Probability-Based Shotgun Cross-Linking Sites Analysis

Young Jin Lee
Ames Laboratory-USDOE and Department of Chemistry, Iowa State University, Ames, Iowa, USA

We recently developed a shotgun tool for cross-linking sites analysis, X!Link, for the sensitive and high-throughput analysis of chemically cross-linked proteins or multiprotein complexes (J. Proteome Res. **2007**, *6*, 3908–3917). Here, we report a further development of the tool using a probability-based scoring system. It calculates explicit E-values, with which sensitive detection of the cross-links is possible with very low false positives, and now can be applied to moderate numbers of protein sequences. Most of the false positives in large scale analysis originate from partial matching where one side of the peptides is correctly matched while the other side is incorrectly matched. Additional E-values were calculated for each peptide and effectively minimized false positives from such partial matching. The usefulness of the new scoring system was demonstrated for a previously published dataset from a cross-linked cytochrome *c* protein, searching against a large database of equine protein sequences. (J Am Soc Mass Spectrom 2009, 20, 1896–1899) © 2009 American Society for Mass Spectrometry

Cross-linking mass spectrometry, chemical cross-linking of intra- or inter-protein close-contacts followed by mass spectrometric determination of cross-linked sites, is a rapidly growing research field and has great potential to supplement X-ray or NMR based structural proteomics [1]. Although the structural information is low-resolution in its nature, it has several advantages superior to the other techniques such as fewer requirements in sample amount and purity, little limitation on protein size, no need for crystallization, and easy application to protein complexes. The current technologies, however, have low sensitivity and throughput, and are not sufficiently advanced to illuminate its hidden potential.

Most of the limitations are attributed to its difficulty in separating cross-linked peptides from other predominant non-cross-linked peptides. Isotopic labeling or comparison with the control has been typically adopted in the search for cross-links [2, 3]. These approaches, based on parent mass spectra to find cross-linked fragments, have limited sensitivity due to the difficulty in comparing low level signals in mass spectra. Recently, we developed a novel method based on a shotgun approach termed X!Link [4]. In this approach, we rely largely on MS/MS spectral interpretation of cross-linked ions rather than the MS level information, which we believe is more efficient, similar to the shotgun approach used in proteomics applications compared with peptide mass fingerprinting [5]. Unlike most others [1–3], we use parent mass spectra only as a primary filtering of candidates, and all candidates are thoroughly evaluated by their MS/MS spectra. We use high mass accuracy precursor spectra to minimize false positives, which is essential because the number of false positives increases rapidly with increasing database size. Our approach, based on high-throughput cross-link search of LC-MS/MS data, showed dramatic improvement in sensitivity and throughput compared with previously reported methods [4].

Here, we report a further development of X!Link using a probability-based scoring system. We use a Poisson distribution to calculate the expected value of random matching, the same approach adopted in open mass spectrometry search algorithm (OMSSA) [6] for single peptide search, but with some modifications. We also implement the calculation of separate E-values for each peptide to avoid the false positive by partial matching, which has been neglected in cross-linking data analysis and could lead to an inadvertent false positive. When we applied the new scoring to a previously published cytochrome *c* cross-linking data, we attained the same sensitivity but without any false positives at E-value cutoffs optimized from a large scale simulation.

Methods

Probability-Based Scoring

We adopt a probability-based scoring system using Poisson distribution and follow OMSSA for most of the equations [6]. For simplification, we combine fragment ions with different charge states into a single theoretical spectrum for charge states up to z-1 with maximum of 4 (z is parent charge state). A probability that a theoretical fragment ion could be matched to any peak in the experimental spectrum by random chance can be calculated as:

Address reprint requests to Professor Y. J. Lee, Ames Laboratory-USDOE and Department of Chemistry, Iowa State University, Ames, IA 50011, USA. E-mail: yjlee@iastate.edu

1044-0305/09/$32.00
doi:10.1016/j.jasms.2009.06.020

Published online July 10, 2009
Received April 24, 2009
Revised June 29, 2009
Accepted June 29, 2009

$$\frac{\text{sum of m/z ranges occupied by each experimental peak}}{\text{m/z scan range in experimental spectrum}} \approx \frac{v \times 2t}{m} \quad (1)$$

