Published in final edited form as:
*Nat Methods*. 2008 April ; 5(4): 315–318. doi:10.1038/nmeth.1192.

# Identification of cross-linked peptides from large sequence databases

**Oliver Rinner**[1], **Jan Seebacher**[2], **Thomas Walzthoeni**[1,3], **Lukas Mueller**[1], **Martin Beck**[1], **Alexander Schmidt**[1,4], **Markus Mueller**[1], and **Ruedi Aebersold**[1,2,4,§]

[1]Institute of Molecular Systems Biology, ETH Zurich, Zurich, Wolfgang-Pauli Strasse 16, 8093 Zurich, Switzerland [2]Institute for Systems Biology, 1441 North 34th Street, Seattle, WA 98103-8904 [3]University of Innsbruck, Innrain 52, A-6020 Innsbruck, Austria [4]Faculty of Science, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

## Abstract

We describe a method to identify cross-linked peptides from complex samples and large protein sequence databases. The advance was achieved by combining isotopically tagged cross-linkers, chromatographic enrichment, targeted proteomics, and a novel search engine called xQuest. This software reduces the search space by an upstream candidatepeptide search before the recombination step; we show that xQuest can identify cross-linked peptides from a total *E. coli* lysate with an unrestricted database search.

## Introduction

Cross-linking of proteins is a powerful method to investigate protein conformation [1, 2]. A cross-link between two peptides is indicative for spatial proximity of the two linked amino acids at the time of cross-linking. To exploit this information, both peptides that are involved in a cross-link need to be identified. Improvement in mass spectrometry (MS) technology has made this conceivable, in particular high mass precision spectrometers and the development of isotopically coded cross-linkers [3-5]. Several studies have proven the general feasibility of this approach[3, 6-9]. Nonetheless, this methodology has only been used for single proteins or small, purified protein complexes so far. Two main obstacles - one being of experimental nature the other computational - have been impeding the application of cross-linking to more complex samples: First, cross-linking experiments give rise to complex samples in which the peptides of interest have a low stoichiometry and low frequency of occurrence relative to unmodified peptides. The presence of a large excess of these non-cross-linked species reduces the yield of product ion mass spectra (MS2) from the targeted cross-linked peptides. Second, the computational challenge is foremost the combinatorial explosion of the search space that results when all peptide:peptide combinations in a database are considered. Consequently, currently available cross-link analysis programs [1, 3, 10-12] are limited to a sequence database size of only a few proteins.

[§]Corresponding author: Email: E-mail: aebersold@imsb.biol.ethz.ch Phone number: +41 44 633 31 70 Fax number: +41 44 633 10 51.

**Author contributions** O.R., most of the software design and programming, conducted most of the experiments, and most of the data evaluation; J.S. synthesis of cross-linker, contribution to software design; T.W., scoring, data evaluation, and parts of the experiments; L.N.M., analysis and alignment of MS1 data; M.B., parts of the experiments, data evaluation; A.S., design of mass spectrometric methods for cross-linked samples; M.M, contribution to scoring; R.A., coordination of the project, contribution of ideas and concept.

We developed a workflow that is capable of identifying cross-linked peptides from complex samples and large sequence databases. The method is based on chromatographical enrichment and targeted sequencing of peptides, which are modified by isotopically coded cross-linkers, and the development of a novel software called xQuest. The key feature of the search engine is a combination of a low stringency search for candidate peptides followed by stringent spectrum matching.

# Results

## xQuest workflow

To identify cross-linked peptides from large protein databases, MS2 spectra from isotopically labeled cross-links are searched by xQuest according to the workflow outlined in Figure 1. Samples containing cross-linked peptides are separated on a reverse phase column coupled to a mass spectrometer (LC-MS). The peptide masses are screened for isotopic pairs based on the presence of a characteristic isotopic shift. MS2 spectra from these pairs are analyzed according to the absence or presence of an isotopic shift between peaks in the fragment ion spectrum. Accordingly, peaks are separated into precross-link (common-peaks) and post-cross-link (xlink-peaks). This peak sorting improves specificity by matching only a subset of all peaks against a fraction of all theoretical fragment ions.

