False discovery rate estimation for cross-linked peptides identified by mass spectrometry

Article *in* Nature Methods · July 2012

DOI: 10.1038/nmeth.2103 · Source: PubMed

8 authors, including:

Franz Herzog
Ludwig-Maximilians-University of Munich
98 PUBLICATIONS **6,422** CITATIONS

SEE PROFILE

Friedrich Förster
Utrecht University
151 PUBLICATIONS **9,255** CITATIONS

SEE PROFILE

False discovery rate estimation for cross-linked peptides identified by mass spectrometry

Thomas Walzthoeni[1,2], Manfred Claassen[3], Alexander Leitner[1], Franz Herzog[1], Stefan Bohn[4], Friedrich Förster[4], Martin Beck[5] & Ruedi Aebersold[1,6]

The mass spectrometric identification of chemically cross-linked peptides (CXMS) specifies spatial restraints of protein complexes; these values complement data obtained from common structure-determination techniques. Generic methods for determining false discovery rates of cross-linked peptide assignments are currently lacking, thus making data sets from CXMS studies inherently incomparable. Here we describe an automated target-decoy strategy and the software tool xProphet, which solve this problem for large multicomponent protein complexes.

To investigate the native protein structure and the topology of protein complexes by using CXMS, proteins are chemically cross-linked in their native state and proteolyzed, and the resulting peptide samples are analyzed by liquid chromatography–tandem mass spectrometry (LC-MS/MS)[1–6]. The fragment-ion spectra of cross-linked peptides are assigned to the corresponding peptide sequences on the basis of database searches[7]. Any confirmed cross-link between two residues provides an upper bound for the distance between these residues in the native protein. Combined with data from complementary structural biology methods and molecular modeling, such distance restraints facilitate the structural analysis of large multicomponent protein assemblies that are not amenable to conclusive atomic-resolution techniques as a whole. This method allowed us and others to solve long-standing problems in structural biology[8–11]. However, the automated validation of sequence database search results from such data sets is a major challenge that has to be addressed before the routine and high-throughput implementation of CXMS.

In standard bottom-up mass spectrometry, the identity of cross-linked peptides is inferred from the quality of the match between observed spectra and predicted sequence-specific fragment-ion patterns, which is expressed as a score. This score does not discriminate unambiguously between correct and incorrect matches; hence, some degree of uncertainty is associated with the identification of each cross-link. Attempts have been made to validate CXMS search results by statistical analysis of search hits matching to random sequences, but the applicability of this strategy has previously been demonstrated only for samples of low complexity and small databases (**Supplementary Table 1**). So far, sequence assignments of cross-linked peptides have been validated mainly by using available score thresholds and manually inspecting the corresponding fragment-ion spectra using different heuristics. This requires a high level of experience, is error prone, is hardly scalable and does not provide an objective measure of false discovery rate (FDR).

Here we introduce several algorithmic strategies to estimate FDRs for cross-linked peptides that take into account the specific requirements of CXMS data sets. We describe an automated approach and introduce a software tool, xProphet, to determine FDRs of large CXMS data sets (**Fig. 1**). The method is based on a target-decoy strategy adapted for database searches of fragment-ion spectra from cross-linked peptides, and it is implemented in the software tools xQuest and xProphet.

As a first step, we extended and improved the scoring function of the xQuest search engine by introducing several novel subscores that were combined with the previously described scoring scheme[12] (**Supplementary Results 1** and **Supplementary Methods**). To optimize this scoring function and to establish a baseline result for the assessment of algorithmic advances with a test data set, we cross-linked eight standard proteins (comprising the '8-mix' data set) separately and mixed them after quenching the cross-linking reaction. This way, only intraprotein cross-links were experimentally possible, whereas all interprotein cross-links identified from the sample were false positive identifications. We searched the 8-mix data set using the previously described scoring scheme[12] against a database containing the target proteins and an additional 100 random *Escherichia coli* sequences to simulate a more complex sample that could give rise to identifiable random matches. We identified 370 true positive cross-linked peptide-spectrum matches (CX-PSMs) corresponding to 46 unique (nonredundant) intraprotein cross-links, all mapping to one of the 8-mix proteins (**Supplementary Table 2** and Online Methods). These identifications were validated using available structures (**Supplementary Results 1**) and constitute a set of high-confidence and high-quality CX-PSMs. We then applied linear discriminant analysis to the test data set to optimize the weights for each individual subscore so that the combined linear discriminant score maximized the separation of true positive and false positive hits. The optimized scoring scheme achieved an excellent separation of false positive from true

[1]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland. [2]Ph.D. Program in Molecular Life Sciences, University of Zurich and ETH Zurich, Zurich, Switzerland. [3]Computer Science Department, Stanford University, Stanford, California, USA. [4]Max Planck Institute of Biochemistry, Martinsried, Germany. [5]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [6]Faculty of Science, University of Zurich, Zurich, Switzerland. Correspondence should be addressed to M.B. (martin.beck@embl.de) or R.A. (aebersold@imsb.biol.ethz.ch).

