Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry

Fan Liu[1,2], Dirk T S Rijkers[3], Harm Post[1,2] & Albert J R Heck[1,2]

We describe an integrated workflow that robustly identifies cross-links from endogenous protein complexes in human cellular lysates. Our approach is based on the application of mass spectrometry (MS)-cleavable cross-linkers, sequential collision-induced dissociation (CID)–tandem MS (MS/MS) and electron-transfer dissociation (ETD)-MS/MS acquisitions, and a dedicated search engine, XlinkX, which allows rapid cross-link identification against a complete human proteome database. This approach allowed us to detect 2,179 unique cross-links (1,665 intraprotein cross-links at a 5% false discovery rate (FDR) and 514 interprotein cross-links at 1% FDR) in HeLa cell lysates. We validated the confidence of our cross-linking results by using a target-decoy strategy and mapping the observed cross-link distances onto existing high-resolution structures. Our data provided new structural information about many protein assemblies and captured dynamic interactions of the ribosome in contact with different elongation factors.

In a cellular context, proteins are highly organized in a variety of assemblies, forming the basis of well-regulated pathways and networks[1]. A rapidly emerging approach, chemical cross-linking combined with MS (XL-MS), provides information on both the structure and the interactions of proteins and protein assemblies[2–5]. In XL-MS, cross-linking reagents, typically consisting of two reactive groups connected by a spacer arm, covalently link two amino acid residues that are in close spatial proximity, and MS is used to identify the linked residues. The spacer-arm length of the cross-linking reagent confers a distance constraint on the two linked residues, providing structural information on proteins, protein complexes and protein-interaction networks.

XL-MS studies have been successfully undertaken to investigate individual purified proteins and protein complexes[6–15] as well as complex protein mixtures[16,17]. However, few attempts have focused on the high-throughput investigation of protein structures and interactions at the proteome-wide level using commercially available cross-linkers (for example, disuccinimidyl suberate and bis(sulfosuccinimidyl) suberate[3]). So far, only dozens to a few hundred cross-links have been identified in *Escherichia coli*[18–20] and *Caenorhabditis elegans*[19] with such methods.

The difficulties in analyzing XL-MS data can be attributed mainly to two factors. The first one is the quadratic expansion of the computational search space, known as 'the *n*-square problem'. Interpeptide cross-links consist of two linear peptides covalently linked together such that the precursor mass obtained from the MS acquisition is the sum of the masses of the two linked peptide moieties and the cross-linker spacer arm. Therefore, the number of possible peptide pairings to be searched increases quadratically with the number of peptides in the database. The second obstacle is the unequal fragmentation efficiency of the two linked peptides. Unlike those of linear peptides, product ion spectra of cross-linked peptides contain fragment ions from both linked peptides. However, ion trap–based CID and higher-energy collisional dissociation (HCD) of cross-links typically favor the formation of product ions from only one of the two constituent peptides[20]. This impairs the confident assignment of cross-links, which depends on high product-ion quality of both linked peptides, especially in complex mixtures.

As an attempt to tackle the *n*-square problem, MS-cleavable cross-linkers, such as disuccinimidyl sulfoxide[21] (DSSO), disuccinimidyl-succinamyl-aspartyl proline[22], bis(succinimidyl)-3-azidomethyl glutarate[23], BuUrBu[24] and cyanurbiotindipropionylsuccinimide[25], have been introduced. This type of cross-linker contains a preferential cleavage site (or sites) in the spacer arm and thus produces signature fragment ions upon MS/MS fragmentation of cross-linked precursors. In recent years, MS-cleavable cross-linkers have been used for proteome-wide cross-linking studies, in which up to a few hundred cross-linked peptides were identified[23,26–28]. However, special requirements for data acquisition (for example, modifications of instrument-operation software for real-time analysis) and/or customized algorithms adapted to the specific cross-linker were required for data analysis. As an example, the software MeroX[29] uses signature fragment ions from MS-cleavable cross-linkers to enhance the confidence of cross-link identification. However, it still uses the vast *n*-square search space and therefore works only with databases containing a limited number of proteins.

Here we describe a generic pipeline for proteome-wide cross-linking studies. We unambiguously identified 2,179 cross-links (1,665 intraprotein cross-links at 5% FDR and 514 interprotein

[1]Biomolecular Mass Spectrometry and Proteomics, Bijvoet Centre for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, University of Utrecht, Utrecht, the Netherlands. [2]Netherlands Proteomics Center, Utrecht, the Netherlands. [3]Medicinal Chemistry and Chemical Biology, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, the Netherlands. Correspondence should be addressed to A.J.R.H. (a.j.r.heck@uu.nl).

RECEIVED 2 MARCH; ACCEPTED 17 AUGUST; PUBLISHED ONLINE 28 SEPTEMBER 2015; DOI:10.1038/NMETH.3603

cross-links at 1% FDR) in HeLa cell lysates using the whole human proteome database (~40,000 entries) as the search space. Furthermore, by mapping identified cross-links on structurally well-characterized proteins or protein complexes, including the ribosome (as well as the ribosome in complex with accessory assemblies such as elongation factors), the proteasome, TriC (TCP-1 ring complex; also known as CCT (chaperonin containing TCP-1)) and the eukaryotic elongation factor 1 (eEF1) complex, we not only validated our approach and the confidence of our data set but also uncovered intriguing new details about the structures and interactions of these protein machineries.

RESULTS

The proteome-wide XL-MS strategy

Our proteome-wide cross-linking strategy incorporates three essential features: the use of MS-cleavable cross-linkers, sequential CID-MS/MS and ETD-MS/MS (CID-ETD-MS/MS) acquisitions for each MS precursor, and a new cross-link search engine, XlinkX. MS-cleavable cross-linkers are used to generate cross-linker–cleaved signature ions in the MS/MS spectra, allowing the user to obtain the accurate mass of each linked peptide during the XlinkX search (described below). CID-ETD-MS/MS acquisitions are applied to overcome the problems of unequal and insufficient fragmentation, thereby enhancing the likelihood of unambiguous identification of each peptide moiety in the cross-linked peptide pairs.

We designed XlinkX to analyze our multipronged data set (**Fig. 1** and Online Methods). XlinkX performs a search in two consecutive steps. In the first step, XlinkX retrieves the precursor mass of each linked peptide on the basis of a unique mass difference (the Δm principle) derived from the gas-phase dissociation of the cross-linker (**Fig. 1a**,**b** and Online Methods). The application of the Δm principle reduces the database size from n^2 to $2n$, enabling the cross-linked peptides to be identified via conventional LC-MS/MS experiments using a full proteome database. In the second step, spectra that contain at least one precursor mass pair are considered as potential cross-linked spectra, and all deduced cross-link pairs are submitted for further peptide-sequence analysis. XlinkX performs product-ion matching for each MS/MS spectrum on the basis of the determined masses of the two linked peptides and all fragment ions generated from the MS precursor. Finally, XlinkX uses an individual peptide probability score (n-score)[30] for spectral quality control and the target-decoy strategy for establishing the FDR. We assessed the software performance of XlinkX and MeroX[29], a recently developed cross-link search engine also designed for cleavable cross-linkers, and found that XlinkX substantially outperformed MeroX when large databases were used (Online Methods, **Supplementary Note 1** and **Supplementary Table 1**).

