BMC Bioinformatics

SOFTWARE

Open Access

ECL 3.0: a sensitive peptide identification tool for cross-linking mass spectrometry data analysis

Chen Zhou[1], Shuaijian Dai[1], Shengzhi Lai[1], Yuanqiao Lin[1], Xuechen Zhang[1], Ning Li[2,3] and Weichuan Yu[1,3*]

*Correspondence:
eeyu@ust.hk

[1] Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China
[2] Division of Life Science, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China
[3] HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Shenzhen, China

Abstract

Background: Cross-linking mass spectrometry (XL-MS) is a powerful technique for detecting protein–protein interactions (PPIs) and modeling protein structures in a high-throughput manner. In XL-MS experiments, proteins are cross-linked by a chemical reagent (namely cross-linker), fragmented, and then fed into a tandem mass spectrum (MS/MS). Cross-linkers are either cleavable or non-cleavable, and each type requires distinct data analysis tools. However, both types of cross-linkers suffer from imbalanced fragmentation efficiency, resulting in a large number of unidentifiable spectra that hinder the discovery of PPIs and protein conformations. To address this challenge, researchers have sought to improve the sensitivity of XL-MS through invention of novel cross-linking reagents, optimization of sample preparation protocols, and development of data analysis algorithms. One promising approach to developing new data analysis methods is to apply a protein feedback mechanism in the analysis. It has significantly improved the sensitivity of analysis methods in the cleavable cross-linking data. The application of the protein feedback mechanism to the analysis of non-cleavable cross-linking data is expected to have an even greater impact because the majority of XL-MS experiments currently employs non-cleavable cross-linkers.

Results: In this study, we applied the protein feedback mechanism to the analysis of both non-cleavable and cleavable cross-linking data and observed a substantial improvement in cross-link spectrum matches (CSMs) compared to conventional methods. Furthermore, we developed a new software program, ECL 3.0, that integrates two algorithms and includes a user-friendly graphical interface to facilitate wider applications of this new program.

Conclusions: ECL 3.0 source code is available at https://github.com/yuweichuan/ECL-PF.git. A quick tutorial is available at https://youtu.be/PpZgbi8V2xI.

Keywords: Proteomics, Cross-linking mass spectrometry, Database searching, Protein feedback

Background

Cross-linking mass spectrometry (XL-MS) is an emerging technology in the field of proteomics that provides insights into protein–protein interactions (PPIs) and protein structures. The significance of XL-MS lies in its ability to detect PPIs that are weak, transient, or difficult to study using other methods such as co-immunoprecipitation or yeast two-hybrid assays. Compared to X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy, XL-MS needs less complex sample preparation and provides much higher throughput in the structural study [1].

However, one of the major challenges of XL-MS is the low peptide fragmentation efficiency in the collision-induced dissociation (CID) process. During the process, longer (or heavier) cross-linked peptides often impede the fragmentation of shorter (or lighter) cross-linked peptides, resulting in less informative fragment ions in the MS/MS spectrum [2, 3]. To address this problem, researchers have proposed to design novel cross-linkers, refine experimental procedures and optimize analysis algorithms.

We proposed a protein feedback idea in the cleavable cross-linking data analysis and used the protein-peptide association to help identify insufficient fragmented peptides [4]. This method has improved the sensitivity in the cleavable cross-linking data.

The implementation of the protein feedback method to the non-cleavable cross-linking data can make more impact in the field because the majority of XL-MS experiments still uses non-cleavable cross-linkers [5, 6]. In this paper, we implement the protein feedback idea in the analysis of non-cleavable cross-linking MS data based on our previously developed non-cleavable data analysis tools [7–9]. We further integrate both non-cleavable and cleavable tools into a unified software program with a user-friendly graphic interface. We call the new program ECL 3.0. Experimental results have demonstrated the superior performance of ECL 3.0 over existing standard analysis tools.

Implementation

In XL-MS, imbalanced fragmentation of cross-linked peptides leads to low-quality XL-MS spectra. Although designing suitable scoring functions can help identify correct peptides, scoring function does not completely solve the problem, especially when no fragmented ions of the shorter or lighter peptide are available. To address this issue, we propose to use additional (global) information in peptide identification.

