Fragmentation Pattern-based Scoring for Peptide Identification By Database Search

*Junhee Hong[‡,1], Seungjin Na[2] and Eunok Paek[*1]*

[1]Department of Computer Science, Hanyang University, Seoul, Republic of Korea.
[2]Institute for Artificial Intelligence Research, Hanyang University, Seoul 04763, Republic of Korea.

INTRODUCTION

In mass spectrometry(MS)-based proteomics, most peptide identifications are performed by matching experimental spectra against theoretical spectra obtained from peptide sequences, using search tools like SEQUEST or Comet. The matching algorithms make the intensities of all mass-to-charge peaks equal when constructing theoretical spectra of peptides. Using such a simple theoretical spectra can result in undesirable peptide-spectrum match (PSM) scores, and consequently it may disturb the identification of high-quality PSMs. We analyzed the fragmentation statistics of tandem MS spectra and used the statistics to calculate the similarity score between a peptide sequence and an experimental spectrum.

METHODS

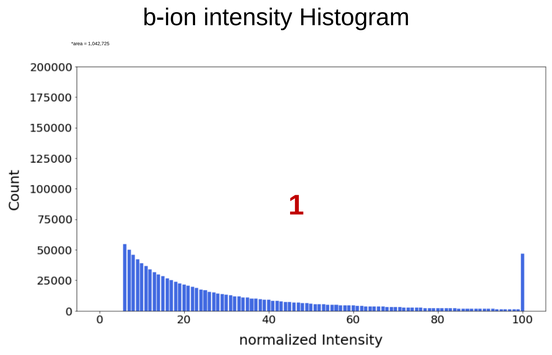
The fragment ion types include b-ion and y-ion, which are dominant fragment ions in tandem MS spectra. After estimating at 1% PSM FDR, we analyzed 2 types of patterns using the PSMs; Pattern1 – the frequency ratio of b-ion and y-ion, Pattern2 – the frequency and the average of intensities according to fragmentation sites. Comet algorithm, a database search tool, was modified to apply the analyzed fragmentation patterns. Three versions were created and their performances were compared with the performance of original Comet in terms of 2 criteria - the number of target PSMs estimated at 1% FDR, and the score distribution of target and decoy PSMs.

RESULTS

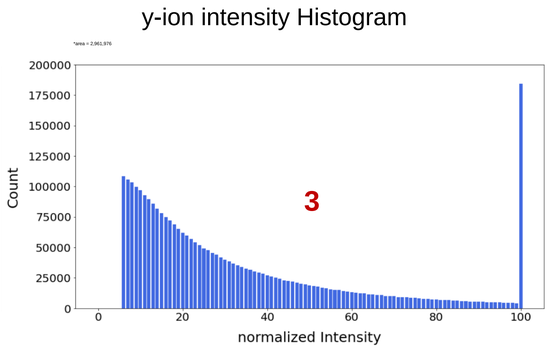
1) Fragmentation patterns
- Q-Exactive tandem MS data of HEK293 cell lysate

