**Proteoform identification using multiplexed top-down mass spectra**

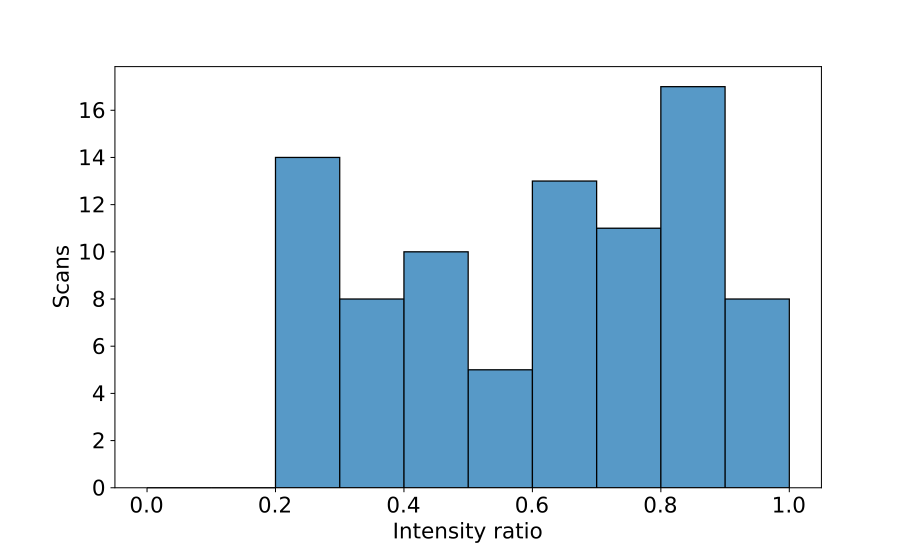
Zhige Wang, Xingzhao Xiong, and Xiaowen Liu

1Deming Department of Medicine, Tulane University, New Orleans, Louisiana, 70112, United States

A diagram of a graph

Description automatically generated with medium confidence

**Fig. 1**: The overview of TopMPI. (a) Primary precursor selection for a multiplexed MS/MS spectrum with two precursors (blue and red) in the MS1 spectrum and the corresponding fragment masses (blue and red) as well as noise ones (black). The red and blue precursors and the fragment masses are identified by database search separately, resulting in a correct identification with red precursor and an incorrect one with the blue precursor. The red precursor is selected as the primary one because its identification has more matched fragment masses. (b) Two rounds of database search. The primary precursor (red) and all the fragment masses in the spectrum are first searched against a protein sequence database for proteoform identification. Then the matched fragments (red) are removed. Finally, the secondary precursor (blue) and the remaining fragment masses are searched against the protein sequence database for proteoform identification.



**Fig. 2**: The distribution of the intensity ratio of the second most intense precursor and the first most intense precursor in the 86 MS/MS spectra with proteoform identifications reported from the *E. coli* data set. The minimum value of the intensity ratio is 0.21.

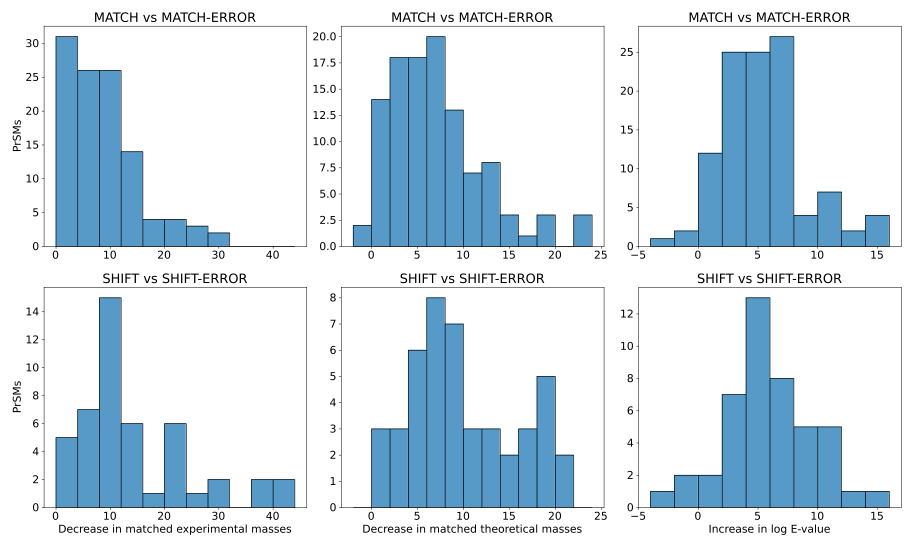
**Fig. 3**: The distribution of the decrease in matched experimental masses, the decrease in matched theoretical masses, and the increase in log base 10 E-value between MATCH and MATCH-ERROR, and between SHIFT and SHIFT-ERROR.