## Identifying cross-links with xQuest

The preprocessed spectra can be searched in two principal modes: an exhaustive enumeration mode for up to 100 proteins and an ion-tag mode for large databases. In the enumeration mode all possible peptide:peptide combinations are stored in a precursor-mass coded index for fast searching. This mode is guaranteed to consider every combination of peptides, but due to the n-squared behavior of the search space it is limited in database size. In the ion-tag mode a candidate peptide search is performed before enumeration of peptide:peptide combinations. This search is based on an ion-index that associates fragment ion masses of all peptides in the database to these peptides (Supplementary Methods). When searching a spectrum this index is queried with the mass/charge (*m/z)* values of the most intense common-peaks. All peptides that are associated with these *m/z* values are then matched against the union of spectrum ions including the xlink-ions. The best matching peptides are retained as candidate-peptides. Only the combinations of these peptides that give rise to the precursor-mass are evaluated as candidate cross-links.

To develop a scoring scheme for optimal separation of true and false positive assignments we acquired a large number of spectral pairs (3,151) from monomeric proteins that were cross-linked with a mix of light and heavy (d12) labeled Disuccinimidyl suberate (DSS). These spectra were searched against a database containing sequences of these recombinant proteins and distractor proteins drawn from the *E. coli* database. All identifications that pointed to an intra-protein cross-link of the correct protein where manually verified for sufficient evidence of a correct assignment (Supplementary Fig. 1 and Supplementary Table 1). These sets of true positive and false positive assignments were analyzed by linear discriminant analysis. Thereby a scoring scheme was defined that was able to discriminate between true positive cross-links and false positive hits (Supplementary results, Fig. 2 and Supplementary Figs. 2 and 3 online).

Inspection of the charge distribution of identified cross-links revealed that the majority had charge states > +3. The most likely explanation for this observation is that cross-linked peptides behave like the sum of two independent peptides with a total of two basic tryptic C-termini. This property was used to direct the mass spectrometer to highly charged ions, increasing the proportion of cross-links among the acquired MS2 spectra. Additionally, isotope tagged precursors of higher charge states that were missed in the data dependent sequencing mode were put into inclusion lists and targeted in another MS run.

The prevalence of more highly charged ions among the cross-linked peptides was also the basis for a physical enrichment that we achieved by fractionation with strong cation exchange (SCX) chromatography under acidic conditions. In the fractions that were eluted with higher salt concentrations the targeted highly charged precursor ions were enriched (Supplementary Results and Supplementary Fig. 4 online).

### Identification of cross-links from total E. coli lysate

Having shown that the computational requirements for a full proteome cross-linking could be met by the xQuest algorithm, we cross-linked the soluble fraction of a total *E. coli* lysate with DSS-d0/d12. Spectra from isotopic pairs were searched in the unrestricted ion-tag mode against the total *E. coli* database. The plausibility of the results was evaluated using identified intra-protein cross-links as an internal control. This class of cross-links is experimentally much more abundant than inter-protein cross-links whilst only a tiny fraction (< 0.1%) of all theoretical peptide:peptide combinations in the *E. coli* database are intra-protein cross-links. However, against these odds, we observed a highly significant enrichment for intra-protein cross-links among the highest scoring hits that are ranked as first hit by xQuest ($X^2$ test; $p <0.01$), whereas this preference is almost absent for the second best hits (Fig. 3a).

To assess sensitivity of the ion-tag search all spectra were searched in enumeration mode taking only intra-protein cross-links into account. With this constraint the enumeration index scales linearly with the database size and spectra can exhaustively searched for cross-links within the same protein or between subunits in homo-oligomeric complexes but not between different proteins. The false positive rate in this search mode was determined by searching a combined forward and decoy database (Fig. 3b). Most of the intra-protein cross-links identified in the enumeration mode where also found in the unrestricted ion-tag mode, with the exception of cross-links where one chain was either very short or poorly covered by matching ions (Supplementary Table 2).