RECEIVED 19 JANUARY; ACCEPTED 23 MAY; PUBLISHED ONLINE 8 JULY 2012; DOI:10.1038/NMETH.2103

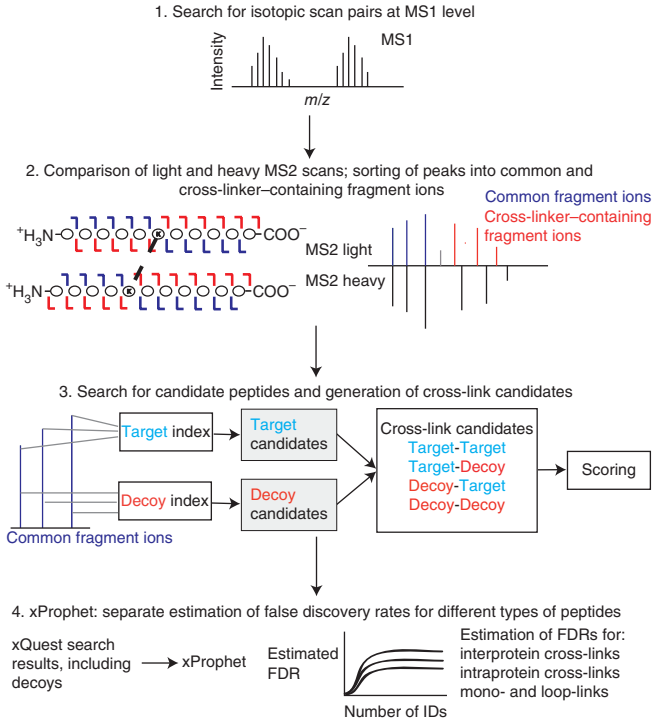

Figure 1 | Identification of cross-linked peptides by xQuest and xProphet. The two peak groups depicted in step 1 represent a scan pair at the MS1 level; the mass difference between the two peak groups indicates the isotopic mass difference of an isotopically labeled cross-linker. In step 2, the red schematic fragment ions contain a cross-linker molecule; the blue ones do not. K indicates lysine residues. Comparison of light and heavy MS2 scans of an isotopic peptide pair (right) illustrates the behavior of the red and blue fragment ions in the fragment-ion spectrum. Steps 3 and 4 depict the xQuest and xProphet workflows, respectively.

positive CX-PSMs in the test data set and exceeded the performance of the previously described xQuest scoring scheme by 92% at CX-PSM level (183:352 CX-PSMs), and by 64% (28:46 CX-PSMs) at the level of nonredundant peptides, respectively (**Fig. 2a** and **Supplementary Figs. 1** and **2**).

For most samples, there is no ground truth for false positive intraprotein cross-links, so we developed a method to determine FDRs for data sets analyzed by the xQuest search engine in cases where the ground truth is not known. We relied on a target-decoy strategy whereby verifiably incorrect 'decoy' sequences are appended to the target-sequence database used by the search engine. The rate of false positive hits mapping to the target database is then estimated from the number of hits mapping to the decoy database[13] (**Fig. 1**).

There are several challenges specific to the identification of CX-PSMs. First, hybrid false positive cross-links, in which one peptide is a correct identification and the second peptide is a random match, have to be taken into account. To test whether hybrid false positive cross-links impose an identification bias, we searched the 8-mix data set against two different databases containing 100 random *E. coli* proteins with and without the eight target proteins. With this method, hybrid false positive hits can be observed only when the target proteins are included. In contrast to entirely random matches—matches in which both cross-linked peptides are wrongly assigned—this test revealed that hybrid false positive cross-links are enriched in the high scoring region (**Fig. 2b**, **Supplementary Fig. 3** and **Supplementary Results 2**).

