MS</td>
<td>MS</td>
<td>No</td>
<td>No</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Pro-Crosslink</td>
<td>[27]</td>
<td>H$_2^18$O isotope labelling</td>
<td>Type ID MS</td>
<td>MS and MS/MS</td>
<td>Yes</td>
<td>No</td>
<td>None</td>
<td>NP</td>
</tr>
<tr>
<td>ProteinXXX/GPMAW</td>
<td>[21, 22]</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>No</td>
<td>No</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>SearchXLinks</td>
<td>[16]</td>
<td>Non-x-linked control</td>
<td>None</td>
<td>MS/PSD</td>
<td>Yes</td>
<td>No</td>
<td>None</td>
<td>NP</td>
</tr>
<tr>
<td>VIRTUAL-MSLAB</td>
<td>[36]</td>
<td>None</td>
<td>None</td>
<td>MS</td>
<td>No</td>
<td>No</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>X-Link</td>
<td>[39]</td>
<td>None</td>
<td>None</td>
<td>MS</td>
<td>Yes</td>
<td>No</td>
<td>None</td>
<td>NP</td>
</tr>
<tr>
<td>X-Link</td>
<td>[28]</td>
<td>MIX isotope labelling</td>
<td>None</td>
<td>MS</td>
<td>No</td>
<td>No</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>XLINK (iXLINK and doXLINK)</td>
<td>[23]</td>
<td>H$_2^18$O and $^3$H isotope labelling; only NHS</td>
<td>Type ID MS</td>
<td>MS and MS/MS</td>
<td>Yes</td>
<td>No</td>
<td>None</td>
<td>P</td>
</tr>
<tr>
<td>X-Links</td>
<td>[31]</td>
<td>None</td>
<td>None</td>
<td>Type ID MS/MS/MS</td>
<td>Yes</td>
<td>No</td>
<td>None</td>
<td>NP</td>
</tr>
<tr>
<td>xQUEST</td>
<td>[24]</td>
<td>$^3$H isotope labelling</td>
<td>None</td>
<td>Type ID MS/MS/MS</td>
<td>Yes</td>
<td>No</td>
<td>None</td>
<td>NP and P</td>
</tr>
</tbody>
</table>

*Entries indicate the requirement of software for any specific cross-linking reagent, labelling method or control sample. MIX: mixed isotope samples [13]; NHS: N-hydroxysuccinimide based cross-linking reagents.

*Entries indicate whether the software can identify a cross-linked peptide in the MS spectrum. ‘Type ID’ identifies software that is capable of discriminating between different cross-link types. The type of experimental data that the software requires to identify the sequences of the cross-linked peptides. The software provides any limit on the maximum distances between cross-linked residues. The software performs any interpretation of the cross-link data in terms of the structure of compound protein assemblies. P, probabilistic, a statistical probability. NP: non-probabilistic, a score relative to a threshold value.
following hydrolysis of the single unreacted functional group in a H$_2^{16}$O/H$_2^{18}$O mixture.

Trypsin incorporates two oxygen atoms from two water molecules at each carboxyl terminal during hydrolysis [25]. If proteolytic cleavage is performed separately in H$_2^{16}$O and in H$_2^{18}$O, and the samples combined before MS analysis, an 8 Da mass differences will be visible in peak pairs of peptide dimers [26]. Doublets separated by 8 Da are highlighted by DetectShift in Pro-Crosslink [27]. Peptide dimers that contain one or both of the original C-termini of the protein will not be labelled by $^{18}$O, and will thus not be flagged in this method.

In the case of MIX samples, proteins purified from cells grown in the presence of $^{14}$N and $^{15}$N labelled amino acids are combined in equal parts. Intra-molecular cross-linked peptides will be detected as peak doublets composed of N$^{15}$/N$^{15}$ and N$^{14}$/N$^{14}$ dimers. Inter-molecular dimers, on the other hand, will be observed as peak triplets composed of N$^{15}$/N$^{15}$, N$^{14}$/N$^{15}$ and N$^{14}$/N$^{14}$ linked peptides. X-Link [28] uses these isotope signatures to detect inter-molecularly cross-linked peptides.