To demonstrate that XlinkX can efficiently identify cross-linked peptides even when a full proteome database is used, we applied XlinkX in a cross-linking study on a full HeLa cell lysate using the MS-cleavable cross-linker DSSO and CID-ETD-MS/MS acquisitions. All high-confidence cross-link identifications are reported (**Supplementary Data 1**).

First, we assessed the number of cross-link identifications at different FDRs (**Supplementary Fig. 1**). For the statistical analysis, we included a total of 404,454 CID and 404,454 ETD spectra. We carried out CID-only analysis by taking CID but not ETD spectra from the same precursor, whereas for CID-ETD analysis we used all fragment ions from both CID and ETD spectra. At 5% FDR, 2,473 unique cross-links were identified via the combined CID-ETD strategy, whereas 1,113 unique cross-links were identified in CID-only analysis, which corresponds to a difference of more than twofold (**Fig. 2a**). As the comparison used MS/MS spectra originating from the same numbers of MS precursors, it can be assumed that the higher number of cross-link identifications in the CID-ETD analysis was a direct consequence of richer fragment ions. Of note, the duty cycle of our CID-ETD fragmentation strategy was reduced by half, as both CID and ETD fragmentation were performed on the same precursor ion. Therefore, the difference between the combined CID-ETD data and the CID-only data in terms of the number of cross-link identifications would likely be smaller if more CID-MS/MS spectra were acquired. When we repeated the analysis with a more stringent FDR cutoff of 1%, the discrepancy between CID-ETD and CID-only data increased to threefold (1,867 cross-links identified with CID-ETD and 594 cross-links identified with CID only) (**Fig. 2a**). The fact that the advantage of combining CID and ETD was more

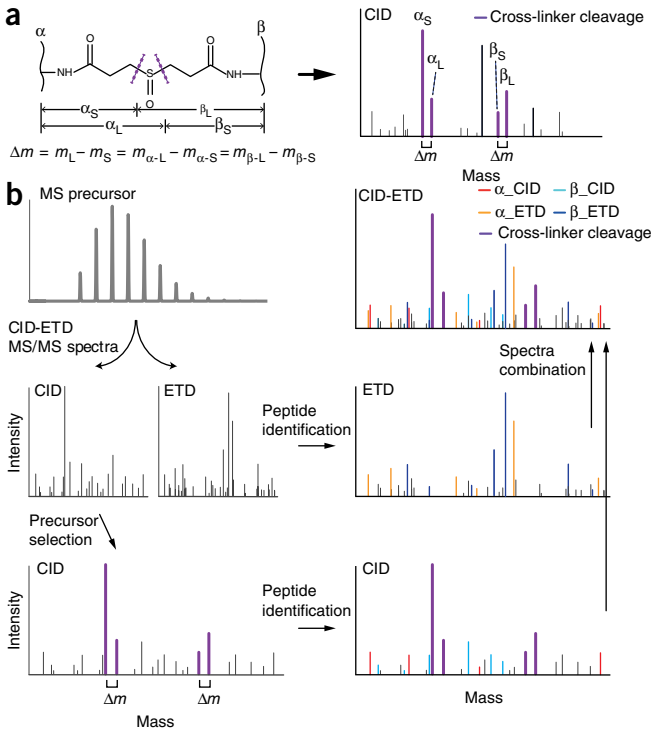

Figure 1 | The proteome-wide XL-MS strategy. (**a**) Schematic structure of a DSSO interpeptide cross-link (left) and its specific fragmentation pattern under CID (right). The four signature MS/MS fragment ions are derived via the equation presented below the structure (i.e., the Δm principle). The variables m_L and m_S represent, respectively, the masses of the longer and shorter arms of the cross-linker after gas-phase dissociation; $m_{\alpha\text{-}L}$, $m_{\alpha\text{-}S}$, $m_{\beta\text{-}L}$ and $m_{\beta\text{-}S}$ are the masses of the four signature peaks. (**b**) The XlinkX workflow to identify interpeptide cross-links. The MS precursor ion is subjected to sequential CID-ETD fragmentation. Only CID spectra are used to obtain the precursor masses of both linked peptides by the Δm principle. The four signature fragment ions resulting from cross-linker cleavage are represented by purple peaks in the MS/MS fragmentation spectra. Subsequently, CID spectra are used to match b- and y-ions, and ETD spectra are used to search for c- and z-ion series.

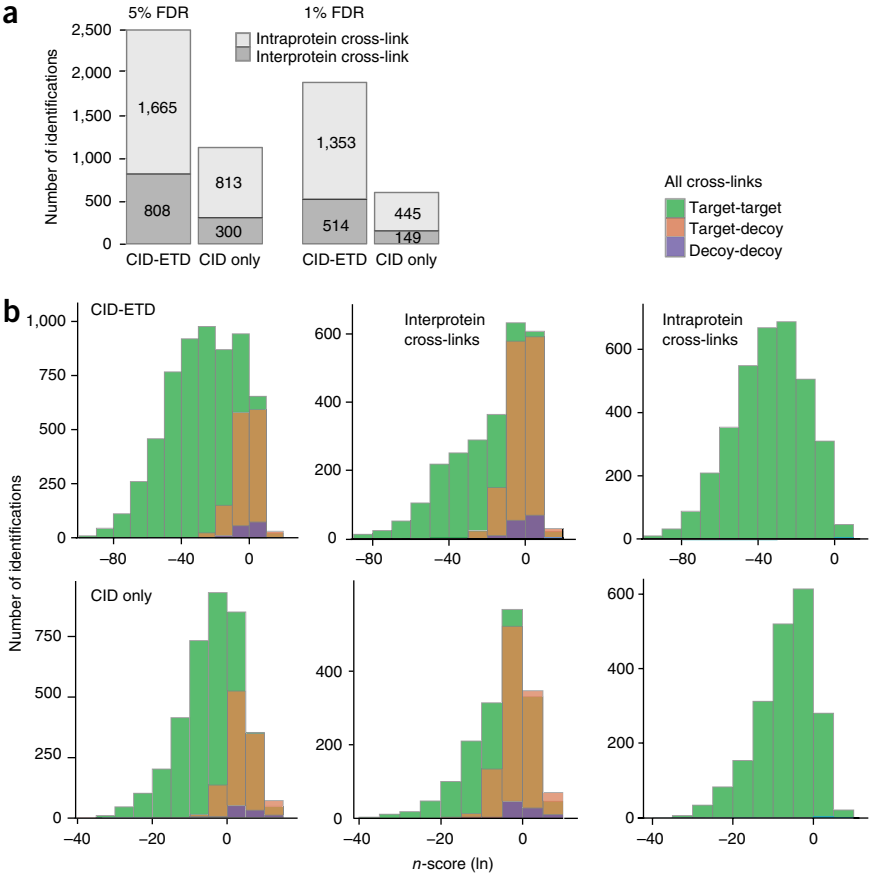

Figure 2 | Statistical analysis of FDR and comparison of CID-ETD and CID-only fragmentation strategies for our HeLa cell lysate cross-linking data set. (**a**) The total number of identified cross-links in different categories. (**b**) Histograms of *n*-scores plotted against the number of identified cross-links. For each of the cross-linked peptides, the higher *n*-score of the two linked peptides was plotted. "Target-target" indicates that both of the linked peptides matched the target database, "target-decoy" indicates that one of the linked peptides matched the decoy database and "decoy-decoy" indicates that both of the linked peptides matched the decoy database.