Suppose two similar peptides are ranked as the top hits by a scoring function with no preference for one over the other. One peptide has many sibling peptides identified from other spectra, whereas the other does not have any sibling peptides being identified elsewhere. Because sibling peptides from the same protein should have higher chances to appear in other MS/MS spectra than some irrelevant peptides, we propose to add more weight to a peptide with many sibling peptides than a peptide without sibling peptides. We define this adjustment as the protein feedback method [4]. For completeness, we briefly summarize the implementations of the protein feedback method below. The reader is referred to [4] for mathematical details:

1. In XL-MS, each query spectrum undergoes preliminary identification of potential peptides using a scoring function. The top N (default 20) cross-linked candidate pairs are then cached for examination. Subsequently, we apply a filtering step to retain

only the highest-scored candidate and those whose score is at least 80% of the top one.

2. The first hit in each spectrum is collected and filtered based on a score cutoff, which is determined empirically and varies depending on the specific scoring function used.
3. The filtered peptides are considered the "true" positives temporarily and are used to build a protein score database. Each of these "true" positives contributes weight to its corresponding protein score [4].
4. After building the protein score database, the top N pairs of each spectrum are re-ranked by their protein scores. And the cross-linked pair with the highest protein score is considered the correct cross-link spectrum match (CSM).
5. After the adjustment of protein feedback in the peptide-spectrum matching process, the target-decoy approach is used to control the false discovery rate (FDR) [10].

ECL 3.0 is an extension of ECL-PF [4], which was focused exclusively on cleavable cross-linking search. In addition to ECL-PF, ECL 3.0 includes Xolik [9], a non-cleavable searching program we previously designed, and incorporates protein feedback. The main code modification in the original ECL-PF cleavable module was the output format. Both ECL-PF and Xolik are standardized to ensure consistent output results.

ECL 3.0 is written in conjunction with Python3 and C++. We provide a testing dataset at http://bioinformatics.hkust.edu.hk to help users get started with the software. Additionally, we have created a detailed tutorial with step-by-step instructions for using ECL 3.0, which can be accessed at youtu.be/PpZgbi8V2xI.

Results and discussion

Figure 1 illustrates the graphical user interface (GUI) of ECL 3.0. Figure 2 shows comparison results between ECL 3.0 and four other state-of-the-art methods. In non-cleavable XL-MS data analysis, ECL 3.0 was compared with Kojak [11] and pLink 2 [12]. In cleavable XL-MS data analysis, ECL 3.0 was compared with MaxLynx [13] and MeroX [14]. The comparison was performed using a synthetic data set [15], and the results from cleavable data analysis were redrawn from [4]. ECL 3.0 achieves significantly higher sensitivity in terms of the number of CSMs with a similar level of precision. For the unique cross-linked peptides, all of the tools have similar levels of sensitivity. This is due to the fact that this data set is of high quality and only contains hundreds of synthesized peptides. We further tested these tools on larger and more complicated real datasets of human proteins. The result (Additional file 1: Fig. S5) reveals that ECL 3.0 can identify 49% more unique cross-linked peptides than other software.

We have done three different types of validation using real data sets in [4], including using the Protein Data Bank (PDB) structure to verify the result, using the protein–protein interactions (PPIs) database (BioGRID and DroID) to verify the protein interactions, and using the artificially created junk data set as negative control to check the false positives. During the revision of this manuscript, we are inspired by the reviewer's suggestions and further use another way to validate the result of ECL 3.0 (in Additional file 1). These validations depict that ECL 3.0 improves the sensitivity of cross-linking peptide identification without sacrificing precision.