We confirmed cross-links of both search modes by spatial proximity where x-ray structures were available. Intra-protein cross-links for 22 monomeric proteins were confirmed (Supplementary Fig. 5). Furthermore, 8 inter-protein cross-links between protein-complex subunits where identified and confirmed in the homo-oligomers Tryptophanase, GroEL, Serine hydroxymethyltransferase, and in two ribosomal subunits and the RNA-polymerase II (Figs. 3c and 3d). These results show for the first time that cross-linked peptides can be identified from samples in which complex protein mixtures were cross-linked and consequently large sequence databases needed to be searched.

## Discussion

Analogously to the shotgun approaches that have collected large numbers of identified sequences and posttranslational modifications, a repository of cross-linked peptides identified from native proteins or protein-complexes would be of tremendous value for structural proteomics and systems biology. Even though homology models can be calculated for a large proportion of the proteome, structural refinement as provided by distance constraints from identified intra-protein cross-links will help to improve the accuracy of such predictions, especially for models that are based on low homology templates[13]. Cross-links between protein subunits can confirm a physical protein interaction and yield spatial constraints to modeling of interaction epitopes. So far this potential has only partly been tapped. The main obstacles have been the computational challenge - caused by the huge search space of all possible peptide:peptide combinations - and the low stoichiometric abundance of cross-linked peptides in digests of cross-linked samples.

In this manuscript we showed that a novel search algorithm and a statistical scoring scheme together with experimental improvements enabled us to identify for the first time cross-linked peptides with high confidence from complex samples. Most of these peptide:peptide links connected lysines within a protein. This class of cross-links is expected to be prevalent in the sample, because the protein surface has generally many solvent exposed lysines whereas the site of a protein:protein interaction might be so small that only few cross-links can form. On the other hand intra-protein cross-links constitute only a tiny fraction of all peptide:peptide combinations of this large database. The prevalence for these hits among the high scoring assignments is hence a strong indication for the specificity of our approach. In order to prove that total database searches for cross-links are feasible we did intentionally not try to restrict the search space by any constraint to the proteins that where considered. It might, however, be useful for certain classes of experiments to limit the database e.g. to proteins that were identified by unmodified peptides.

The complexity of a protein digest from a cross-linked sample is beyond what can be resolved without prior fractionation. Consequently, we identified only cross-links in the presumably most abundant proteins. In order to achieve a sufficient depth of analysis and to enrich for inter-protein cross-links it will be necessary to further enrich for cross-linked peptides. We showed that SCX fractionation enriches for highly charged isotopic pairs, which are mostly cross-links. Thereby they can be partially separated from mono-links and small non-modified peptides. Nonetheless, this is only a partial solution because it co-purifies also unrelated species such as very basic or very large peptides. In order to increase the yield of high quality assignments of cross-linked peptides, they will need to be purified directly. There have been suggestions for enrichment strategies based on affinity tagged cross-linkers, which, however, are not available in isotopically coded form yet or have isotopic signatures too small for electrospray ionization[14, 15]. Having shown that high throughput identification of cross-linked peptides is now feasible it should spark interest to invest in advanced enrichment strategies for cross-linked peptides. Given the enormous number potential cross-links within and between proteins this would be a virtual inexhaustible source of information for structural biology and protein:protein interaction studies.

## Material and Methods

Cross-linking of standard-proteins; Cross-linking of *E. coli* lysate; Reduction, alkylation, and digestion; SCX cleanup, Isoelectric focusing; LC-MS; xQuest software; Linear discriminate analysis and determination of Sensitivity and Selectivity. See Supplementary Methods for detailed descriptions.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

OR was supported by fellowships of the Deutsche Forschungsgemeinschaft and the Roche Research Foundation, MB was supported by a long-term fellowship of the European Molecular Biology Organization (EMBO) and a Marie Curie fellowship of the European Commission. AS was supported in part by a grant from F. Hoffmann-La Roche Ltd (Basel, Switzerland) provided to the Competence Center for Systems Physiology and Metabolic Disease. The work was supported in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, under contract No. N01-HV-28179 and with funds from ETH Zurich.