pronounced at a lower FDR cutoff shows that the CID-ETD strategy performs best when highly confident cross-link assignments are required. This is particularly useful for the discovery of novel interaction partners in a full-proteome context, where unambiguous validation of interprotein cross-links is essential.

Next, we demonstrated the benefit of sequential CID-ETD fragmentation over CID-only strategies at the level of both a single spectrum (**Supplementary Fig. 2**) and the whole data set. For the latter, we calculated the *n*-score distributions of true and false positive hits and found that true positive matches (target-target) were clearly separated from false positive hits (target-decoy and decoy-decoy) in the low–*n*-score region in both CID-only and CID-ETD data (**Fig. 2b**). However, the distribution became very different when we examined the interprotein and intraprotein cross-links separately. For intraprotein cross-links, almost all identifications (regardless of *n*-score) were target hits. This is not surprising, as XlinkX searches and assigns each linked peptide independently of its partner, and therefore the chance that both peptides will result in the same decoy protein is extremely low. In contrast, for interprotein identifications, decoy hits were enriched with both fragmentation strategies. Evidently, the CID-ETD approach was more efficient than the use of CID alone for distinguishing target and decoy hits in the low–*n*-score region. This clearly demonstrates that the combination of informative b- or y-ions and c- or z-ions in the CID-ETD strategy results in a strong improvement in the number of true positive identifications and thus is beneficial for separating target from decoy matches.

On the basis of the statistical analysis of the distribution of target and decoy hits, we decided to use a more stringent FDR cutoff for interprotein cross-links than for intraprotein cross-links. Thus the reported 2,179 cross-links were filtered at 5% FDR for intraprotein cross-links (1,665 cross-links) and at 1% FDR for interprotein cross-links (514 cross-links) (**Fig. 2a**).

Structural validation of and new insight into cross-linking data on the 80S ribosome

To further authenticate our cross-linking data, we assessed the spatial constraints derived from the detected cross-links by comparing them with high-resolution three-dimensional structures.

We mapped our cross-linking data onto the recently published cryo–electron microscopy (cryo-EM) structure of the human 80S ribosome (PDB entry 4V6X), which consists of the 80S core ribosomal proteins in complex with elongation factor 2 (eEF2) and the plasminogen-activator inhibitor 1 RNA-binding protein[31] (SERBP1). We identified a total of 97 unique Lys-Lys cross-links in the 80S ribosomal complex (67 intraprotein cross-links and 30 interprotein cross-links). The maximum Cα-Cα distance between lysine residues that DSSO can cross-link is 23.4 Å. Out of 97 cross-links, 79 cross-links could be mapped on the structurally available regions (**Fig. 3a**); 75 (out of 79) Cα-Cα distances were less than 23.4 Å, and 2 Cα-Cα distances were less than 28.4 Å (23.4 Å + 5 Å tolerance for protein flexibility in solution). The two remaining mapped cross-links had Cα-Cα distances of 34.6 Å and 66.6 Å, respectively, and exceeded the maximum distance constraint.

One anomalous cross-link was an intraprotein cross-link between K498 and K512 of eEF2 (**Supplementary Fig. 3**). eEF2 promotes the GTP-dependent translocation of the ribosome and is transiently bound to the core ribosome during each elongation cycle. The function of eEF2 indicates the existence of multiple conformational states of this protein *in vivo*. Although the observed cross-link violated the structure of ribosome-bound eEF2 presented in the cryo-EM map, it is conceivable that it reflects a conformationally different (e.g., ribosome-unbound) state of the protein.

The other anomalous cross-link was an interprotein cross-link between SERBP1-K140 and RPS28-K16. SERBP1 is an mRNA-binding protein that has an extremely flexible and extended structure. Of its 408 amino acids, only two segments, spanning residues 139–188 and 281–303, were positioned on the 80S