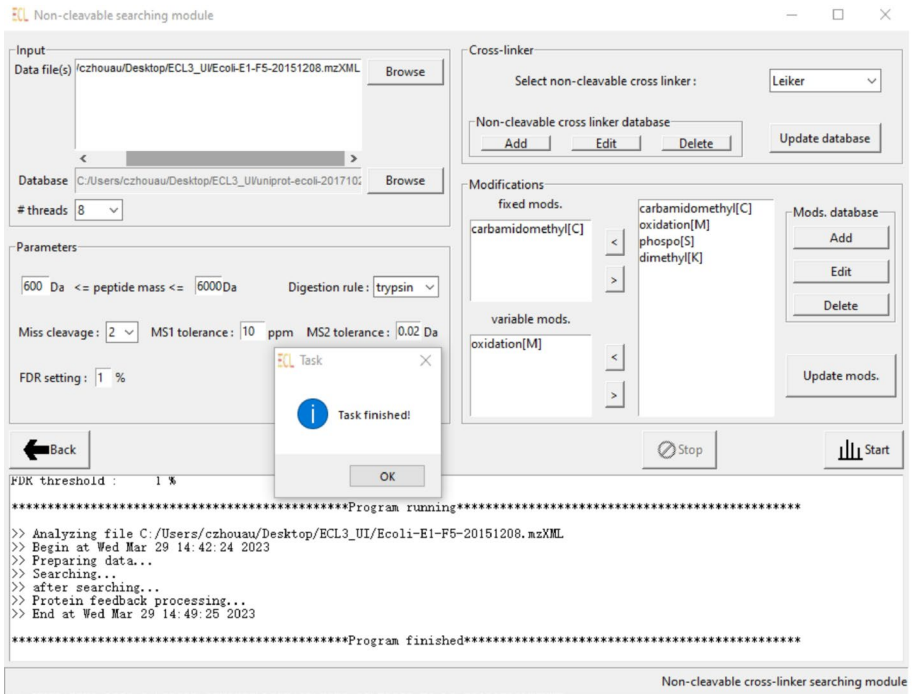

Fig. 1 A snapshot of graphical user interface (GUI) in ECL 3.0

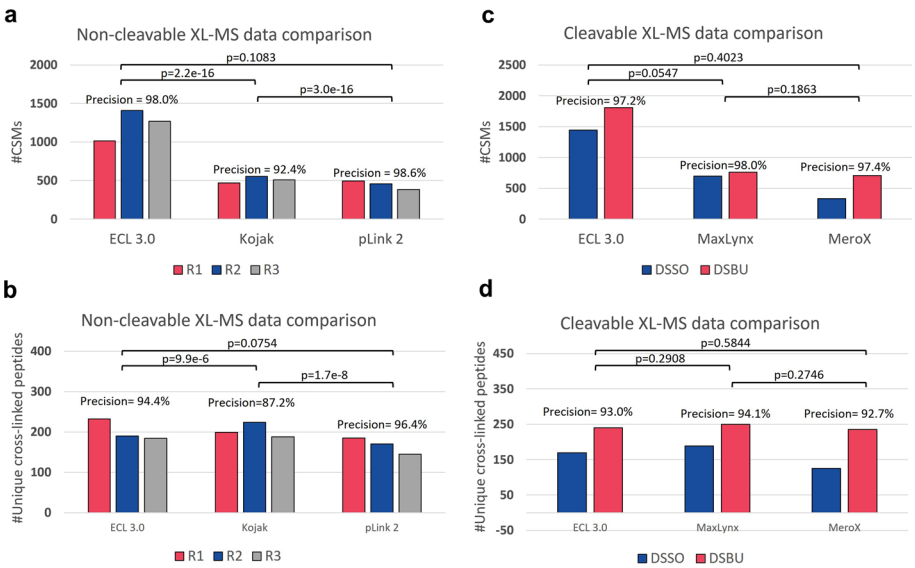

Fig. 2 Performance comparison results. **a**, **b** The performance of ECL 3.0, Kojak, and pLink 2 in non-cleavable data analysis using three technical replicates (R1, R2, and R3) of a synthetic peptide sample (PXD014337) [15], which uses disuccinimidyl suberate (DSS) as the cross-linker. The number of cross-linked peptide spectrum matches (CSMs), the number of unique cross-linked peptides and precision (TP/TP+FP) were calculated for each software tool. ECL 3.0 identified over twice as many results as other tools in the CSMs number and similar results in the unique cross-linked peptides. Using the (one-sided) Fisher's exact test, the *p*-value indicated that ECL 3.0 has significantly better precision than Kojak and has similar precision as pLink 2. **c**, **d** Identification results of ECL 3.0, MaxLynx, and MeroX using cleavable data. Results using two different cross-linkers (disuccinimidyl sulfoxide (DSSO) and disuccinimidyl dibutyric urea (DSBU)) are shown. ECL 3.0 identified over twice as many results as other tools in CSMs numbers and similar results in the unique cross-linked peptides. The precision of these three tools does not have a significant difference according to the (one-sided) Fisher's exact test (*p*-value)

Details of the parameter settings and further experimental comparisons can be found in the Additional file 1, where we further utilized two E. coli data sets [12, 16] and four human data sets [17–20].

Conclusion

ECL 3.0 is a comprehensive cross-linking mass spectrometry data analysis tool. It enables us to unravel more intriguing PPIs and protein structure conformations by using the protein feedback mechanism. ECL 3.0 is available with GUI version and is open source at https://github.com/yuweichuan/ECL-PF.git.

Availability and requirements

Project name: ECL 3.0

Project home page: https://github.com/yuweichuan/ECL-PF.git

Project alternative page: https://zenodo.org/record/8176558

Operating system(s): Windows.

Programming languages: Python and C++.