Here, t, m, and v represent mass tolerance, m/z scan range (high m/z–low m/z), and the number of peaks in the experimental spectrum. The average number of random matching, μ, between experimental and theoretical peaks can be calculated by multiplying the random matching probability, $2tv/m$, with the number of theoretical peaks, h:

$$\mu = \frac{2tvh}{m} \quad (2)$$

The Poisson distribution for x number of matches and μ mean value is given by

$$P(x, \mu) = \frac{\mu^x}{x} e^{-\mu} \quad (3)$$

OMSSA adopts an additional constraint to increase the specificity by accepting the matching results only when at least one of the top n ($n = 3$ by default) peaks is detected. In our previous paper [4], we adopted a similar but more generalized approach requiring at least l major peaks among the top n peaks. The optimum values for l and n were suggested as 3 and 10, respectively. The probability distribution is adjusted for this additional filtering [6]:

$$P^{l\prime}(x, \mu) = \frac{1}{Q}(1 - (1 - q)^x)P(x, \mu) \quad (4)$$

where the normalization factor Q is

$$Q = \sum\nolimits_x (1 - (1 - q)^x)P(x, \mu) \quad (5)$$

Here, q is the probability to have l calculated peaks among the top n peaks and can be calculated as,

$$q = \frac{{}_nC_l}{{}_vC_l} \quad (6)$$

where ${}_nC_l$ is the number of l combinations from a set of n, defined as $n!/((n - l)!l!)$.

According to our calculations, major peak filtering of at least one peak out of the top three ($l = 1$ and $n = 3$) and three peaks out of the top 10 ($l = 3$ and $n = 10$) have almost the same sensitivity, but they exclusively detect 1%~2% more cross-links than each other (data not shown). For the best sensitivity, both major filterings are included in the algorithm: X!Link tests whether there is at least one peak out of the top three and, if it fails, it tests the existence of at least three peaks out of the top 10.

The E-value can be calculated as following:

$$E(y, \mu) = N(1 - (\sum\nolimits_{x=0}^{y-1} P^{l\prime}(x, \mu))^N) \quad (7)$$

Here, N is the number of all the possible cross-links that has passed the primary ion filtering and y is the number of successful product ion matches for the particular cross-link. When fragmentations from one of the two peptides dominate the MS/MS spectrum, a low E-value could be obtained even when the other peptide is a random match. To avoid such false positives, additional E-values are calculated for each peptide. The same eq 2 and eq 3 are used but with h and x corresponding to fragments from only one peptide, and major peak filtering is omitted, namely q = 1 in eq 4 and eq 5. To distinguish these E-values, the E-value for the cross-link is denoted as E_X and E-values for each peptide are defined as E_α and E_β. Typically, E_α and E_β are higher than E_X because the fragment ions corresponding to the other counter peptide are regarded as noise.

Results and Discussion

Application to Cytochrome c *Cross-Link Data*

We applied the new X!Link algorithm to the previously reported cytochrome *c* cross-link data [4]. In short, this data were generated from a 2-h gradient LC-MS/MS with LTQ-FT (Thermo Finnigan, San Jose, CA) for a trypsin digest of BS3 (Bis[sulfosuccinimidyl] suberate)-cross-linked cytochrome *c*. The Mascot generic file (Mgf) deisotoped by Mascot distiller (Matrix Science) was used as an input file for X!Link. The X!Link output is compiled into a short list by sorting out multiple assignments, combining the same cross-links, and manual validation. E-value filtering was made following the suggestion in the next section ($E_X < 0.03$ and $E_\alpha/E_\beta < 0.3$). Supplementary Table 1, which can be found in the electronic version of this article, summarizes all unambiguously identified cross-links thus assigned. The sensitivity was not affected significantly by the new scoring system, but no false positives were found with the given filtering parameters. All cross-links have Lys inter-α-carbon distances less than 24 Å (the length of fully expanded BS3 and two Lys side chains) and were mostly less than 20 Å. There are some changes from the previous report: a total of 25 inter-peptide cross-links in 215 MS/MS spectra were found compared with the previous 21 cross-links in 180 spectra. The differences mostly resulted from neglecting the intra-cross-linked heme group in the previous report. The changes are discussed in detail in Supplementary Information 3.