## References

1. Young MM, et al. High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. Proc Natl Acad Sci U S A 2000;97:5802–5806. [PubMed: 10811876]

2. Sinz A. Chemical cross-linking and mass spectrometry for mapping three-dimensional structures of proteins and protein complexes. J Mass Spectrom 2003;38:1225–1237. [PubMed: 14696200]

3. Seebacher J, et al. Protein cross-linking analysis using mass spectrometry, isotope-coded cross-linkers, and integrated computational data processing. Journal of proteome research 2006;5:2270–2282. [PubMed: 16944939]

4. Ihling C, et al. Isotope-Labeled Cross-Linkers and Fourier Transform Ion Cyclotron Resonance Mass Spectrometry for Structural Analysis of a Protein/Peptide Complex. J Am Soc Mass Spectrom. 2006

5. Muller DR, et al. Isotope-tagged cross-linking reagents. A new tool in mass spectrometric protein interaction analysis. Anal Chem 2001;73:1927–1934. [PubMed: 11354472]

6. Dihazi GH, Sinz A. Mapping low-resolution three-dimensional protein structures using chemical cross-linking and Fourier transform ion-cyclotron resonance mass spectrometry. Rapid Commun Mass Spectrom 2003;17:2005–2014. [PubMed: 12913864]

7. Huang BX, Kim HY, Dass C. Probing three-dimensional structure of bovine serum albumin by chemical cross-linking and mass spectrometry. J Am Soc Mass Spectrom 2004;15:1237–1247. [PubMed: 15276171]

8. Novak P. Unambiguous assignment of intramolecular chemical cross-links in modified mammalian membrane proteins by fourier transform-tandem mass spectrometry. Anal Chem 2005;77:5101–5106. [PubMed: 16097745]

9. Pearson KM, Pannell LK, Fales HM. Intramolecular cross-linking experiments on cytochrome c and ribonuclease A using an isotope multiplet method. Rapid Commun Mass Spectrom 2002;16:149–159. [PubMed: 11803535]

10. Gao Q, et al. Pro-CrossLink. Software tool for protein cross-linking and mass spectrometry. Anal Chem 2006;78:2145–2149. [PubMed: 16579592]

11. Schilling B, Row RH, Gibson BW, Guo X, Young MM. MS2Assign, automated assignment and nomenclature of tandem mass spectra of chemically crosslinked peptides. J Am Soc Mass Spectrom 2003;14:834–850. [PubMed: 12892908]

12. de Koning LJ, et al. Computer-assisted mass spectrometric analysis of naturally occurring and artificially introduced cross-links in proteins and protein complexes. The FEBS journal 2006;273:281–291. [PubMed: 16403016]

13. Back JW, de Jong L, Muijsers AO, de Koster CG. Chemical cross-linking and mass spectrometry for protein structural modeling. J Mol Biol 2003;331:303–313. [PubMed: 12888339]

14. Sinz A, Kalkhof S, Ihling C. Mapping protein interfaces by a trifunctional cross-linker combined with MALDI-TOF and ESI-FTICR mass spectrometry. J Am Soc Mass Spectrom 2005;16:1921–1931. [PubMed: 16246579]

15. Trester-Zedlitz M, et al. A modular cross-linking approach for exploring protein interactions. J Am Chem Soc 2003;125:2416–2425. [PubMed: 12603129]