**Post-fragmentation reporter ions**

This approach requires the use of either a disulphide or a special type of cross-linker, termed a ‘protein interaction reporter’ (PIR) [29]. In the case of disulphide bonds, in-source decay often results in the loss of (H$_2$ + H$^+$) [30]. SearchXLinks identifies peptides linked by a disulphide bridge by searching for a group of three peaks, where the combined mass of two peaks equals that of the third, minus the mass of (H$_2$ + H$^+$) [16]. X-Links [31] is specifically designed to use with a PIR cross-linking reagent [29]. PIR linkers fragment during collision induced dissociation (CID) in a defined fashion, releasing a signature reporter ion and the modified single peptides. Where X-Links detects the PIR-derived reporter ion in an MS/MS spectrum, the spectrum is further scrutinized to identify two peaks where the combined mass plus that of the reporter ion matches a peak in the MS spectrum, which is then flagged as a cross-linked peptide. This approach was successfully used to identify interaction partners and interaction sites in vivo [32].

**Matching peaks to a library of possible peptide dimers**

Where the proteins in a complex are known, many programmes follow the route of creating a library of all possible cross-linked peptides based on the specificity of the proteolytic enzyme selected, the chemical composition of the cross-linking reagent, the reactive amino acid residue, and the allowed post-translational modifications selected by the user. Peaks in the experimental MS spectra that match the mass of entries in this library are flagged as possible cross-linked di-peptides. The programs ASAP [33], CLPM, MS-Bridge [34], MSX-3D [35], PROWL’s PeptideMap, VIRTUALMSLAB [36] and X-Link generate a list of such matched peaks. Some programmes can make use of data from subsequent MS/MS analyses of flagged peptides to confirm the presence of the di-peptide (Crux [37], MS2Assign [38], SearchXLinks, Pro-Crosslink, iXLINK, X-Links, GPMAP, XLInk [39], CrossSearch and xQUEST). Programs such as MS2Links [40] and MS2PRO [41] that are used in a top-down proteomics approach omit the initial MS step, using only data from the MS/MS analysis.

Several groups [42–44] have used existing peptide search engines intended for single peptides, such as MASCOT [45] or X!Tandem [46], to match the experimental spectra of cross-linked peptides.

**Generating the library of peptide dimers**

The degree to which the user can customize the theoretical library of possible cross-linked peptides, in terms of the allowed post-translational modifications or the chemical composition and residue specificity of the cross-linking reagent, for example, differ between software packages (summarized in Supplementary Table S1). This limits the type of experimental data that each software tool can analyse. SearchXLinks does not support any user-defined modification types. ASAP, CLPM, MS2Assign, MS2Links, MSX-3D, VIRTUALMSLAB, SearchXLinks, iXLINK and xQUEST allow a limited number of post-translational modifications. CLPM permits a maximum of 10 custom modification types. MS2Assign, MS2Links, MSX-3D, VIRTUALMSLAB, iXLINK and xQUEST allow a limited number of post-translational modifications. CLPM permits a maximum of 10 custom modification types. MS2Assign, MS2Links, MSX-3D, PROWL’s PeptideMap and Pro-Crosslink allow the user to specify custom cross-linking reagents and reactive amino acid side-chains. As a novel ability, XLINK also provides for custom amino acids, but was developed solely for the amine specific, CID-cleavable PIR cross-linker. MassMatrix, ProteinProspector’s MS-Bridge and PROWL’s
PeptideMap are applicable only to disulphide bridge cross-links. Crux, Pro-Crosslink, X-Link and X!Link assume the use of trypsin as a protease, and does not allow selection of a different cleavage enzyme. CLPM, GPMAW, PROWL’s PeptideMap, X-Links and iXLINK, on the other hand, support any protease with a defined target sequence. As a further refinement, PROWL’s PeptideMap will allow proteases that cleave the bond on the N-terminal as well as C-terminal side of the residue recognized by the protease (e.g. Thermolysin [48]).