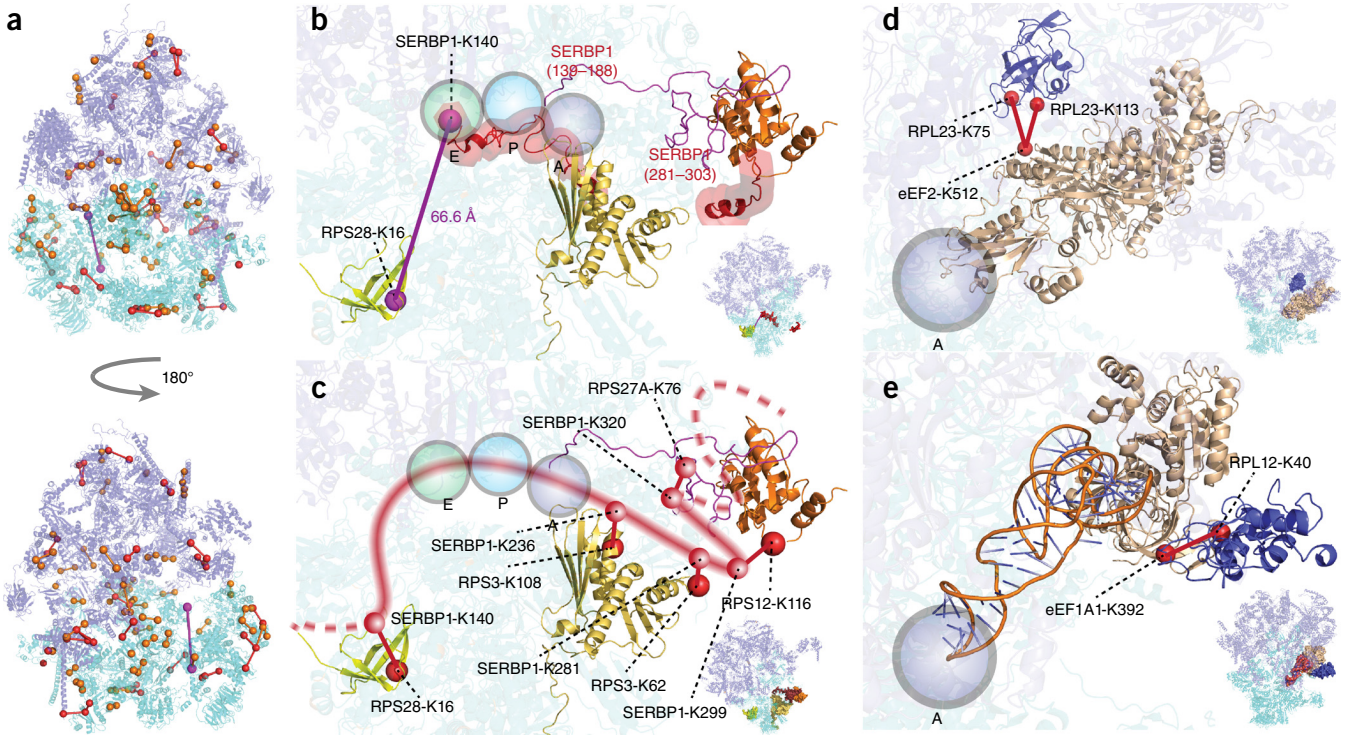

Figure 3 | Mapping detected cross-links on the 80S ribosome. (**a**) All identified cross-links involving ribosomal proteins were mapped onto the available part of the human 80S cryo-EM structure (PDB 4V6X). The small subunit (40S) is shown in cyan, and the large subunit (60S) is shown in purple. Cross-links within the DSSO maximum distance constraint of 28.4 Å are indicated by orange (intrapeptide cross-links) and red (interpeptide cross-links) lines. Cross-links exceeding the maximum distance constraints are represented by magenta lines. (**b**) A zoomed-in view of the anomalous cross-link between SERBP1-K140 and RPS28-K16. SERBP1 (red), RPS28 (yellow-green), RPS3 (yellow), RPS12 (orange) and RPS27A (magenta) are highlighted. The three colored circles (green, blue and purple from left to right) indicate E, P and A tRNA binding sites. (**c**) A schematic positioning of SERBP1 (shown as a red line) on the 80S cryo-EM structure based on our cross-linking data. The protein color scheme is the same as in **b**. (**d**) Cross-link mapping of eEF2-K512 with RPL23-K75 and eEF2-K512 with RPL23-K113 onto the human 80S cryo-EM structure (PDB 4V6X). RPL23 is shown in blue, and eEF2 is shown in light brown. (**e**) Cross-link mapping of eEF1A-K392 with RPL12-K40 onto the rabbit 80S cryo-EM structure (PDB 4UJE). RPL12 is not included in the rabbit 80S structure; therefore human RPL12 was manually positioned on the rabbit 80S ribosome on the basis of a structural alignment using Pymol (version 1.5.0.4). RPL12 and eEF1A1 are shown in blue and light brown, respectively. The tRNA is shown in orange. All high-resolution structures were viewed in Pymol (version 1.5.0.4).

cryo-EM map[31] (**Fig. 3b**). Previous studies have indicated that SERBP1 residues 139–188 pass through the P- and A-site tRNA binding sites and interact with the E-site tRNA[31]. However, our cross-linking data suggest that this segment could alternatively be much closer to RPS28. Considering prior knowledge that SERBP1 passes by the P-, A- and E-site tRNA binding sites and the five newly identified cross-links between SERBP1 and core ribosomal proteins from our study, we propose an alternative localization of SERBP1 on the 80S ribosome (**Fig. 3c**). When we compared our SERBP1 positioning to the cryo-EM structure of rabbit 80S in complex with a 34-nt mRNA fragment[32] (PDB entry 4UJE), we found that our proposed SERBP1 position fit very well with the mRNA localization in that structure (**Supplementary Fig. 4**). This is also in agreement with the function of SERBP1 as the mRNA-binding protein during protein translation. Furthermore, we obtained additional information on the conformation of the C terminus of SERBP1 from five intraprotein cross-links (**Supplementary Fig. 5**). These data suggest that SERBP1 possesses a globular structure at the C terminus (as well as at the mRNA exit site), as opposed to the extended conformation of the remaining regions.

Our cross-linking data also led to the identification of interprotein cross-links between ribosomal core proteins and translation elongation factors, including one cross-link with eEF1-α1 and two cross-links with eEF2 (**Fig. 3d**,**e**). In the cell, the binding of elongation factors to ribosomal core proteins is a transient process. During each cycle of translation elongation, diverse elongation factors transiently bind to ribosomal core proteins to facilitate the elongation process. We mapped the identified cross-links to the cryo-EM structures of the human 80S-eEF2 complex[31] (PDB entry 4V6X) and the rabbit 80S–eEF1A-Val-tRNA-GMPPNP complex[32] (PDB entry 4UJE) and found that our data were in good agreement with these published structures (**Fig. 3d**,**e**). Taking the ribosome as an illustrative example, here we have (1) demonstrated that all mapped cross-links (79 cross-links in total) can be explained on the basis of known structural and functional data, (2) provided new structural insights into the interaction of the SERBP1 protein and (3) shown the capability of cross-linking in probing distinct transient ribosome complexes, a unique feature of studies performed on whole lysates, as opposed to purified samples.

Proteome-wide XL-MS reveals details on structures and interactions of various protein assemblies

We next surveyed our cross-linking data in a more global manner. First, we observed many cross-links from structurally