Pattern1 The frequency ratio of b-ion and y-ion

b-ion intensity Histogram

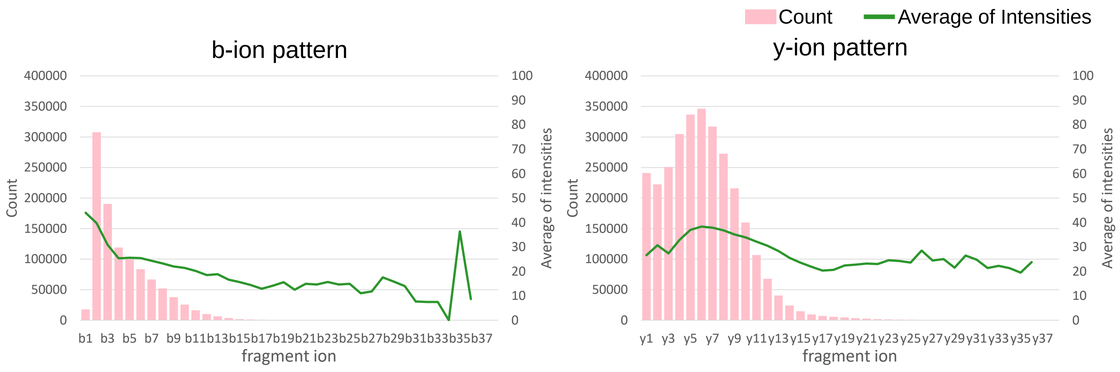


y-ion intensity Histogram

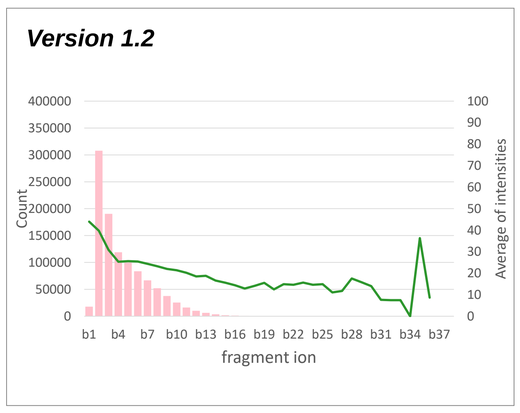


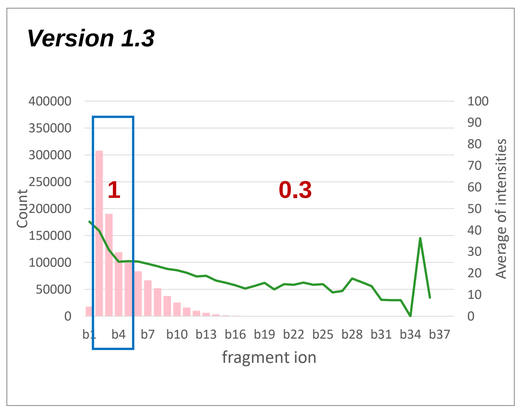
Both histograms above show that the frequency ratio between b-ion and y-ion is about 1:3; the area of b-ion is 1,042,725 and the area of y-ion is 2,961,976. We constructed version 1.1 using the pattern – b-ion : y-ion = 1 : 3

Pattern2 The frequency and the average of intensities by fragment ion's location



It shows the frequency and the average of b- and y- ion intensities. It can be seen that the frequency of the b-ions shows the most prominent pattern. We constructed version 1.2 and version 1.3 using the b-ion's frequency pattern.

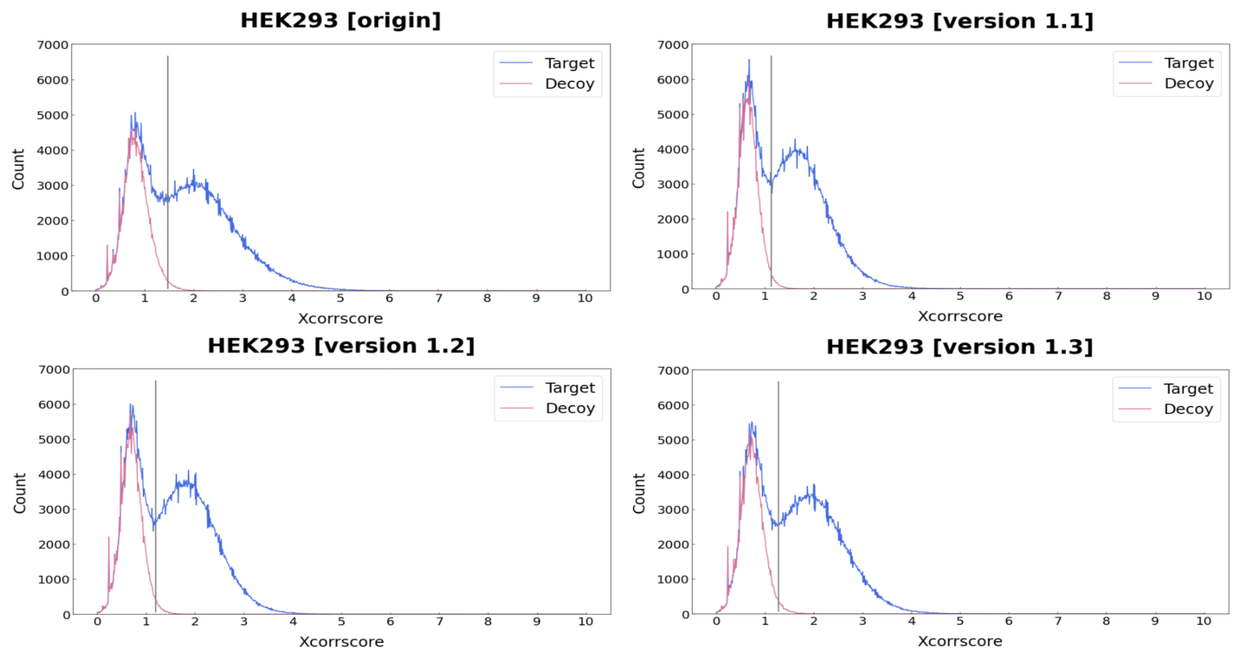




Version 1.2 is a scoring method in which a frequency pattern of all of the b-ions is applied. In version 1.3, the weight from b2 to b5 is 1, while the remaining b-ions are given a weight of 0.3.