It has been suggested that a NHS (N-hydroxysulfosuccinimide) containing cross-linker, like BS3 that we used for cytochrome *c* cross-linking, might cross-link Tyr or Ser as side reactions in addition to the primary amines [7]. We performed an X!Link search of the same cytochrome *c* cross-link data for possible cross-linking of Lys-Tyr and Lys-Ser. No Lys-Ser cross-links were detected, while 85 possible cases were detected for Lys-Tyr cross-links. How-

ever, all the potential Lys-Tyr cross-links were confirmed to be incorrect after careful manual inspection, and most of them are mis-assignment of Lys-Lys cross-links. Even though most of the incorrect Lys-Tyr cross-links have higher E-values than Lys-Lys cross-links, some have lower E-values due to accidental random matching of some of the fragments. Therefore, it is always very important to manually validate the exact cross-linking sites. No Lys-Ser or Lys-Tyr cross-links were identified, suggesting that non-primary amine cross-linking products by the NHS cross-linker is absent or minimal at least in our experimental conditions.

Simulation of a Complex Protein Mixture

X!Link was originally designed for one or two protein sequence(s) [4]. The new probability-based scoring system, however, is independent of the charge state and database size (the database size effect is included in N in eq 7) allowing its application to multiple protein sequences. The possible application to a complex protein mixture was tested by searching the cytochrome *c* cross-link dataset against two additional protein sequence databases, a database with 346 equine protein sequences (supplied by Thermo Finnigan for BioWorks 3.2) and a database of the reversed sequences. The E-value, the chance of a random matching, is expected to be very low for most of the true positives. Figure 1a compares the old and new scores for the false positives obtained from X!Link search against the reversed equine sequence (decoy database). In total, 22,712 cross-linked peptides passed the old filtering of 1.0; however, only four could pass the new filtering of E_X less than 0.03. It clearly demonstrates the effectiveness of filtering false positives with the new scoring system. Figure 1b shows the number of matching cross-links as a function of E_X cutoff for the cytochrome *c* sequence, the equine database, and the decoy database of reversed equine sequences. As expected, the search against the true protein sequence (cytochrome *c*) gives mostly low E_X values, while the search against the false database (reversed equine sequence) gives high E-values with no matching below 0.001. In a typical proteomics analysis, the cutoff value is often determined from the search against the reversed sequence, i.e., 0.001 in this case. The equine database (forward equine sequence) contains both true (cytochrome *c*) and false (other equine proteins) sequence(s), and the search result includes the distribution of both true and false positives as shown separately in (**b**). The true positive from the equine sequence has the similar distribution with that against cytochrome *c* sequence, but with a little higher E-values resulting from the larger value of N in eq 7. In contrast to the reversed equine database, some of false positives from equine database have low E-values down to an E_X cutoff of 1×10^{-10}. It indicates that the cutoff cannot be simply determined from the search against the reversed sequences, unlike single peptide search in proteomics. As can be seen in Supplementary Table 2, these false positives are mostly coming from partial matching: one peptide is correct but the other one results from incorrect random matching. We could solve this problem by adding additional filtering in each peptide, E_α and E_β. When we applied E_α/E_β filtering of 0.3 or 0.1 ((**b**) and Supplementary Table 2), the false positive was dramatically removed and there was no false positives with E_X of 0.0001 or below.