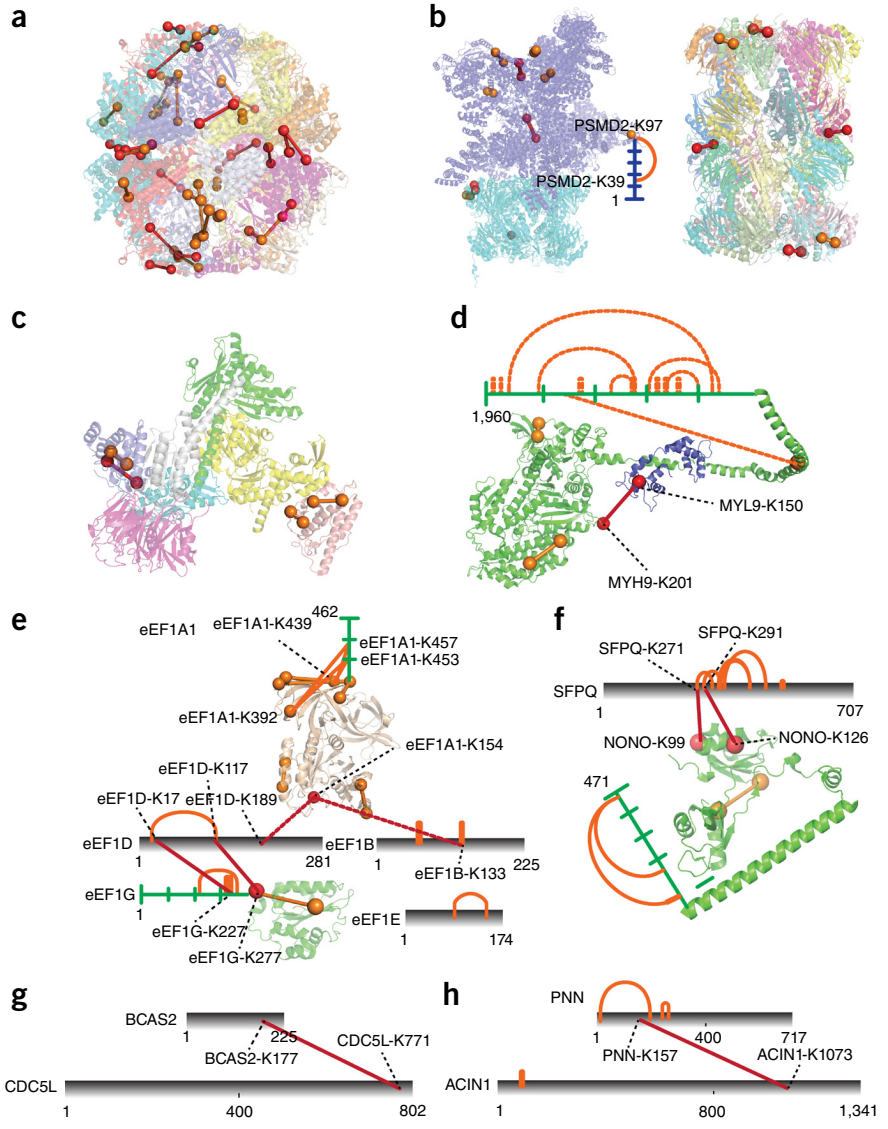

Figure 4 | Cross-link analysis of selected protein complexes. For all panels, all cross-links that could be mapped onto the high-resolution structures are within the DSSO maximal distance constraint of 28.4 Å. Interprotein cross-links are labeled by red lines, and intraprotein cross-links are labeled by orange lines. (**a**) The TriC (CCT) complex in yeast (PDB 4V94). The two identical subunits are labeled with the same color. (**b**) Left, the 26S proteasome of yeast (PDB 4CR4). The 20S core subunit is shown in cyan, and the 19S regulatory subunit is shown in purple. Right, the rat 20S proteasome (PDB 1RPY). (**c**) The bovine Arp2/3 complex (PDB 1K8K). (**d**) Myosin in complex with its regulatory domain (chicken; PDB 3J04). Cross-links on the myosin C terminus are labeled with dashed lines because they may represent contacts between the myosin monomers or within one monomer. (**e**) Cross-links showing the subunit connectivity within the eEF1 complex. aEF1A of *Aeropyrum pernix* (PDB 3VMF) is shown in light brown, and human eEF1G (PDB 1PBU) is shown in green. eEF1A-K154 cross-linked with eEF1B-K133 or eEF1D-K189 is labeled with dashed lines, as one of the linked peptides is shared by eEF1B and eEF1D. (**f**) Interaction of SFPQ and NONO indicated by two detected cross-links. Human NONO (PDB 3SDE) is shown in green. (**g**) Interaction of BCAS2 and CDC5L. (**h**) The identification of a novel protein interaction between PNN and ACIN1. Cross-linking maps were generated with xiNET[35].

well-studied protein assemblies, including the molecular chaperonin TriC, the 26S proteasome, the actin-related protein 2/3 complex, and myosin in complex with its regulatory light chain (**Fig. 4a–d**). These protein assemblies were selected for structural analysis because high-resolution structures of the respective proteins or close homologs were available. Mapping of cross-linked residues on the homologous structures was achieved by protein-sequence alignment. Our results show that all cross-links in the resolved parts of the proteins were in good agreement with the reported structures. We also were able to identify many cross-links from domains lacking X-ray or cryo-EM structures, such as the N terminus of 26S proteasome non-ATPase regulatory subunit 2 and the C terminus of myosin 9 (**Fig. 4b,d**).

Furthermore, our cross-linking data provide new information on previously unknown residue contacts in various protein assemblies (**Fig. 4e–g**). One example is eEF1 (**Fig. 4e**). Our data provide evidence for direct interaction in the cellular context between eEF1D and eEF1G by revealing interprotein cross-links between eEF1D-K17 and eEF1G-K227, as well as between eEF1D-K117 and eEF1G-K277.

Lastly, we identified a set of cross-links that suggested novel protein interactions. One example (**Fig. 4h**) was from a pair of potentially interacting proteins: pinin (PNN) and apoptotic chromatin condensation inducer in the nucleus (ACIN1). Both of these function before or during the pre-mRNA splicing process and have been identified as auxiliary components of the splicing-dependent multiprotein exon-junction compledeposited at the splice junction on mRNAs[33]. Our data provide putative evidence for their direct interaction in cell lysates via an interprotein cross-link between PNN-K157 and ACIN1-K1073.

DISCUSSION

XL-MS has emerged as a valuable technique for the structural analysis of purified proteins and reconstituted protein assemblies, but investigations of protein structures and interactions in a cellular context have been hampered, primarily by the lack of a generic approach that works robustly on complex samples. Here we show that the combination of MS-cleavable cross-linkers, CID-ETD fragmentation strategies and the dedicated XlinkX algorithm provides a generally applicable workflow that effectively overcomes two major obstacles. Using this approach, we generated a large cross-linking data set with high confidence from whole human cell lysates. We show examples illustrating the versatility of XL-MS for studying flexible or disordered proteins, capturing transient interactions and uncovering novel protein interactions.

To assess the sensitivity of our method at a cellular level, we compared all proteins involved in cross-links to data from a previously published study on absolute cellular protein quantities[34]. Not surprisingly, this analysis showed that more than 90% of proteins in our cross-linking study were among the top 500 most abundant proteins in HeLa cells. This clearly indicates that the sensitivity of XL-MS is still limited for analyses of highly complex samples. Opportunities for further improvements are foreseeable at both the protein and the peptide level, such as prefractionation of protein assemblies by size-exclusion chromatography, digestion of proteins with complementary proteases, and more efficient approaches for enriching interpeptide cross-links (for example, the use of affinity-tagged cross-linkers).

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors thank R.A. Scheltema and P. Lössl for critical reading of the manuscript. This work was partly supported by Proteins@Work, a program of the Netherlands Proteomics Centre financed by the Netherlands Organisation for Scientific Research (NWO) as part of the National Roadmap Large-scale Research Facilities of the Netherlands (project number 184.032.201).