An experiment was conducted to output the result by modifying the Comet source code provided as an open source project.

2) Comet Search Results
- HEK293 (Q-Exactive), Adult Lung (Orbitrap-Velos) and Mouse Brain (Q-Exactive)

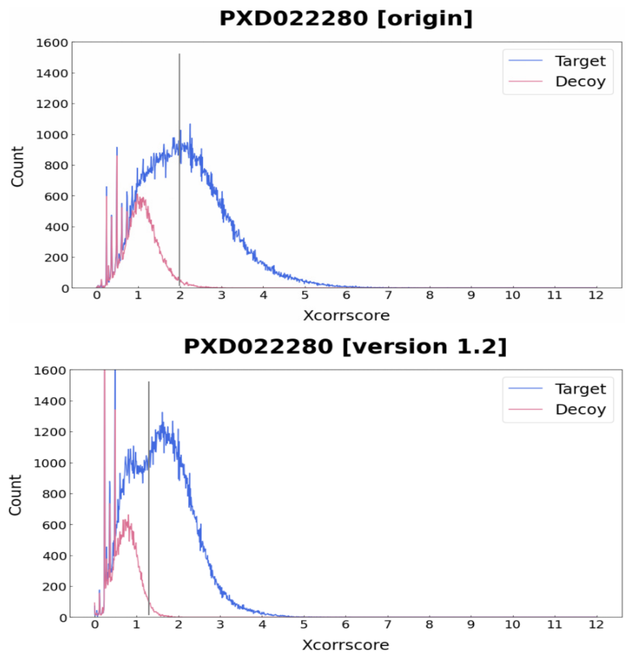


	origin	Version 1.1	*Version 1.2	Version 1.3
Target (fdr 0.01/total)	**481,835** / 840,185	**497,304** / 838,795	**504,560 /** 840,977	**499,191** / 839,649
Decoy (fdr 0.01/total)	4,823 / 269,450	5,275 / 270,817	5,227 / 268,616	5,318 / 269,965

It can be seen that all modified versions distinguish between target and decoy slightly better than the original Comet. Moreover, all modified versions resulted in more target PSMs (all estimated at 1% PSM FDR) than origin. All modified versions have improved the identification performance by about 5% compared to the original Comet. In particular, after estimating at 1% PSM FDR, *version 1.2 identified the most target PSMs (504,560). Although it is not shown, the same experiment was performed on other datasets (Adult Lung, Mouse scans) and the results were similar. We chose *version 1.2 as the final pattern to apply to the database search algorithm.

3) Comet Search Result with Large Database
- Human HCT116 colon cancer cell line (Q-Exactive)

*We search with combination databases of nanopore cDAN-seq and UniProt, splicing isoforms included (1066M)



	origin	Version 1.2
Target (fdr 0.01/total)	**121,208** / 228,113	**146,048** / 234,050
Decoy (fdr 0.01/total)	1,252 / 53,263	1,475 / 47,276

In the score distributions, the version 1.2 separates target and decoy PSMs more clearly. Also, version 1.2 could identify much more target PSMs (146,048) at 1% PSM FDR, leading to about 20% improvement.

CONCLUSIONS

We analyzed the fragment ions from PSMs identified at 1% FDR and generated three statistical models for the fragmentation patterns. The frequency ratio of b- and y-ions is used to construct version 1.1, and the frequency of fragmentation is used to construct version 1.2 and version 1.3. All of the modified Comet scoring algorithm(cross-correlation) resulted in increased identifications compared to the original Comet. Especially, version 1.2 showed the highest improvement of over 10~20% on the number of matched targets compared to the other versions and the original Comet. This research shows that fragmentation patterns could affect the scoring function and thus the result of peptide identification from database search.

REFERENCES

(1) Eng, Jimmy K., Tahmina A. Jahan, and Michael R. Hoopmann. "Comet: an open-source MS/MS sequence database search tool." *Proteomics* 13.1 (2013): 22-24.
(2) Eng, Jimmy K., et al. "A fast SEQUEST cross correlation algorithm." *Journal of proteome research* 7.10 (2008): 4598-4602.
(3) Elias, Joshua E., and Steven P. Gygi. "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry." *Nature methods* 4.3 (2007): 207-214.
(4) Eng, Jimmy K., Ashley L. McCormack, and John R. Yates. "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database." *Journal of the american society for mass spectrometry* 5.11 (1994): 976-989.

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-01373, Artificial Intelligence Graduate School Program(Hanyang University), National Research Foundation of Korea (NRF) grants (2019M3E5D3073568), and by the BK21 FOUR (Fostering Outstanding Universities for Research) project of the National Research Foundation of Korea Grant.