We designed a mathematical model for the estimation of false positive cross-linked peptides above a certain score threshold that takes this property of hybrid false positives into account (equation (1), Online Methods), whereby #(Decoy) denotes all cross-links containing at least one decoy peptide and #($\text{Decoy}_{\text{D-D}}$) denotes cross-links where both peptides match to a decoy sequence

$$\hat{E}[\#(\text{FP})] = \#(\text{Decoy}) - 2 \times \#(\text{Decoy}_{\text{D-D}}) \quad (1)$$

Second, equal relative proportions of target to decoy hits are ensured by retaining equal numbers of target and decoy candidate

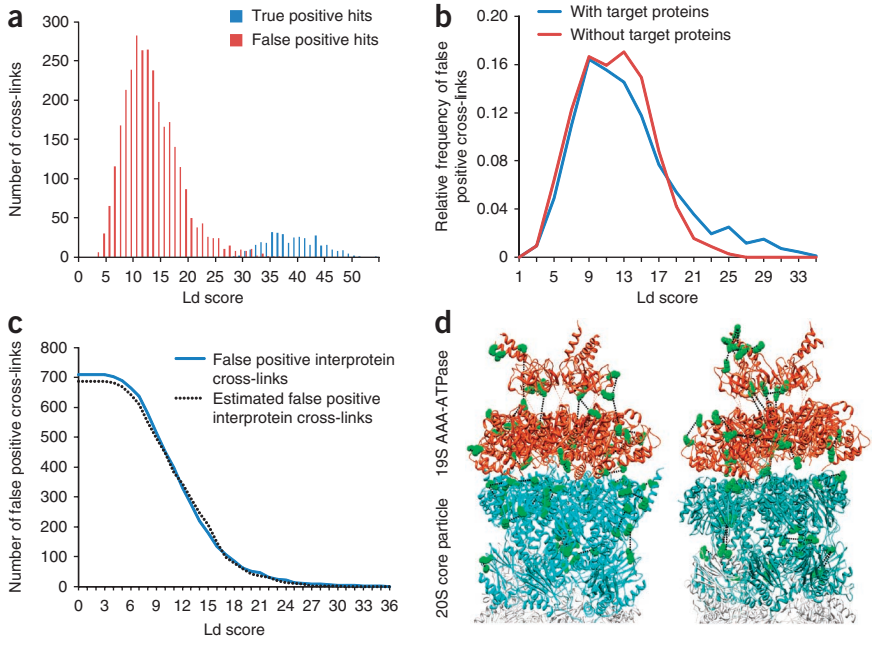

Figure 2 | Validation and application of target-decoy model for cross-linked peptides. (**a**) Separation of true positive from false positive hits in the 8-mix data set after application of the improved and optimized scoring scheme. Ld, linear discriminant. (**b**) Score distribution of false positive interprotein cross-links of the 8-mix data set searched against 100 random *E. coli* proteins with and without target proteins. The elevated right tail accounts for hybrid false positive hits (blue curve). (**c**) Cumulative histogram of the score of actual and target-decoy estimated false positive interprotein cross-links of the 8-mix data set. (**d**) Intra- (left) and interprotein (right) cross-links identified at an FDR of 5% mapped onto the atomic model of the 26S proteasome. Cross-linked lysine residues are indicated in green. Red subunits correspond to subunits of the 19S AAA-ATPase module; blue subunits correspond to α and β ring proteins of the 20S core particle; white subunits correspond to the β′ subunits of the 20S core particle.

peptides before the recombination step to cross-link candidates (**Fig. 1**). Third, FDRs for the different possible types of cross-linker–modified peptides, specifically mono-links, loop-links and cross-links—the latter of which is further segregated into intraprotein cross-links and interprotein cross-links—have to be calculated separately because of their different a priori probabilities for matching. Therefore xProphet estimates the FDR individually for each peptide type.

We then validated this multitiered target-decoy approach. In the case of the 8-mix data, the false positive interprotein cross-link distribution estimated by the target-decoy model accurately matched the actual distribution of false positive interprotein cross-links (**Fig. 2c**). At an FDR cutoff of 5%, as determined by xProphet, 101 unique intraprotein cross-links, 225 mono-links and 59 loop-links were identified. Among these identifications, all intraprotein cross-links mapped to one of the eight target proteins, and three mono-links and three loop-links mapped to an *E. coli* entry, which corresponds to an FDR of 2.1% (for mono- and loop-links), assuming that all hits to the eight target proteins are correct (**Supplementary Table 2** and **Supplementary Fig. 4a**).