An important consideration when comparing experimental MS spectra of cross-linked peptides to a library of all possible pairs is the presence of various combinations of cross-linked species in the mass spectrum (see Figure 1). Crux, MS2Assign, X-Links and iXLINK can accommodate different classes of cross-linked peptides. CrossSearch, GPMAW, X-Link and X!Link will flag dead-end linkers, while CLPM and VIRTUALMSLAB also allow identification of intra-molecular cross-links.

In many studies reported to date cross-linking was carried out on complexes formed by only two proteins with known sequences [49–52]. Crux, MassMatrix, MSX-3D, VIRTUALMSLAB, X-Links and xQUEST will accept any number of protein subunits in the studied complex. Other programs accept only one (ASAP, PROWL’s PeptideMap, MS2Links and MS2PRO) or two (CLPM, CrossSearch, GPMAW, Pro-Crosslink, X-Link, X!Link) proteins. Nucleic acids also form part of many protein complexes such as in chromatin, ribosomes and snRNPs. The theoretical libraries generated by CLPM and MS2Links can include both protein and nucleic acid sequences.

Matching experimental peaks to the theoretical library
All the programmes reviewed here will score two peaks as a match when the $m/z$ values of an observed peptide peak and a theoretical peak are within a specified range. Some programmes match the experimental peaks to the theoretical library, ignoring absent experimental peaks (ASAP, CLPM, Crux, Pro-Crosslink, VIRTUALMSLAB, X-Links, iXLINK and xQUEST) [14, 23, 24, 27, 31, 33, 36, 37] while others match each theoretical peak to the list of experimental peaks (MS2Assign, MSX-3D, SearchXLinks, X-Link, X!Link) [16, 28, 35, 38, 39]. Software can search for either both average and mono-isotopic masses (ASAP, GPMAW, MS2Assign and MSX-3D) or only for mono-isotopic masses (CLPM, CrossSearch, MS2Links, Pro-Crosslink, SearchXLinks, VIRTUALMSLAB, iXLINK and xQUEST). In the experience of the authors, the implementation of the different peak matching methodologies did not translate to significant performance differences.

The identity of the peptides in the cross-linked di-peptide is derived from the highest scoring theoretical peptide in the theoretical library (Step 2, Figure 2). Programs such as IdentifyXLink in Pro-Crosslink, MassMatrix, MS2Assign, MS2Links, SearchXLinks, X!Link, X!LINK and xQUEST allow additional MS/MS verification of the peptides involved in the cross-link.

Identification of the cross-linked residues in the di-peptide
After identification of the peptides in the cross-linked dimer, the residue that is cross-linked must be identified (Step 3, Figure 2). In the case where only a single residue in each peptide is reactive towards the cross-linking reagent, the problem is trivial. However, if a greater number of cross-linkable residues are present in a di-peptide, the problem requires further analysis. Currently available software achieves this by comparison of the di-peptide product ion scan to fragment libraries, each generated according to preset fragmentation rules from the putative di-peptide in a particular cross-linker configuration (reviewed in [53]). No programme has implemented de novo sequencing as an approach, although xQUEST identifies uncross-linked peptides in this way.

Generating an MS/MS fragment library
Many programs (CrossSearch, Crux, GPMAW, MassMatrix, MS2Assign, MS2Links, MS2PRO, Pro-Crosslink, SearchXLinks, X!Link, X!LINK and xQUEST) use defined models of peptide fragmentation [54–66] (reviewed in [53, 67]) to perform in silico fragmentation of the identified di-peptide precursor ion. In this way, the matched entries in the theoretical library of di-peptides are expanded to also include the product ion fragments for each entry. In the case of multiple possible cross-linked combinations for a given di-peptide, a different theoretical fragmentation spectrum is generated for each possibility. The fundamental chemistry behind peptide fragmentation is not yet fully understood, but a
well-defined set of empirical fragmentation rules is known [53]. All programmes utilize a fixed set of such rules, although MassMatrix allows the definition of additional custom fragmentation rules [47].