AUTHOR CONTRIBUTIONS

F.L. and A.J.R.H designed the research; D.T.S.R synthesized the cross-linker; F.L. and H.P. performed the experiments; F.L. developed the XlinkX software and analyzed the data; F.L. and A.J.R.H. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Robinson, C.V., Sali, A. & Baumeister, W. The molecular sociology of the cell. *Nature* **450**, 973–982 (2007).
2. Bruce, J.E. *In vivo* protein complex topologies: sights through a cross-linking lens. *Proteomics* **12**, 1565–1575 (2012).
3. Rappsilber, J. The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *J. Struct. Biol.* **173**, 530–540 (2011).
4. Sinz, A. Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein-protein interactions. *Mass Spectrom. Rev.* **25**, 663–682 (2006).
5. Sinz, A. The advancement of chemical cross-linking and mass spectrometry for structural proteomics: from single proteins to protein interaction networks. *Expert Rev. Proteomics* **11**, 733–743 (2014).
6. Chen, Z.A. *et al.* Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. *EMBO J.* **29**, 717–726 (2010).
7. Lauber, M.A. & Reilly, J.P. Structural analysis of a prokaryotic ribosome using a novel amidinating cross-linker and mass spectrometry. *J. Proteome Res.* **10**, 3604–3616 (2011).
8. Kao, A. *et al.* Mapping the structural topology of the yeast 19S proteasomal regulatory particle using chemical cross-linking and probabilistic modeling. *Mol. Cell. Proteomics* **11**, 1566–1577 (2012).
9. Leitner, A. *et al.* The molecular architecture of the eukaryotic chaperonin TRiC/CCT. *Structure* **20**, 814–825 (2012).
10. Erzberger, J.P. *et al.* Molecular architecture of the 40S·eIF1·eIF3 translation initiation complex. *Cell* **158**, 1123–1135 (2014).
11. Martinez-Rucobo, F.W. *et al.* Molecular basis of transcription-coupled pre-mRNA capping. *Mol. Cell* **58**, 1079–1089 (2015).
12. Arlt, C., Ihling, C.H. & Sinz, A. Structure of full-length p53 tumor suppressor probed by chemical cross-linking and mass spectrometry. *Proteomics* **15**, 2746–2755 (2015).
13. Plaschka, C. *et al.* Architecture of the RNA polymerase II–Mediator core initiation complex. *Nature* **518**, 376–380 (2015).
14. Benda, C. *et al.* Structural model of a CRISPR RNA-silencing complex reveals the RNA-target cleavage activity in Cmr4. *Mol. Cell* **56**, 43–54 (2014).
15. Greber, B.J. *et al.* Architecture of the large subunit of the mammalian mitochondrial ribosome. *Nature* **505**, 515–519 (2014).
16. Guerrero, C., Milenkovic, T., Przulj, N., Kaiser, P. & Huang, L. Characterization of the proteasome interaction network using a QTAX-based tag-team strategy and protein interaction network analysis. *Proc. Natl. Acad. Sci. USA* **105**, 13333–13338 (2008).
17. Herzog, F. *et al.* Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry. *Science* **337**, 1348–1352 (2012).
18. Rinner, O. *et al.* Identification of cross-linked peptides from large sequence databases. *Nat. Methods* **5**, 315–318 (2008).
19. Yang, B. *et al.* Identification of cross-linked peptides from complex samples. *Nat. Methods* **9**, 904–906 (2012).
20. Trnka, M.J., Baker, P.R., Robinson, P.J.J., Burlingame, A.L. & Chalkley, R.J. Matching cross-linked peptide spectra: only as good as the worse identification. *Mol. Cell. Proteomics* **13**, 420–434 (2014).
21. Kao, A. *et al.* Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes. *Mol. Cell. Proteomics* **10**, M110.002212 (2011).
22. Soderblom, E.J. & Goshe, M.B. Collision-induced dissociative chemical cross-linking reagents and methodology: applications to protein structural characterization using tandem mass spectrometry analysis. *Anal. Chem.* **78**, 8059–8068 (2006).
23. Buncherd, H., Roseboom, W., de Koning, L.J., de Koster, C.G. & De Jong, L. A gas phase cleavage reaction of cross-linked peptides for protein complex topology studies by peptide fragment fingerprinting from large sequence database. *J. Proteomics* **108**, 65–77 (2014).
24. Müller, M.Q., Dreiocker, F., Ihling, C.H., Schäfer, M. & Sinz, A. Cleavable cross-linker for protein structure analysis: reliable identification of cross-linking products by tandem MS. *Anal. Chem.* **82**, 6958–6968 (2010).
25. Petrotchenko, E.V., Serpa, J.J. & Borchers, C.H. An isotopically coded CID-cleavable biotinylated cross-linker for structural proteomics. *Mol. Cell. Proteomics* **10**, M110.001420 (2011).
26. Weisbrod, C.R. *et al.* *In vivo* protein interaction network identified with a novel real-time cross-linked peptide identification strategy. *J. Proteome Res.* **12**, 1569–1579 (2013).
27. Chavez, J.D., Weisbrod, C.R., Zheng, C., Eng, J.K. & Bruce, J.E. Protein interactions, post-translational modifications and topologies in human cells. *Mol. Cell. Proteomics* **12**, 1451–1467 (2013).
28. Navare, A.T. *et al.* Probing the protein interaction network of *Pseudomonas aeruginosa* cells by chemical cross-linking mass spectrometry. *Structure* **23**, 762–773 (2015).
29. Götze, M. *et al.* Automated assignment of MS/MS cleavable cross-links in protein 3D-structure analysis. *J. Am. Soc. Mass Spectrom.* **26**, 83–97 (2015).
30. Meng, F. *et al.* Informatics and multiplexing of intact protein identification in bacteria and the archaea. *Nat. Biotechnol.* **19**, 952–957 (2001).
31. Anger, A.M. *et al.* Structures of the human and *Drosophila* 80S ribosome. *Nature* **497**, 80–85 (2013).
32. Budkevich, T.V. *et al.* Regulation of the mammalian elongation cycle by subunit rolling: a eukaryotic-specific ribosome rearrangement. *Cell* **158**, 121–131 (2014).
33. Tange, T.Ø., Shibuya, T., Jurica, M.S. & Moore, M.J. Biochemical analysis of the EJC reveals two new factors and a stable tetrameric protein core. *RNA* **11**, 1869–1883 (2005).
34. Kulak, N.A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**, 319–324 (2014).
35. Combe, C.W., Fischer, L. & Rappsilber, J. xiNET: cross-link network maps with residue resolution. *Mol. Cell. Proteomics* **14**, 1137–1147 (2015).

ONLINE METHODS

Synthesis and characterization of DSSO. DSSO was synthesized according to a published procedure, and the observed NMR data (not shown) were in line with those reported[21].