To validate the target-decoy approach for a realistic sample consisting of a large multiprotein complex, we generated a CXMS data set of the 26S proteasome, a protein complex comprising up to 33 different subunits. We assessed the plausibility of the identified cross-links by measuring the Euclidean C-α pair distances within the known parts of the structure (20S core particle and the AAA-ATPase module of the 19S regulatory particle, **Fig. 2d**)[10]. At an estimated FDR of 5% or 10%, we identified 198 and 223 cross-linked peptides, respectively. These included 85 interprotein and 113 intraprotein cross-links at an FDR of 5% and 102 interprotein and 121 intraprotein cross-links at an FDR of 10% (**Supplementary Table 3**). Out of the interprotein and intraprotein cross-links identified at an FDR of 5%, 102 could be mapped onto the known part of the proteasome structure (52 interprotein cross-links, 50 intraprotein cross-links). At a false discovery rate of 5%, as determined by the target-decoy approach, 4 out of 102 cross-links violated the distance threshold and 1 additional homotypic cross-link (a cross-link consisting of two identical peptides) could be deduced as a false positive hit. Therefore, the FDR of the structurally verified cross-links corresponded to an FDR of 4.8% (**Supplementary Fig. 4b**). At an estimated FDR of 10%, 52 intraprotein cross-links and 59 interprotein cross-links could be mapped onto the structurally known part of the proteasome. Of these, 5 out of 111 cross-links violated the distance threshold, and an additional 2 cross-links mapped from proteasome subunits to unrelated proteins present in the sample. These data correspond to an FDR of 7.0% of the structurally validated cross-links. Taken together, the validation results from both data sets show that our target-decoy algorithm estimates FDRs accurately within a satisfactory error range in experimental scenarios of different complexity.

The software tools xQuest and xProphet are released as an installable package, including a new web interface (**Supplementary Fig. 5**) that enables a straightforward usage of CXMS with standard equipment (**Supplementary Results 3**). They are publicly accessible from http://proteomics.ethz.ch/.

Thanks to other recent advances in the field—including enrichment strategies for cross-linked peptides[12,14], improvements in mass spectrometry instrumentation in terms of sequencing speed and sensitivity, and the availability of panels of (isotope labeled) cross-linking reagents[15]—cross-linking data sets have become far more comprehensive, more straightforward to generate and of higher quality than in the past. The development of a generic target-decoy model for the computational analysis of such data sets is therefore a crucial step in establishing chemical cross-linking as a routine, high-throughput technique.

METHODS

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

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by the European Union 7th Framework project PROSPECTS (Proteomics Specification in Space and Time, grant HEALTH-F4-2008-201648) and ERC advanced grant 'Proteomics v3.0' (grant no. 233226) of the European Union to R.A. We thank O. Rinner and A.I. Nesvizhskii for constructive discussions.

AUTHOR CONTRIBUTIONS

T.W. and A.L. performed cross-linking experiments and mass spectrometry analysis. T.W. and M.C. developed the target-decoy algorithm. T.W. analyzed the data and designed and wrote the software. S.B. purified 26S proteasomes; F.F. provided the structural model of the 26S proteasome; and T.W., A.L., S.B., M.C., F.H., F.F., M.B. and R.A. discussed and designed experiments. All authors contributed to writing the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nmeth.2103. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Young, M.M. *et al. Proc. Natl. Acad. Sci. USA* **97**, 5802–5806 (2000).
2. Rappsilber, J. *et al. Anal. Chem.* **72**, 267–275 (2000).
3. Leitner, A. *et al. Mol. Cell. Proteomics* **9**, 1634–1649 (2010).
4. Rappsilber, J. *J. Struct. Biol.* **173**, 530–540 (2011).
5. Sinz, A. *Mass Spectrom. Rev.* **25**, 663–682 (2006).
6. Back, J.W. *et al. J. Mol. Biol.* **331**, 303–313 (2003).
7. Mayne, S.L. & Patterton, H.G. *Brief. Bioinform.* **12**, 660–671 (2011).
8. Lasker, K. *et al. Proc. Natl. Acad. Sci. USA* **109**, 1380–1387 (2012).
9. Chen, Z.A. *et al. EMBO J.* **29**, 717–726 (2010).
10. Bohn, S. *et al. Proc. Natl. Acad. Sci. USA* **107**, 20992–20997 (2010).
11. Jennebach, S. *et al. Nucleic Acids Res.* doi:10.1093/nar/gks220 (6 March 2012).
12. Rinner, O. *et al. Nat. Methods* **5**, 315–318 (2008).
13. Elias, J.E. & Gygi, S.P. *Nat. Methods* **4**, 207–214 (2007).
14. Leitner, A. *et al. Mol. Cell Proteomics* **11**, M111.014126 (2012).
15. Petrotchenko, E.V. & Borchers, C.H. *Mass Spectrom. Rev.* **29**, 862–876 (2010).