Most programmes calculate all possible ions resulting from a single fragmentation on the backbone of putative di-peptide precursors, that is the a, b, c, x, y and z ion series (GPMAW, MS2Assign, MS2Links, Pro-Crosslink and SearchXLinks). In others, the theoretical fragmentation library is constrained to the more abundant b and y product ions of these di-peptide precursors (Crux, MS2PRO, iXLINK, X!Link and xQUEST in ion-tag mode). Many programmes also model a variety of additional fragment ion types. MS2Assign, MS2Links, MS2PRO and Pro-Crosslink include a search for the immo- nium ions for the amino acids H, M, W, Y and F. GPMAW considers NH3 and H2O loss, whereas Crux, MS2Assign, MS2Links, Pro-Crosslink and doXLINK also include CO and, with the exception of Crux, CO2 loss. Only SearchXLinks and PROWL’s PeptideMap incorporate the proline effect, where breakage of the bond on the C-terminal side of P is typically not observed. Pro-Crosslink and X!Link model double fragmentation of a peptide, and the top-down methodology programs MS2Links and MS2PRO support a greater number of successive fragmentation steps. While CID fragmentation is the norm for MS3D software, SearchXLinks also incorporates in-source decay (ISD) and post-source decay (PSD), while MS2Links incorporates sustained off-resonance irradiation CID (SORI-CID).

Matching MS/MS spectra

Confirming the di-peptide identity and the cross-linking configuration that produced an experimental MS/MS spectrum involves the comparison of the experimental spectrum to the predicted spectra from a library of possible precursor ions (Step 3, Figure 2). However, predicted and observed spectra are seldom a perfect match. Many predicted products may be present in the experimental spectrum, but of such a low intensity as to go unobserved. Experimental spectra may also contain contaminants absent in the predicted spectrum.

To improve the efficiency of such comparisons, noise peaks can be excluded beforehand. While most instrument platforms are able to do this, some MS3D packages (MS2Assign, MS2Links, Pro-Crosslink and xQuest) incorporate pre-comparison noise filtering.

xQUEST slides a window of m/z 1000 across each spectrum, including only the 250 most intense peaks within each window. The simplest method is the inclusion of peaks that are above an absolute (MS2Assign, Pro-Crosslink) or relative (MS2Links, Pro-Crosslink) intensity threshold. Though not automated, X-Links displays indicators of spectrum quality and allows users to manually exclude peaks. A number of stand-alone spectra filtering packages such as Decon2LS [68] can also be used.

The method used to score the closeness-of-fit between the spectra is of critical importance to maximize true positive matches. A score can reflect a formal, statistical probability of a match with a certain confidence level (probabilistic), or it can be a value on an arbitrary, unbounded scale, with a minimum threshold required to qualify as an acceptable match (non-probabilistic). Some programmes combine the two forms of scoring. MassMatrix calculates both a non-probabilistic and two probabilistic scores [47, 69]. xQUEST makes use of a non-probabilistic function that includes a probabilistic term [24]. Crux converts an initially non-probabilistic score into a probability estimate [37]. The scores of true positive di-peptide assignments tend to be significantly lower than those of identified single peptides. In fact, the MassMatrix user manual recommends a probability threshold of ~0.2 for the assignment of di-peptides [47].

Non-probabilistic scoring

Non-probabilistic scoring of MS/MS spectrum matches is implemented in Crux, MS2Assign, MS2PRO, SearchXLinks, Pro-Crosslink, X-Links, X!Link and xQUEST. Most programmes calculate these scores using simple scoring functions such as the number of theoretical fragments assigned (SearchXLinks), the number of experimental peaks successfully assigned (MS2Assign, SearchXLinks, X!Link), the percentage of peaks assigned (Pro-Crosslink), the number of peaks assigned, normalized to the number of amino acids in the precursor di-peptide (X!Link), or the sum of the intensities of all assigned peaks (SearchXLinks). Thus the score of MS2Assign, MS2PRO, Pro-Crosslink and X!Link is directly proportional to the number of assigned peaks. SearchXLinks uses each equation indicated above as well as similar equations for specific fragments as terms in its scoring function. An additional term for the number of matches assigned to consecutive ions in the b or y fragment series can cause the
SearchXLinks score to scale exponentially against the number of matches. The user can customize the scoring function by specifying a weighting for each term. X-Links uses the uniqueness of a mass within its theoretical library as a match score. XQUEST uses an initial filter that considers only the single-chain b and y ions lacking a cross-linkable site, followed by a more complex scoring scheme that involves a cross correlation function (first introduced in SEQUEST [70]), the percentage of ion intensities in the matched spectrum that is contributed by matched peaks, and a term equal to the negative log10 of the probability of a random match. Crux uses a similar, normalized cross-correlation function.