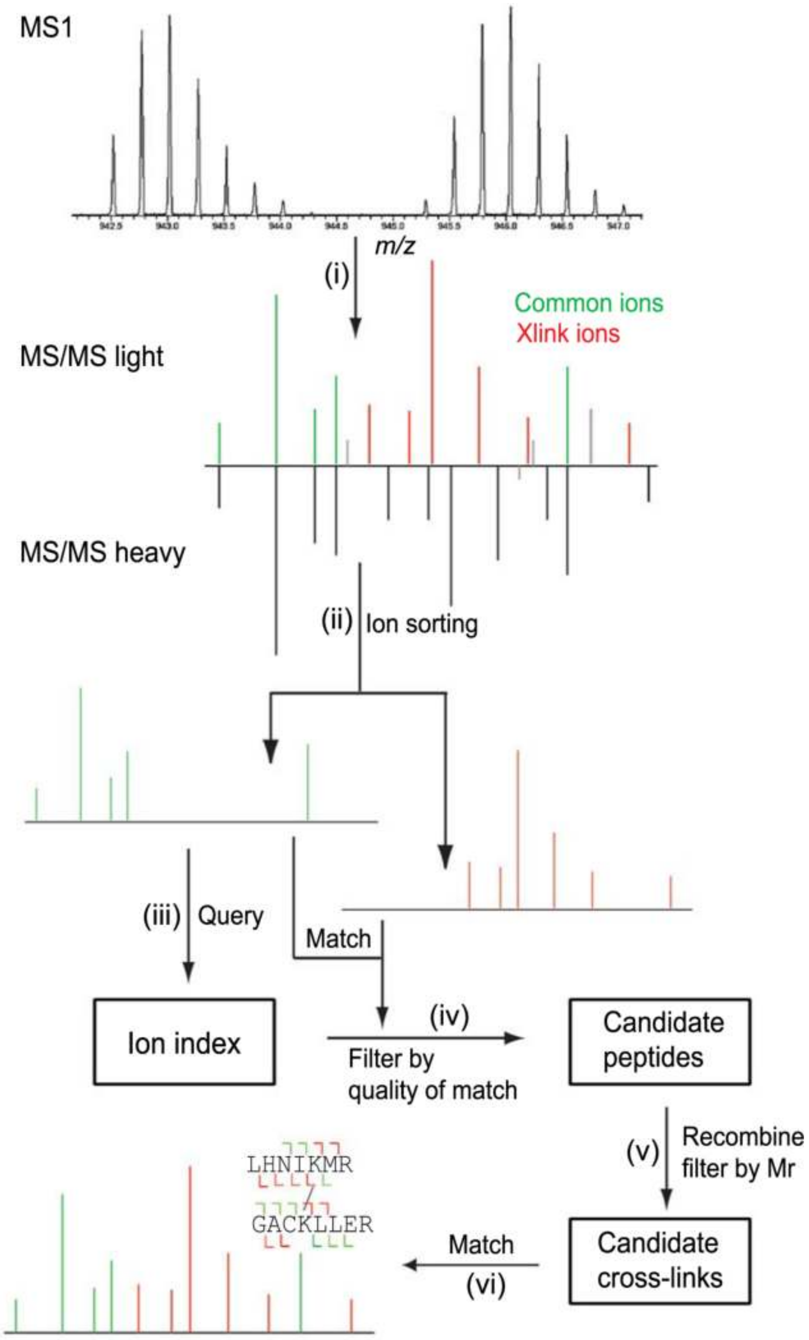

**Figure 1.**

Workflow of cross-link search with xQuest in the ion-tag mode. The light and heavy form of isotopic peptide pairs are detected in precursor ion spectra and subjected to separate MS2 sequencing (**i**). MS2 spectra are compared; fragment ions that are present in both spectra (green) are labeled common-ions; ions with a characteristic isotopic shift (red) are xlink-ions (**ii**). Common-ions are used to query the ion-index which contains all peptide sequences that can give rise to a specific fragment-ion *m/z* (**iii**). Candidate peptides from this ion-index are matched against the union of common- and xlink-ions. Candidate peptides, which match sufficiently well to the spectrum are retained (**iv**) and recombined to cross-links (**v**). Only peptide combinations that match the precursor ion mass are retained and scored against the reconstructed MS2 spectrum (**vi**).