Cell culture, cross-linking and proteolytic digestion. HeLa cells (from ATCC; cells were tested for mycoplasma by Hoechst staining) were cultured in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum and 1% penicillin-streptomycin until they reached 80% confluence. Cells were harvested by trypsinization and washed three times with PBS. For whole-cell lysate preparation, the cell pellet was resuspended in cross-linking buffer (20 mM HEPES, 150 mM NaCl, 1.5 mM $MgCl_2$, 0.5 mM DTT, pH 7.8) containing complete EDTA-free protease-inhibitor mixture (Roche). Cells were lysed by sonication on ice for three cycles (30 s in each cycle with 30-s intervals) with 50% amplitude power. Cell debris was then removed by centrifugation at 13,800*g* for 20 min at 4 °C. For nuclear-extract preparation, the cell pellet was first lysed in cellular membrane lysis buffer (10 mM HEPES, 10 mM KCl, 1.5 mM $MgCl_2$, 0.5 mM DTT, 0.4% NP-40, pH 7.8) for 10 min on ice. Then nuclei were pelleted by centrifugation for 10 min at 3,200*g*. The nuclear pellet was resuspended in cross-linking buffer and lysed the same way as the whole cell by sonication. The protein concentration was measured via Bradford assay (Bio-Rad). A total of 500 µg of protein for each sample (whole-cell lysate and nuclear extract) was used for cross-linking, and the protein concentration was adjusted to 1 mg/ml with the cross-linking buffer.

The DSSO cross-linker was first dissolved in dimethyl sulfoxide and then added to the cell lysate to a final concentration of 1 mM. Cross-linking was performed for 1 h at room temperature and quenched with 20 mM Tris-HCl, pH 8.0. Cross-linked proteins were denatured with 8 M urea, reduced with 4 mM DTT at 56 °C for 30 min, and alkylated with 8 mM iodoacetamide at room temperature for 30 min in the dark, after which additional 4 mM DTT was added. A first digestion was carried out with Lys-C at an enzyme-to-protein ratio of 1:75 (wt/wt) at 37 °C for 4 h. The samples were diluted four times and further digested with trypsin at an enzyme-to-protein ratio of 1:100 (wt/wt) at 37 °C overnight. Protein digests were desalted using Sep-Pak C18 cartridges (Waters), dried and stored at −20 °C for further use.

Strong cation exchange (SCX) fractionation of cross-linked peptides. The desalted digests were reconstituted in 10% formic acid and then loaded onto a Zorbax BioSCX–Series II column (0.8 mm inner diameter, 50 mm length, 3.5 µm). SCX solvent A consisted of 0.05% formic acid in 20% acetonitrile, and solvent B consisted of 0.05% formic acid, 0.5 M NaCl in 20% acetonitrile. The SCX gradient was as follows: 0–0.01 min (0–2% B); 0.01–8.01 min (2–3% B); 8.01–14.01 min (3–8% B); 14.01–28 min (8–20% B); 28–38 min (20–40% B); 38–48 min (40–90% B); 48–54 min (90% B); and 54–60 min (0% B). A total of 50 fractions were collected for each sample, dried and stored at −20 °C before further use.

LC-MS/MS analysis. Twenty of the later SCX fractions (fractions 21–40), which predominantly contained the longer and higher charged peptides ($z > 3$), were analyzed using an ultra-HPLC Proxeon EASY-nLC 1000 system (Thermo Fisher Scientific) coupled on-line to an ETD-enabled LTQ Orbitrap Elite mass spectrometer (Thermo Fisher Scientific). Reversed-phase separation was accomplished using a 100-µm inner diameter 2-cm trap column (packed in-house with ReproSil-Pur C18-AQ, 3 µm) (Dr. Maisch) coupled to a 50-µm inner diameter 50-cm analytical column (packed in-house with Poroshell 120 EC-C18, 2.7 µm) (Agilent Technologies). Mobile-phase solvent A consisted of 0.1% formic acid in water, and mobile-phase solvent B consisted of 0.1% formic acid in acetonitrile. The flow rate was set to 100 nL/min. A 180-min gradient was used (7–30% solvent B within 151 min, 30–100% solvent B within 3 min, 100% solvent B for 5 min, 100–7% solvent B within 1 min and 7% solvent B for 20 min).

For the MS/MS experiment, we selected the five most abundant precursors and subjected them to sequential CID-MS/MS and ETD-MS/MS acquisitions. All spectral data were acquired in the Orbitrap mass analyzer. For the MS scans, the scan range was set to 350–1,500 m/z at a resolution of 60,000, and the automatic gain control (AGC) target was set to 1×10^6. For the MS/MS scans, the resolution was set to 15,000, the AGC target was set to 1×10^5, the precursor isolation width was 2 Da, and the maximum injection time was set to 500 ms. The CID normalized collision energy was 35%; the charge-dependent ETD reaction time was enabled, and the ETD AGC target was set to 1×10^5.

Data analysis. The raw data files were converted to .mgf files using Thermo Scientific Proteome Discoverer 1.4 software (Thermo Fisher Scientific) with the add-on node MS2-Spectrum Processor as previously described[11]. The in-house-developed algorithm XlinkX was used for the main search. The following settings were used in XlinkX: precursor ion mass tolerance, 10 ppm; product ion mass tolerance, 20 ppm; fixed modification, Cys carbamidomethylation; variable modification, Met oxidation; allowed number of mis-cleavages, 3. All MS/MS spectra were searched against the full UniProt human database (retrieved in July 2013, containing 41,008 target protein entries) using a concatenated target-decoy database.

XlinkX description. We developed a novel search engine, termed XlinkX, that can identify (1) disulfide-bridged peptides (through integration with our previously published disulfide-bridged peptide search engine SlinkS)[36] and (2) cross-links obtained with non-cleavable cross-linkers (such as disuccinimidyl suberate and bis(sulfosuccinimidyl) suberate), ETD cleavable cross-linkers (such as 3,3'-dithiobis(sulfosuccinimidyl propionate)) and CID- or HCD-cleavable cross-linkers (such as DSSO[21], succinamyl-aspartyl-proline[22] and disuccinimidyl-succinamyl-valyl proline[37]). XlinkX is able to assign fragment ions from MS/MS experiments using CID, HCD, ETD, and EThcD fragmentation. XlinkX supports the cross-link search via either 'enumeration' mode (where all the peptide-pair combinations are considered (i.e., the n-square database)) or 'linear peptide' mode (where the identification of each linked peptide is based on the individual precursor mass and the corresponding product ions). Because of the database expansion problem, 'enumeration' mode can be applied only when the database contains fewer than 50 proteins; 'linear peptide' mode can be used with a database of any size (albeit with the use of MS-cleavable cross-linkers).