ONLINE METHODS

Chemical cross-linking of eight standard proteins (8-mix data set). Eight standard proteins—bovine catalase, rabbit creatine kinase, rabbit fructose-bisphosphate aldolase, bovine serum albumin, chicken ovotransferrin, rabbit pyruvate kinase, bovine lactotransferrin and bovine serotransferrin (all from Sigma-Aldrich)—were separately dissolved in 20 mM HEPES, pH 8.3, at protein concentrations of 2 mg/ml. We cross-linked 100 µg of each protein 1 mM DSS (disuccinimidyl suberate, H12/D12, from Creative Molecules, dissolved in dimethylformamide from Thermo Scientific). DSS H12/D12 (a 1:1 molar ratio mixture of DSS-H12 and DSS-D12) is a homobifunctional, isotopically coded cross-linker (**Supplementary Fig. 6**). The reactive groups are *N*-hydroxysuccinimide (NHS) esters that mainly react with primary amino groups to form stable amide bonds. The main targets of DSS when applied to proteins are the ε-amino group of lysines and the N-terminal amino group of proteins. The cross-linking reaction was carried out for 30 min at 37 °C in a thermomixer (Eppendorf) at 750 r.p.m. The reaction was quenched for 20 min at 37 °C by adding ammonium bicarbonate to a final concentration of 50 mM. The proteins were pooled and evaporated to dryness in a vacuum centrifuge; dissolved in 50 µl 8 M urea and reduced with 2.5 mM Tris(2-carboxyethyl)phosphine hydrochloride (TCEP, Pierce) at 37 °C for 30 min; and subsequently alkylated with 5 mM iodoacetamide (Sigma-Aldrich) for 30 min at room temperature in the dark. For digestion, the samples were diluted to 1 M urea and digested by adding 2% (w/w) trypsin (Promega). Digestion was carried out at 37 °C overnight and stopped by acidification to 1% (v/v) formic acid. Peptides were purified using C-18 Sep-Pak columns (Waters) according to the manufacturer's protocol. Enrichment of cross-linked peptides by peptide size-exclusion chromatography and LC-MS/MS analysis was carried out as described previously[14].

Chemical cross-linking and analysis of 26S proteasome. Cross-linking of purified 26S proteasome samples was carried out as described previously[10]. In brief, 26S proteasomes were affinity purified from *Schizosaccharomyces pombe* using a 3×Flag tag placed at the C terminus of RPN11. Purified proteasome samples were concentrated to ~1 mg/ml, and then 50 µl of sample was cross-linked with 1 mM DSS (H12/D12, Creative Molecules) and processed as described above. C-α distances were calculated using UCSF Chimera (v. 1.4.1)[16] and the model of the 26S proteasome, including structures of the 20S core particle (20S CP) and the 19S AAA-ATPase module. Six cross-links within the cavity of the 19S-20S core particle were excluded because they showed a larger distance than that allowed by the cross-linking reagent. However, the longer distance can be explained by a translocation mechanism that is similar to that of the bacterial HsIU[10].