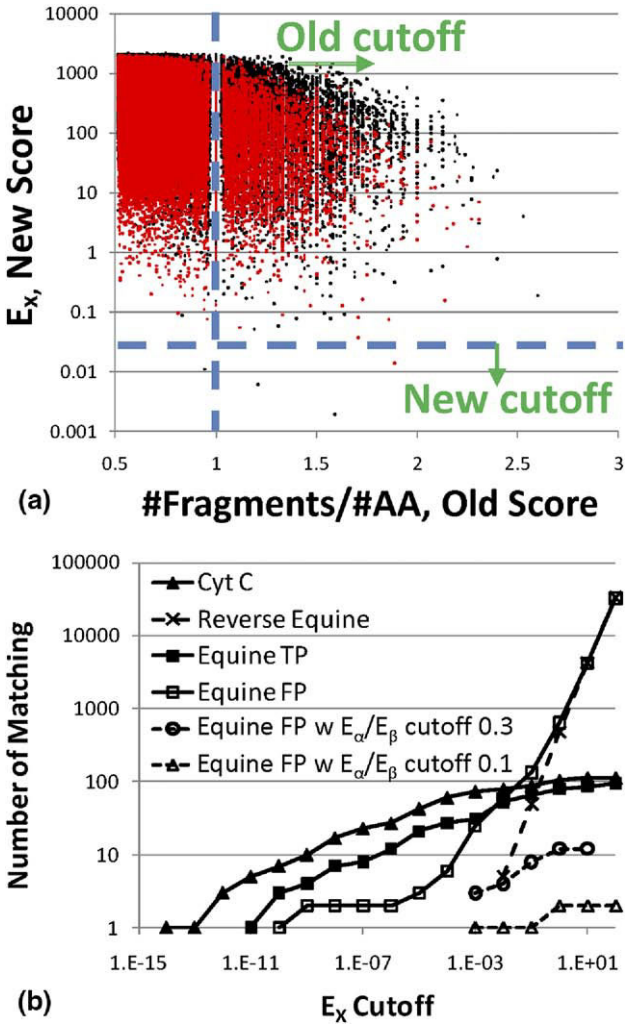

Figure 1. (**a**) Effective removal of false positives with the new score. Comparison between the old score (the number of fragments/the number of amino acids) and the new score (E_X, E-value for cross-link) for the X!Link search result of cytochrome *c* cross-link data against the decoy database of reversed equine sequences. Red and black dots denote the charge state of 3 and 4, respectively. (**b**) False positives by partial matching and their removal with E-value for each peptide. The number of matched cross-linked ions are shown as a function of E_X cutoff for the X!Link search result of the cytochrome *c* cross-link data against cytochrome *c* sequence (Cyt C), equine protein database (Equine), and the decoy database of reversed equine sequences (Reverse Equine). True (TP) and false positive (FP) distributions are separated for the equine database search. False positive in equine database search has low E_X value down to 1×10^{-10} due to partial matching, which is effectively removed with additional filtering at each peptide level (E_α/E_β less than 0.3 or 0.1). Only unique cross-linked ions are counted in the figure.

Sensitivity and selectivity were calculated for both cytochrome *c* and equine database at various E_X and

E_α/E_β cutoffs. Figure 2 shows an example calculated at E_α/E_β cutoff of 0.3. The best optimization was made at the E_X and E_α/E_β cutoff of 0.03 and 0.3, respectively, where the sensitivity is 100 and 28% for cytochrome *c* and equine sequence(s), respectively, and the selectivity is still 88% for equine sequence. Therefore, we believe we have almost 100% sensitivity and selectivity for a small protein like cytochrome *c*. When searching against a larger protein or multiple protein sequences, the same criteria might be used to minimize false negatives but careful manual inspection needs to be performed for the boundary score cross-links; i.e., those with 0.0001~0.03 for E_X and/or 0.03~0.3 for E_α/E_β.

Conclusion

A reliable scoring system is developed for the shotgun cross-linking site analysis with statistical evaluation of random matching. The new scoring system has two major improvements in cross-linking analysis. First, it dramatically removes false positives when applied to a large number of protein sequences, demonstrating its applicability to a complex mixture. It still may not be directly used for a very large protein database; however, we can reduce the number of proteins by single peptide searching using common proteomics tools and then perform cross-link searching against only those identified proteins. Second, we developed a novel concept of calculating E-values for each peptide and significantly removed false positives induced by partial matching. It should be noted that a short peptide will give high E_α/E_β values and may not pass the filtering criteria. For the same reason, short peptide sequences cannot be used to identify proteins in a large protein database search. For a small protein or for only a few protein sequences, however, a short peptide could still survive the filtering and the corresponding cross-links could be detected. For example, some KK and GKK peptides give E_α/E_β value lower than 0.3 and could be detected in our X!Link search against cytochrome *c* sequence (Supplementary Table 1). The new algorithm will be especially useful for the cross-linking studies of protein mixtures, including protein–protein interactions in multi-protein complex and in vivo cross-linking site analysis. Such applications involve very complex samples and low abundant cross-linked peptides. The solution should come from not only the improvement in the algorithm, but also in the experimental approaches, such as exclusion of low charge state ions in MS/MS data acquisition [3, 8, 9], SCX enrichment [8], and partial blocking of reactive Lys sites [10]. Our new program is available by request.