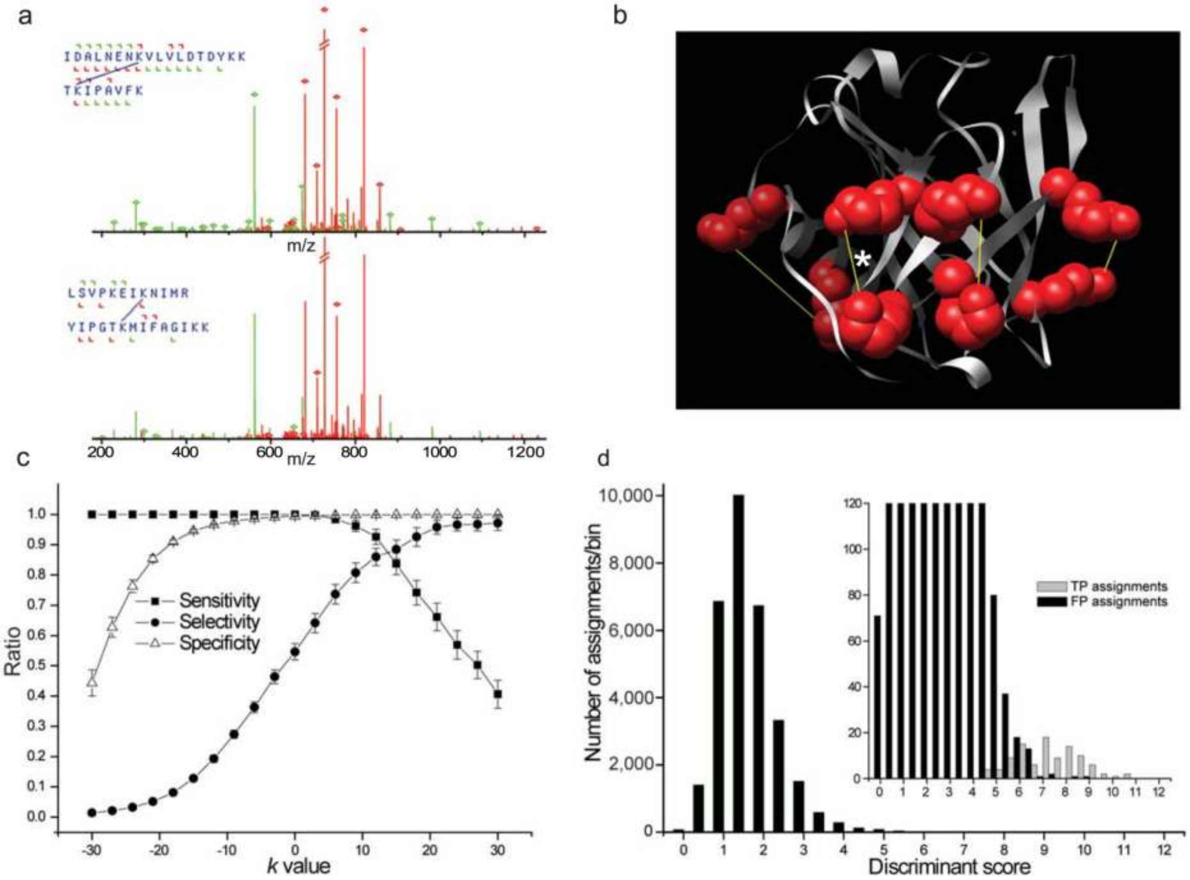

**Figure 2.**

Linear discriminant analysis (LDA) separates true positive MS2 assignments to standard protein cross-links from false positive hits in a large peptide:peptide database. (**a**) Example spectrum of [M+5H]5+ charged cross-link assigned to β-Lactoglobulin (upper spectrum). The lower spectrum shows the next best random match to unrelated *E. coli* proteins. Matches (diamonds) are indicated with a mass tolerance of 0.2 Da for common-ions (green) and 0.3 Da for *x-link* ions (red). **(b)** Spectral assignments were verified for spatial plausibility with 3D structures or homology models as shown for bovine β-Lactoglobulin. Yellow lines indicate cross-links between the ε-amino groups. The spectrum of the cross-link indicated by the star is shown in **a**. (**c**) Sensitivity (TP/(TP+FN)), selectivity (1-FP/(TP+FP)), and specificity (TN/(TN+FP)) of assignments for different k-values (FP: false positive, TP: true positive, FN: false negative, TN: true negative). Error bars indicate s.d. of the bootstrap resampling distributions (n = 1,000). (**d**) Distribution of discriminant scores with weights derived from the LDA. Inset shows that the distribution of true positive assignment scores is separated from the scores of the large number of false positive assignments.