Data analysis with xQuest. The RAW data files were converted to mzXML files using ReadW (version 4.0.2, from TPP). Precursor masses (*m/z*), charge states and retention times of all MS/MS scans were extracted from the mzXML scan headers. MS/MS scan pairs were searched with the xQuest (v.2.1.1) pipeline using the mass shift of DSS (12.07321 Da), ±10 p.p.m. precursor mass tolerance and ±3 min retention time tolerance. The spectra were then searched using target and decoy FASTA databases. For the 8-mix data set, the *E. coli* strain K12 sequence database (organism no. 83333) and sequences of the eight standard proteins were retrieved from UniProt/SwissProt. For pyruvate kinase, two separate entries were created for the two isoenzymes, and the known signal peptides as annotated in UniProt were removed from the primary sequence. We randomly selected 100 *E. coli* sequences from the database and concatenated them with the eight target protein sequences and the sequence of trypsin. The selected sequences are listed in **Supplementary Table 2**. For the 26S proteasome data set, known proteasome subunits for *S. pombe* and sequences of unrelated proteins (trypsin and creatine kinase) that were present in the sample were retrieved from UniProt/SwissProt. The decoy databases were derived by reversing the target database and subsequently shuffling the peptide sequences conserving tryptic cleavage sites. For the xQuest search, the following search parameters were used: maximum number of missed cleavages (excluding the cross-linking site) = 2, peptide length = 5–50 amino acids, fixed modifications = carbamidomethyl-Cys (mass shift = 57.02146 Da), variable modification methionine oxidation (15.99491 Da), mass shift of the light cross-linker = 138.06808 Da, mass shift of mono-links = 156.07864 Da and 155.09643 Da, MS1 tolerance = 10 p.p.m. and MS2 tolerance = 0.2 Da for common ions and 0.3 Da for cross-linker–containing ions.

Target-decoy analysis by xProphet. xProphet first parses xQuest search results and allows prefiltering of the identifications according to several criteria. For FDR calculation, we used xProphet (v. 2.5) with the following parameters: calculation of statistics using unique IDs, filter by p.p.m. using −4 to +7 and filter by δ score <0.95. xProphet analysis is carried out on the top-ranking search hit of each spectrum. In the first step, peptides are filtered and sorted cumulative into score bins of 0.1 score units. In a second step, for each score bin, the false discovery rate (FDR) is estimated, and the hits are annotated with the corresponding FDR values. Thereby FDRs are estimated individually for the following types: mono- and loop-links (cumulative), interprotein cross-links and intraprotein cross-links. An output XML (xproph_out.xml) file is generated that can be viewed and processed using the xQuest viewer (**Supplementary Fig. 5**). The viewer allows browsing and exporting of the data, inspecting spectra and further manual filtering by the user.

Generation of the training data set and linear discriminant analysis (LDA). To generate a training data set of true positive and false positive identifications (true positive set, false positive set), LC-MS/MS data were searched with xQuest using the original scoring scheme and a target and a decoy database each containing an additional 100 random proteins of *E. coli*. The following criteria were used to generate the true positive set: type of cross-link = intraprotein cross-link, error p.p.m. = −4 to +7, δ score = <0.95, linear discriminant (Ld) score >30. We identified 370 XL-PSMs, all corresponding to one of the 8 standard proteins. Accordingly, the false positive set was generated by selecting all interprotein cross-links (target and decoy) using the same criteria as for the true positive set without any score threshold. The false positive set consists of 3,040 hits. For the LDA, the following subscores were considered: MatchOdds score, cross-correlation of cross-linker–containing ions score (xcorrx), cross-correlation of common ions score (xcorrb), wTIC score and intsum score. LDA was performed as previously described[12]. The weights obtained by LDA were applied to the training data set, and the improvement

doi:10.1038/nmeth.2103

was evaluated by counting the XL-PSMs for intraprotein cross-links up to the score that the highest false positive interprotein cross-link achieved (see also **Supplementary Fig. 2**).

Validation of cross-links. Cross-links of the 8-mix data set were validated on the following PDB files using UCSF Chimera (v. 1.4.1)[16]: bovine catalase (1TH3), rabbit creatine kinase (2CRK), rabbit fructose-bisphosphate aldolase (1ADO), bovine serum albumin (Modbase[17] model A5PJX3), chicken ovotransferrin (1AIV), rabbit pyruvate kinase (1PKN), bovine lactotransferrin (1BLF) and bovine serotransferrin (Modbase model Q29443).

FDR estimation. As discussed in the main text, a complication arises from the fact that CXMS data sets contain a heterogeneous set of identifications (mono-links, loop-links, intraprotein cross-links and interprotein cross-links). Each of these identification types has different error characteristics due to the different a priori probabilities for selection as a search hit. Therefore we calculate the global FDRs separately for each of these species (mono-links, loop-links (considered as one type) and cross-links (intraprotein cross-links and interprotein cross-links, considered as two individual types).