Other requirements: Python 3.6 or higher.

License: MIT

Any restrictions to use by non-academics: None

Abbreviations

XL-MS	Cross-linking mass spectrometry
PPIs	Protein–protein interactions
MS/MS	Tandem mass spectrum
CSMs	Cross-link spectrum matches
NMR	Nuclear magnetic resonance
CID	Collision-induced dissociation
FDR	False discovery rate
GUI	Graphical user interface
DSS	Disuccinimidyl suberate
DSSO	Disuccinimidyl sulfoxide
DSBU	Disuccinimidyl dibutyric urea

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05473-z.

Additional file 1. Further comparisons, validations, and supplementary figures.

Acknowledgements

We thank the friendly reminder from one anonymous reviewer during the submission of our earlier paper [4] that the protein feedback idea can be extended to non-cleavable cross-linking peptide identification. We also thank two reviewers for constructive suggestions which have helped us to improve the manuscript.

Author contributions

CZ proposed the protein feedback concept and designed the tool. SD performed data analysis. SD, SL and XZ optimized the algorithm. CZ and YL developed the graphical user interface (GUI). CZ, NL and WY drafted the manuscript. NL and WY supervised the work.

Funding

This work was supported in part by T12-101/23-N, R4012-18, C6021-19EF, 16102422, 16103621, 16101920,16101819, and 16306919 from the Research Grant Council (RGC) of the Hong Kong S.A.R. government of China, MHP/033/20 from the Innovation and Technology Commission (ITC) of the Hong Kong S.A.R., the Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone project (HZQB-KCZYB- 2020083), the internal grants 3030-009 and BGF.001.2023 from HKUST.

Availability of data and materials

All the data sets used in the comparison can be found on the public proteomic database (PRIDE Archive) at https://www.ebi.ac.uk/pride/archive/. The source code is available at https://github.com/yuweichuan/ECL-PF.git.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
W.Y. is an Associate Editor of BMC Bioinformatics. All authors declare that they have no competing interests.

Received: 2 June 2023 Accepted: 11 September 2023
Published online: 20 September 2023