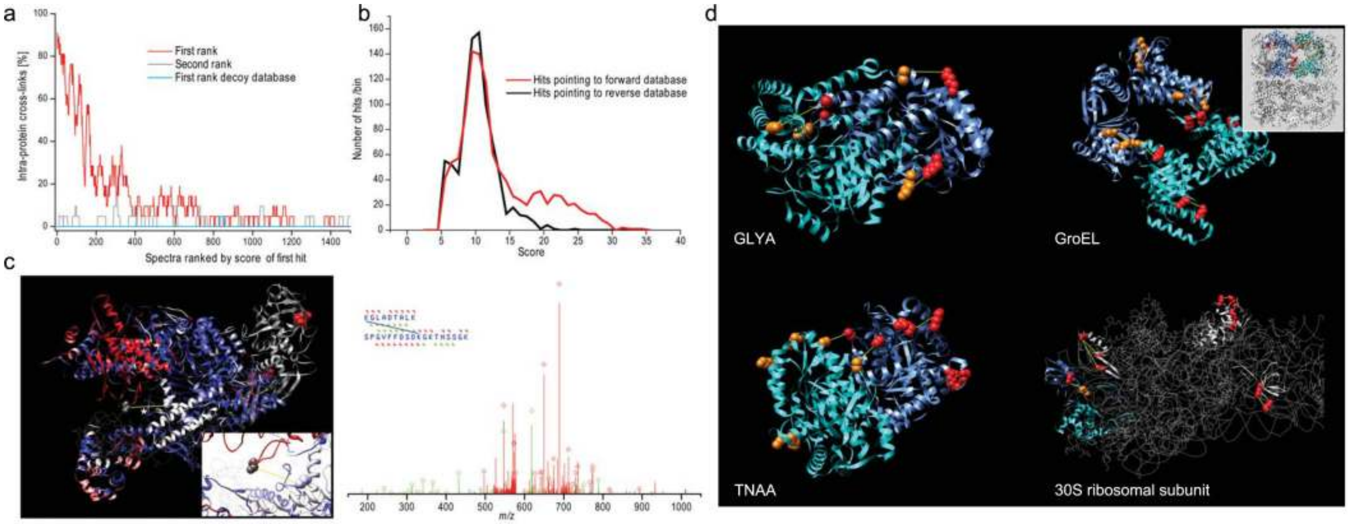

**Figure 3.**

Inter-protein and intra-protein cross-links were identified from a total *E. coli* lysate, searched against the total *E. coli* protein database. (**a**) Top ranked search hits to cross-links show a highly significant preference for intra-protein cross-links. Search hits ranked at second place show only a slight preference. Almost no intra-protein cross-links were identified with the reversed sequence database. The curves where smoothed with a sliding average window of size 20. (**b**) An exhaustive search for intra-protein cross-links was performed in enumeration mode against the full *E. coli* database and a decoy database. True positive hits are clearly separated from false positive hits to cross-links in the decoy database. (**c**) Several cross-links were found for the RNA polymerase. A homology model was created based on the x-ray structure of *Thermus aquaticus* (Taq); thin line is the backbone of the Taq structure. Subunits B and B' are color coded for template homology (red: high homology, blue: low homology). A cross-link between B and B' was identified in a low homology region of the complex. The link is shown between the conserved lysine in B and a proline, which corresponds most closely to the lysine in the *E. coli* sequence. Inset shows a magnification of the linked region. The spectrum of the cross-link is shown and annotated as described above. (**d**) Examples of cross-links identified within and between complex subunits (colored red-blue and orange-cyan respectively) of Serine hydroxymethyltransferase (GLYA), GroEL (inset shows whole complex), Tryptophanase (TNAA) and the small ribosomal subunit.