For each of the three species types, we compute the FDR for an arbitrary score threshold according to equation (2), where $P(y|\mathrm{fp})$ denotes the probability that a certain type is a false positive; $P(\mathrm{fp})$, the probability of a false positive hit of any type; and $P(y)$, the probability of a search hit being of a certain type[18]

$$\mathrm{FDR}(y) = \frac{P(y\,|\,\mathrm{fp}) \times P(\mathrm{fp})}{P(y)} \tag{2}$$

We estimate FDR(y) by estimating the probabilities $P(y|\mathrm{fp})$, $P(\mathrm{fp})$ and $P(y)$ by the ratio of the estimates for the expected counts of the respective identification sets. Specifically, we introduce the estimate for the expected counts of false positives for a specific type, $\hat{E}(y|\mathrm{fp})$; all expected false positives in the data set, $\hat{E}(\mathrm{all}|\mathrm{fp})$; target hits of a specific type, $T(y)$; and all target hits, $T(\mathrm{all})$. Plugging these variables into equation (2) yields the estimate $\widehat{\mathrm{FDR}}(y)$

$$\widehat{\mathrm{FDR}}(y) = \frac{\left(\frac{\hat{E}(y\,|\,\mathrm{fp})}{\hat{E}(\mathrm{all}\,|\,\mathrm{fp})}\right)\left(\frac{\hat{E}(\mathrm{all}\,|\,\mathrm{fp})}{T(\mathrm{all})}\right)}{\frac{T(y)}{T(\mathrm{all})}} \tag{3}$$

Equation (3) can be simplified to achieve the following formulation for the estimate:

$$\widehat{\mathrm{FDR}}(y) = \frac{\hat{E}(y\,|\,\mathrm{fp})}{T(y)} \tag{4}$$

In summary, equation (4) shows that the estimated FDR ($\widehat{\mathrm{FDR}}(y)$) for a specific type of identification (y) can be computed from the ratio of the expected number of observations of a specific type given that the identifications are false positive identifications ($\hat{E}(y\,|\,\mathrm{fp})$) and the total number of target hits of the corresponding type ($T(y)$).

For the mono- and loop-link type of modified peptides, $\hat{E}(y\,|\,\mathrm{fp})$ above an arbitrary score threshold can be estimated directly from the decoy counts of these types because the background frequency of target and decoy hits is 1:1. Such a direct estimation of $\hat{E}(y\,|\,\mathrm{fp})$ is not possible for cross-linked peptides because the expected false positive hits are a mixture of random false positive hits ($\mathrm{FP}_{\mathrm{!TC\text{-}!TC}}$) and hybrid false positive hits ($\mathrm{FP}_{\mathrm{TC\text{-}!TC}}$) as described in the main text and the **Supplementary Results 2**. We describe how to estimate $\hat{E}(y\,|\,\mathrm{fp})$ for cross-linked peptides in the following paragraph.

Estimation of expected false positive cross-link counts from decoy counts. The expected number of false positive hits is computed by individually estimating the number of different types of spurious cross-links. In the following we distinguish between three (not necessarily disjunctive) types of spurious cross-link identifications. This distinction will allow for an FDR estimate of all cross-link identifications while accounting for the different error contributions of each identification type. Type TC-!TC identifications are composed of one correct target peptide and one peptide that is not a correct target peptide. Type !TC-!TC identifications are composed of two peptides that are both not correct target peptides (none of these types can be counted directly because TC and !TC cannot be distinguished). Type D-D identifications are composed of two peptides that are both mapping to the decoy database (this type can be counted directly). Note that type D-D identifications are a subset of type !TC-!TC identifications. For the following derivation, we assume a set of cross-link identifications above some arbitrary fixed score threshold.

The total number of false positive hits #(FP) of this set can be decomposed as follows:

$$\#(\mathrm{FP}) = \#(\mathrm{FP}_{\mathrm{TC\text{-}!TC}}) + \#(\mathrm{FP}_{\mathrm{!TC\text{-}!TC}}) \tag{5}$$