Fig. 4: the search results of the simulated pseudo-multiplexed dataset, labeling the x-axis and y-axis with the logged based 10 E-value of the original non-multiplexed spectra that were retrieved to form the pseudo-multiplexed spectra. The color labels the number of identifications reported for that pseudo-multiplexed spectrum under an E-value cutoff of 0.01.

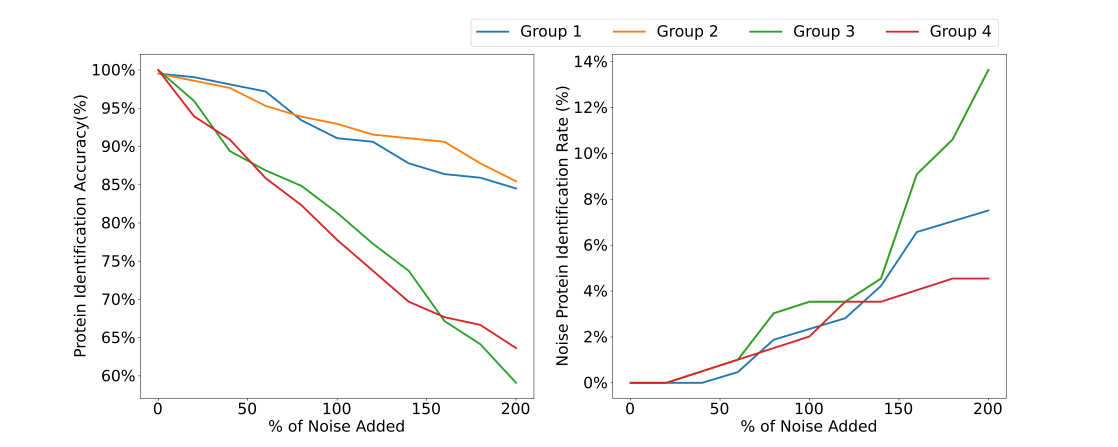


Fig. 5: The above image illustrates the search results of the simulated noised pseudo-multiplexed datasets. The left figure shows the change in the percentage of PrSMs whose protein accession is the same as their base spectrum. And the figure on the right shows the percentage of PrSMs whose protein accession is the same as their noise spectrum.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | # of Spectra | Method | # of Target PrSMs | E-value of Last PrSM |
| Pseudo-Multiplexed | 1095 | Single PrSM | 2173 | 579.82 |
| Pseudo-Multiplexed | 1095 | Separated | 1095 + 1074 = 2169 | 0.52 & 5.64 |
| *E. coli* | 10320 | Single PrSM | 2015 | 0.01 |
| *E. coli* | 10320 | Separated | 2011 + 226 =  2237 | 0.11 & 0.02 |
| Ovarian Cancer | 6875 | Single PrSM | 4450 | 0.07 |
| Ovarian Cancer | 6875 | Separated | 3746 + 716 =  4462 | 0.17 & 0.03 |

Table 1. Performance comparison between the two methods of FDR computation.

|  |  |  |  |
| --- | --- | --- | --- |
| Methods | # of Spectra with Masses Removed | # of Target PrSMs | E-value of Last PrSM |
| Method 0 | 0 | 2011 | 0.11 |
| Method 1 | 226 | 1959 | 0.07 |
| Method 2 | 194 | 1959 | 0.07 |
| Method 3 | 2528 | 1965 | 0.04 |

Table 2. Performance comparison between the four methods of E-value recalculation.