In the following, we provide a detailed description of the algorithm design of the 'linear peptide' mode, using the MS-cleavable

cross-linker DSSO as an example. DSSO embeds two CID- or HCD-cleavable C-S bonds adjacent to the sulfide[21]. When a DSSO interpeptide cross-link (where α and β are the two cross-linked peptides and the cross-link is indicated as α-DSSO-β) is subjected to CID or HCD fragmentation, either of the two MS-cleavable sites may preferentially cleave. Because both α and β peptides can be attached with either the longer (L) or the shorter (S) arm of the cross-linker, this cross-linker cleavage results in the formation of four signature fragment ions (α-S, α-L, β-S and β-L) in the MS/MS spectra (**Fig. 1a**). In the initial study of the DSSO cross-linker by Huang's group, these signature peaks were selected for MS^3 acquisitions to provide peptide-sequence information[21]. XlinkX, however, identifies cross-links from MS-cleavable cross-linkers directly from MS^2 spectra. This MS^2 strategy is also applied by the software MeroX[29], which makes use of a signature peak–sensitive scoring algorithm. However, because MeroX uses the signature peaks for scoring only, it still searches against the vast *n*-square database for identification. Consequently, it works only with databases containing a limited number of proteins. In contrast, XlinkX uses the four signature peaks to directly extract the precursor mass of each constituent peptide of a cross-link and thus reduces the search space from n^2 to $2n$. This algorithm design tremendously reduces the database size, allowing us to identify cross-links directly from the whole human database (a detailed comparison between MeroX and XlinkX performance is given in **Supplementary Note 1** and **Supplementary Table 1**).

XlinkX performs the main search in two consecutive steps: (1) determining the precursor mass and (2) matching the fragment masses of each linked peptide (**Fig. 1b**). In the first step, XlinkX retrieves the precursor mass of each linked peptide by searching for a unique mass difference (Δm) among all possible MS^2 fragment ion pairs in the CID spectrum; this is referred to as the Δm principle.

$$\Delta m = m_{\mathrm{L}} - m_{\mathrm{S}} = m_{\alpha\text{-L}} - m_{\alpha\text{-S}} = m_{\beta\text{-L}} - m_{\beta\text{-S}} \quad (1)$$

where m_{L} and m_{S} are the masses of the longer and shorter arms of the cross-linker after gas-phase dissociation, respectively, and $m_{\alpha\text{-L}}$, $m_{\alpha\text{-S}}$, $m_{\beta\text{-L}}$ and $m_{\beta\text{-S}}$ are the masses of the four signature peaks.

The two fragment ion pairs that match equation (1) correspond to the four signature fragments originating from the specific MS-induced cleavage of the cross-linker. Thus, the precursor masses of the two linked peptides (the masses of α and β) can be calculated by

$$m_{\alpha} = m_{\alpha\text{-L}} - m_{\mathrm{L}} = m_{\alpha\text{-S}} - m_{\mathrm{S}} \quad (2a)$$

$$m_{\beta} = m_{\beta\text{-L}} - m_{\mathrm{L}} = m_{\beta\text{-S}} - m_{\mathrm{S}} \quad (2b)$$

where m_{α} and m_{β} are the masses of the two linked peptides α and β.

Here three important features are highlighted. First, although both CID and HCD are able to fragment the DSSO cross-linker, CID performs slightly better in preserving all four signature peaks and was therefore used throughout our experiments (a detailed comparison of CID and HCD performance is given in **Supplementary Note 2**, **Supplementary Table 2** and **Supplementary Fig. 6**). Second, although we performed sequential CID-ETD-MS/MS on each MS precursor ion, only the fragment ions from the CID-MS/MS spectrum were used to deduce the precursor mass. This is because the four signature peaks were present exclusively in the CID spectra, and taking ETD fragment ions into consideration would thus only increase the chance of random matches. Third, the precursor-mass deduction benefits significantly from high-quality CID-MS/MS spectra because the precursor masses (m_{α} and m_{β}) can be retrieved only when all four signature peaks (α-S, α-L, β-S and β-L) are observed.

After the signature peaks have been matched, spectra that contain at least one precursor mass pair are considered as potential cross-linked spectra, and all deduced cross-link pairs from the above equations are submitted for further peptide-sequence analysis. XlinkX performs product-ion matching of each MS^2 spectrum on the basis of the determined masses of the two linked peptides (m_{α} and m_{β}) and all fragment ions generated from the MS precursor. In this step all fragment ions are included in the search; the CID spectrum is used to search exclusively for b- and y-ion series, and the ETD spectrum is used to search exclusively for c- and z-ion series.

XlinkX uses a probability score to calculate the confidence of each candidate sequence; this score is adapted and modified from the *p*-score calculation in ProSightPC (Thermo Fisher Scientific). It reflects the probability of obtaining at least as good a match between the observed fragment list and any random sequence by chance. A low *p*-score means that the probability of obtaining a given set of fragments matching a random sequence is low, and thus it is unlikely that random chance is the cause of the association. The *p*-score is calculated according to the previously published formula[30]

$$p(n) = 1 - \sum_{i=0}^{n-1} \frac{e^{-xf}(xf^{i})}{i!} \quad (3)$$

where *n* is the number of matching fragments and *x* is the probability of an observed fragment ion matching a random theoretical fragment ion by chance and is equal to

$$x = \frac{1}{111.1} \times 4 \times (m_t \times 2) \quad (4)$$

where 111.1 is the average mass of an amino acid, 4 is the number of fragment ions generated from each peptide bond cleavage (b-, y-, c- and z-ion series), and m_t is the user-defined mass tolerance.

f is the calculated total number of MS^2 fragment ions for each linked peptide. For example, *f* for peptide α can be explained as

$$f = \frac{l_{\beta}}{l_{\alpha} + l_{\beta}} \times f_{\mathrm{total}} \quad (5)$$

where l_{α} and l_{β} are the lengths of peptides α and β, respectively, and f_{total} is the total number of MS^2 fragment ions.

The *n*-score is a measurement that takes the size of the database into consideration in the probability estimation. It is calculated as

$$n\text{-score} = p(n) \times N \quad (6)$$

where $p(n)$ is the *p*-score and *N* is the number of proteins in the target database. In the XlinkX result file, the *n*-score of each linked peptide is reported.

doi:10.1038/nmeth.3603

XlinkX employs the target-decoy strategy to establish the FDR. The combined database (containing target and decoy sequences) is used for cross-link identifications, and the FDR is calculated as

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}} \quad (7)$$

where TP represents the true positive matches and FP represents the false positive matches. In the case of cross-link identification, FP is the number of cross-links with at least one of the linked peptides matching the decoy database (target-decoy + decoy-target + decoy-decoy) and TP is the number of cross-links with both of the linked peptides matching the target database (target-target).

Results were filtered by an FDR of 5% for intraprotein cross-links and 1% for interprotein cross-links. All identifications and their annotated spectra are reported in **Supplementary Data 1**. Raw files are publically available via the Chorus data repository (https://chorusproject.org) with project I.D. number 890.

Code availability. XlinkX is publically available at http://sourceforge.net/projects/xlinkx/ and as **Supplementary Software**.

36. Liu, F., van Breukelen, B. & Heck, A.J. Facilitating protein disulfide mapping by a combination of pepsin digestion, electron transfer higher energy dissociation (EThcD), and a dedicated search algorithm SlinkS. *Mol. Cell. Proteomics* **13**, 2776–2786 (2014).
37. Liu, F. & Goshe, M.B. Combinatorial electrostatic collision-induced dissociative chemical cross-linking reagents for probing protein surface topology. *Anal. Chem.* **82**, 6215–6223 (2010).