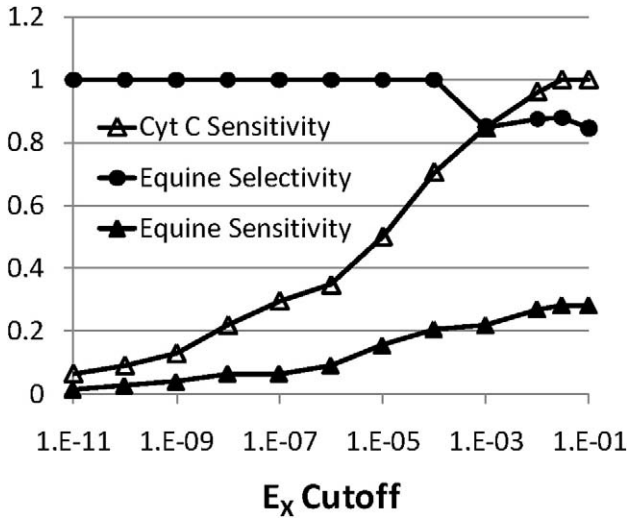

Figure 2. Sensitivity and selectivity as E_X cutoff for the X!Link search against cytochrome *c* sequence and equine database with $E_\alpha/E_\beta < 0.3$. Sensitivity and selectivity are defined as TP/(TP + FN) and 1-FP/(TP + FP), respectively, where TP, FN, and FP are true positive, false negative, and false positive, respectively. The total number of TP was assumed as 78, the number of unique cross-linked ions at E_X cutoff of 0.03 and E_α/E_β cutoff of 0.3 in the search against cytochrome *c* sequence.

Acknowledgments

The author acknowledges partial supports for this work by grants from Iowa State University and Ames Laboratory-USDOE. The Ames Laboratory is operated for the United States Department of Energy by Iowa State University under contract no. DE-AC02-07CH11358.

Appendix A
Supplementary Material

Supplementary material associated with this article may be found in the online version at doi:10.1016/j.jasms.2009.06.020.

References

1. Sinz, A. Chemical Cross-Linking and Mass Spectrometry to Map Three-Dimensional Protein Structures and Protein–Protein Interactions. *Mass Spectrom. Rev.* **2006,** *25*, 663–682.
2. Huang, B. X.; Kim, H. Interdomain Conformational Changes in Akt Activation Revealed by Chemical Cross-linking and Tandem Mass Spectrometry. *Mol. Cell Proteom.* **2006,** *5*, 1045–1053.
3. Seebacher, J.; Mallick, P.; Zhang, N.; Eddes, J. S.; Aebersold, R.; Gelb, M. H. Protein Cross-Linking Analysis Using Mass Spectrometry, Isotope-Coded Cross-Linkers, and Integrated Computational Data Processing. *J. Proteome Res.* **2006,** *5*, 2270–2282.
4. Lee, Y. J.; Lachner, L.; Nunnari, J.; Phinney, B. S. Shotgun Cross-Linking Analysis for Studying Quaternary and Tertiary Protein Structures. *J. Proteome Res.* **2007,** *6*, 3908–3917.
5. Eng, J. K.; McCormack, A. L.; Yates, J. R. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom.* **1994,** *5*, 976–989.
6. Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open Mass Spectrometry Search Algorithm. *J. Proteome Res.* **2004,** *3*, 958–964.
7. Swaim, C. L.; Smith, J. B.; Smith, D. L. Unexpected Products from the Reaction of the Synthetic Cross-Linker 3,3′-Dithiobis(Sulfosuccinimidylpropionate), DTSSP with Peptides. *J. Am. Soc. Mass Spectrom.* **2004,** *15*, 736–749.
8. Rinner, O.; Seebacher, J.; Walzthoeni, T.; Mueller, L.; Beck, M.; Schmidt, A.; Mueller, M.; Aebersold, R. Identification of Cross-Linked Peptides from Large Sequence Databases. *Nat. Methods* **2008,** *5*, 315–318.
9. Singh, P.; Shaffer, S. A.; Scherl, A.; Holman, C.; Pfuetzner, R. A.; Freeman, T. J. L.; Miller, S. I.; Hernandez, P.; Appel, R. D.; Goodlett, D. R. Characterization of Protein Cross-Links via Mass Spectrometry and an Open-Modification Search Strategy. *Anal. Chem.* **2008,** *80*, 8799–8806.
10. Guo, X.; Bandyopadhyay, P.; Schilling, B.; Young, M. M.; Fujii, N.; Aynechi, T.; Guy, R. K.; Kuntz, I. D.; Gibson, B. W. Partial Acetylation of Lysine Residues Improves Intraprotein Cross-Linking. *Anal. Chem.* **2008,** *80*, 951–960.