References

1. Yu C, Huang L. Cross-linking mass spectrometry (xl-ms): an emerging technology for interactomics and structural biology. Anal Chem. 2018;90:144–65.
2. Trnka MJ, Baker PR, Robinson PJ, Burlingame A, Chalkley RJ. Matching cross-linked peptide spectra: only as good as the worse identification. Mol Cell Proteom. 2014;13:420–34.
3. Iacobucci C, Sinz A. To be or not to be? five guidelines to avoid misassignments in cross-linking/mass spectrometry. Anal Chem. 2017;89:7832–5.
4. Zhou C, Dai S, Lin Y, Lian S, Fan X, Li N, Yu W. Exhaustive cross-linking search with protein feedback. J Proteom Res. 2023;22:101–13.
5. Steigenberger B, Albanese P, Heck A, Scheltema R. To cleave or not to cleave in xl-ms? J Am Soc Mass Spectrom. 2019;31:196–206.
6. Iacobucci C, Piotrowski C, Aebersold R, Amaral BC, Andrews P, Bernfur K, Borchers C, Brodie NI, Bruce JE, Cao Y, et al. First community-wide, comparative cross-linking mass spectrometry study. Anal Chem. 2019;91:6953–61.
7. Yu F, Li N, Yu W. ECL: an exhaustive search tool for the identification of cross-linked peptides using whole database. BMC Bioinform. 2016;17:217–24.
8. Yu F, Li N, Yu W. Exhaustively identifying cross-linked peptides with a linear computational complexity. J Proteom Res. 2017;16:3942–52.
9. Dai J, Jiang W, Yu F, Yu W. Xolik: finding cross-linked peptides with maximum paired scores in linear time. Bioinformatics. 2019;35:251–7.
10. Walzthoeni T, Claassen M, Leitner A, Herzog F, Bohn S, Förster F, Beck M, Aebersold R. False discovery rate estimation for cross-linked peptides identified by mass spectrometry. Nat Methods. 2012;9:901–3.
11. Hoopmann MR, Zelter A, Johnson RS, Riffle M, MacCoss MJ, Davis TN, Moritz RL. Kojak: efficient analysis of chemically cross-linked protein complexes. J Proteom Res. 2015;14:2190–8.
12. Chen Z-L, Meng J-M, Cao Y, Yin J-L, Fang R-Q, Fan S-B, Liu C, Zeng W-F, Ding Y-H, Tan D, et al. A high-speed search engine plink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. Nat Commun. 2019;10:3404–15.
13. Yılmaz S, Busch F, Nagaraj N, Cox J. Accurate and automated high-coverage identification of chemically cross-linked peptides with maxlynx. Anal Chem. 2022;94:1608–17.
14. Iacobucci C, Götze M, Ihling CH, Piotrowski C, Arlt C, Schäfer M, Hage C, Schmidt R, Sinz A. A cross-linking/mass spectrometry workflow based on ms-cleavable cross-linkers and the merox software for studying protein structures and protein–protein interactions. Nat Protoc. 2018;13:2864–89.
15. Beveridge R, Stadlmann J, Penninger JM, Mechtler K. A synthetic peptide library for benchmarking crosslinking-mass spectrometry search engines for proteins and protein complexes. Nat Commun. 2020;11:742–50.
16. Stieger CE, Doppler P, Mechtler K. Optimized fragmentation improves the identification of peptides cross-linked by ms-cleavable reagents. J Proteom Res. 2019;18:1363–70.
17. Graziadei A, Schildhauer F, Spahn C, Kraushar M, Rappsilber J. SARS-CoV-2 Nsp1 N-terminal and linker regions as a platform for host translational shutoff. bioRxiv; 2022.
18. Chen Y, Zhou W, Xia Y, Zhang W, Zhao Q, Li X, Gao H, Liang Z, Ma G, Yang K, et al. Targeted cross-linker delivery for the in situ mapping of protein conformations and interactions in mitochondria. Nat Commun. 2023;14:3882–97.
19. Ryl PS, Bohlke-Schneider M, Lenz S, Fischer L, Budzinski L, Stuiver M, Mendes MM, Sinn L, O'reilly FJ, Rappsilber J. In situ structural restraints from cross-linking mass spectrometry in human mitochondria. J Proteom Res. 2019;19:327–36.
20. Chang Y-G, Lupton CJ, Bayly-Jones C, Keen AC, D'Andrea L, Lucato CM, Steele JR, Venugopal H, Schittenhelm RB, Whisstock JC, et al. Structure of the metastatic factor p-rex1 reveals a two-layered autoinhibitory mechanism. Nat Struct Mol Biol. 2022;29:767–73.

Publisher's Note