Probabilistic scoring
Three distinct approaches have been implemented in iXLINK, MassMatrix and Crux to derive a score related to a statistical probability. iXLINK employs a Bayesian scoring scheme based solely on established fragmentation principles that are implemented as nested probability functions [71]. These consider factors such as the consistency of deduced amino acid composition with observed immonium ions, the number and mass deviation (normalized to instrument accuracy) of matched b and y ions, the probability that unmatched peaks are noise, as well as the fraction of complementary (b–y pair) and contiguous (bₙ and bₙ₊₁) assignments in a fragmentation series.

MassMatrix performs hypothesis testing on MS/MS spectrum matches, based on a probability distribution for random matching that can be binomial or non-parametric, which is estimated from the experimental data [69]. Crux presumes a Weibull distribution for the probability of a random match, and estimates a probability score by fitting the non-probabilistic score to this distribution [37].

Structure modelling
To date, no MS3D analysis software seamlessly integrates the identification of cross-linked peptides and residues with the verification or refinement of a structural model (Step 4, Figure 2). In many structural studies that reported MS analysis of chemically cross-linked proteins, distance constraints revealed by the position and reach of the cross-linker were entered into separate modelling programs such as VMD-XPLOR [72], or were directly incorporated into subsequent homology modelling or docking studies [33, 49, 73–75]. MSX-3D [35], however, determines whether published structural models described in a PDB format [76] are consistent with observed peptide cross-links and spatial reach of the chemical cross-linking reagent. MSX-3D also visualizes the model and the spatial reach of the selected cross-linking reagent in a web applet.

Data input and output
In line with trends in the biological sciences towards high-throughput methodologies, MS3D studies also require the analysis of ever larger data sets. This is not supported in many of the older programs where data needs to be manually copied and pasted (CrossSearch, MS-Bridge and PeptideMap), or where only one spectrum can be uploaded at a time (ASAP, CLPM, MS2Assign, MSX-3D, Pro-Crosslink and GPMAW’s ProteinXXX). CLPM, Crux, SearchXLinks, X!Link and XLINK have command line interfaces that allow the scripting of serial analyses in batch. However, with XLINK data files must be manually transferred into and out of a specific working folder for each analysis. Many of the newer MS3D packages, though, claim the handling of high data volume to be an explicit design goal (X!Link, X-Links and xQUEST). Specific data volumes, in terms of the number of spectra, have been demonstrated for Crux (3314 spectra) [37], X!Link (approximately 5000 spectra) [39] and xQUEST (3592 spectra) [24]. However, throughput bottlenecks can also arise with the server. In the case of xQUEST the authors found a variable upper limit to the size of the data file that could successfully be uploaded via the xQUEST web form.

Software release
Many MS3D software packages have been released as web-based services (CLPM, CrossSearch, MS2Assign, ProteinProspector’s MS-Bridge, MSX-3D, PROWLS’s PeptideMapASAP, SearchXLinks and xQUEST) or as packages that can be downloaded (Crux, MassMatrix, Pro-Crosslink, X-Links, XLINK) or requested from the developer (CLPM, VIRTUALMSLAB, X-Link, X!Link, XLINK). The platform, dependencies and availability of the different software tools reviewed are listed in Supplementary Table S2.

DISCUSSION
MS3D is a valuable addition to the arsenal of tools in structural biology. As an MS-based technique, MS3D is faster and less costly that traditional
methods, and able to analyse samples at the nano-
gram level [77]. The major advantage of MS3D is the
size of molecular assemblies that are amenable to
structural investigation. Developments in data ana-
alysis and cross-linking reagents are also broadening
the scope of MS3D-type approaches and improving
workflow efficiency. This is likely to increase the
throughput of solved structures for mega-Dalton
protein assemblies, complementing established tech-
niques such as XRC and NMR.