The expected count of each false positive type given the decoy counts can be estimated as follows:

$$\hat{E}[\#(\mathrm{FP}_{\mathrm{TC\text{-}!TC}})] = \#(\mathrm{Decoy}_{\mathrm{TC\text{-}!TC}}) \times r_{\mathrm{TC\text{-}!TC}} \tag{6}$$

$$\hat{E}[\#(\mathrm{FP}_{\mathrm{!TC\text{-}!TC}})] = \#(\mathrm{Decoy}_{\mathrm{!TC\text{-}!TC}}) \times r_{\mathrm{!TC\text{-}!TC}} \tag{7}$$

where $\#(\mathrm{Decoy}_a)$ denotes the number of decoy cross-link identifications of type a, and r_a denotes the respective target-decoy frequencies. The target-decoy frequencies for the individual types are given by virtue of target-decoy database construction. For TC-!TC and !TC-!TC identifications we have target-decoy frequencies $r_{\mathrm{TC\text{-}!TC}}$ of 1:1 (equation (6)) and $r_{\mathrm{!TC\text{-}!TC}}$ 1:3 (equation (7)). The target-decoy frequencies result from the combinatorial composition of the individual types; the random type !TC-!TC comprises four different, equally likely cases, T-T, T-D, D-T and D-D, where T denotes a target hit and D, a decoy hit. The ratio between target hits and decoy hits is therefore 1:3. For the hybrid type, only two species are generated: TC-T and TC-D, which reflects a 1:1 ratio for this type.

The number of false positives cannot be read out directly because the counts $\#(\mathrm{Decoy}_{\mathrm{TC\text{-}!TC}})$ and $\#(\mathrm{Decoy}_{\mathrm{!TC\text{-}!TC}})$ cannot be counted directly. Although decoy cross-link identifications can be recognized from the occurrence of at least one decoy peptide, it is in general not possible to assign it to either type TC-!TC or !TC-!TC because target peptides cannot be recognized as (in)correct with certainty. Therefore the quantities $\#(\mathrm{Decoy}_{\mathrm{TC\text{-}!TC}})$ and $\#(\mathrm{Decoy}_{\mathrm{!TC\text{-}!TC}})$ have to be estimated from the subset of decoy identifications constituted by two decoy peptides (type D-D).

By virtue of the target-decoy database construction and combinatory considerations for cross-link identifications, it can be seen

that the expected count of decoy identifications exclusively consisting of spurious peptides (!TC-!TC) can be estimated from D-D identifications as follows because the random hits are distributed equally among the decoy types D-D, T-D and D-T:

$$\hat{E}[\#(\text{Decoy}_{\text{!TC-!TC}})] = 3 \times \#(\text{Decoy}_{\text{D-D}}) \quad (8)$$

Considering that the total number of decoy identifications #(Decoy) is the sum of #(Decoy$_{\text{TC-!TC}}$) and #(Decoy$_{\text{!TC-!TC}}$), it further follows:

$$\hat{E}[\#(\text{Decoy}_{\text{TC-!TC}})] = \#(\text{Decoy}) - 3 \times \#(\text{Decoy}_{\text{D-D}}) \quad (9)$$

The expected number of false positive cross-link identifications can now be estimated using the decoy identification calculations from equations (8) and (9). By plugging these estimates into equations (6) and (7), respectively, and using the results to compute the total number of false positives (equation (5)), we obtain equation (10)

$$\hat{E}[\#(\text{FP})] = \#(\text{Decoy}) - 2 \times \#(\text{Decoy}_{\text{D-D}}) \quad (10)$$

Equation (10) shows how the number of expected false positive hits for a given cross-link type (either intraprotein or interprotein cross-link) can be estimated using the corresponding decoy counts of the individual type. As an example, at a certain score threshold, the following counts are observed for decoy interprotein cross-links: the total interprotein decoy cross-links equal 12 (#(Decoy)), and the number of cross-links with both peptides being decoys (#(Decoy$_{\text{D-D}}$) equals 3. According to equation (10), the expected number of false positive hits $\hat{E}$[#(FP)] for this type at the given score threshold is $\hat{E}[\#(\text{FP})] = 12 - 2 \times 3 = 6$.

These estimates can be plugged into equation (4) along with the number of target hits of the corresponding type. In the example, if 100 target hits of a certain type (say, interprotein cross-links) are identified above a specific score threshold, and the number of expected false positive interprotein cross-link hits $\hat{E}(y|\text{fp})$ as estimated based on decoy counts is 6, then the estimated FDR can be computed according to equation (4): $\widehat{\text{FDR}}(y) = 6/100 = 0.06 = 6\%$.

16. Pettersen, E.F. *et al. J. Comput. Chem.* **25**, 1605–1612 (2004).
17. Pieper, U. *et al. Nucleic Acids Res.* **30**, 255–259 (2002).
18. Efron, B. & Tibshirani, R. *Genet. Epidemiol.* **23**, 70–86 (2002).

doi:10.1038/nmeth.2103