To meet the computational demands of the tech-
nique, available software for MS3D analysis has
continued to increase in sophistication. With an
increased emphasis on high-throughput approaches,
many of the more recent programmes are designed
to accommodate larger experimental datasets and
generate larger theoretical libraries for comparison.
The need for confident identification of low abun-
dance cross-linked di-peptides within complex pep-
tide mixtures has been addressed with a more
rigorous statistical treatment of the process of match-
ning experimental against theoretical spectra. This
allows low-abundance di-peptides to be detected
within larger, more complex data sets with fewer
false positives.

While a number of powerful analytical features
have appeared in MS3D software, these are restricted
to a few specific packages which tend to cater for
very specific experimental methodologies. These
typically involve different uses of isotope labelling
and un-cross-linked controls. However, speciality
MS3D software also caters for protein–nucleic acid
complexes (CLPM, MS2Links), top-down MS3D
(MS2Links, MS2PRO), MALDI data (XLINK),
ISD fragmentation (SearchXLinks) and PIR cross-
linker methodology (X-Links).

GPMAW and VIRTUALMSLAB offer versatile
and user-friendly graphic interfaces but do not
have complete MS/MS-matching functionality. MSX-3D
uniquely allows the verification of determined structures by MS3D but, lacking MS/MS
analysis, should only be used to analyse simple ex-
perimental spectra. Pro-Crosslink, SearchXLinks and
XLink perform sound but rudimentary MS/MS
analysis. For the best detection of cross-linked spe-
cies, Crux, xQUEST, XLINK and MassMatrix all
implement sophisticated scoring schemes. Of these
xQUEST and XLINK also exploit isotope labelling.
Crux and XLINK can be run via scripts through a
command line interface, and xQUEST is specially
gearied to process large data sets, but does
demonstrate a slight assignment bias towards single-
chain peptides. MassMatrix has the distinct advantage
that its scoring scheme can be recalibrated by the user
with new datasets.

Currently, an important deficiency is the absence
of a single, integrated software package that also
allows direct structural interpretation and modelling
of protein complexes based on the spatial constraits
revealed by the MS mapping data.

SUPPLEMENTARY DATA
Supplementary data are available online at http://
bib.oxfordjournals.org/.

Key Points
- The mass spectrometric analysis of cross-linked di-peptides,
generated from a chemically cross-linked protein complex fol-
lowing proteolytic cleavage, provides spatial information on the
structural arrangement of constituent subunits in a molecular
assembly, where the mass of the complex may exceed the limit
of traditional structural techniques.
- Rigorous methods for the computational analysis of the complex
MS spectra are essential due to the low yield of cross-linked pep-
tide dimers, as well as the large combinatorial space of potential
products, including those with different combinations of residues
that can be cross-linked, and multiple possible post-translational
modifications.
- Experimental approaches using isotope labelling, MS controls
and reporter ions are possible to simplify the identification
of cross-linked peptide dimers against a background of un-
cross-linked, singly linked and intra-cross-linked peptides.
- Where the sequences of the proteins in a complex are known,
bioinformatics tools identify peptide dimers in the experimental
spectrum by a comparison to a theoretical library of all possible
cross-linked peptides, including a selection of post-translational
modifications.
- The residues involved in an inter-molecular cross-link are identi-
fied by comparing the MS fragmentation spectrum of the dimer
to a library of theoretical spectra predicted for all the possible
combinations of cross-links, if there are more than one, for the
identified di-peptide.
- Bioinformatics software currently available for the analysis of
cross-linked peptides do not yet allow the seamless integration
of the results into a molecular modelling package for refinement
of the structure.

FUNDING
National Research Foundation, South Africa and the
Advanced Biomolecular Research Cluster;
University of the Free State (to H.-G.P.); a post-
graduate bursary from the National Bioinformatics
Network, South Africa (to S.L.N.M.).
References


