National Institute of Standards and Technology
U.S. Department of Commerce

Mass Spectrometry Data Center

Improved algorithms for identifying phosphopeptides in peptide tandem mass spectral libraries

Sergey L. Sheetlin, Dmitrii V. Tchekhovskoi , Zheng Zhang, Stephen E. Stein
Biomolecular Measurement Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

Overview

- The NIST web-site chemdata.nist.gov contains multiple peptide tandem mass spectral libraries including two phosphopeptide libraries. A tool for searching experimental spectra in these libraries, NIST MSPepSearch, is also available for download.
- This study focuses on improving the performance of MSPepSearch for phosphopeptides identification by utilizing their unique fragmentation behavior and the Hybrid search [1].
- Currently, the suggested method works with non-phosphorylated libraries and can find phosphopeptides with one phosphorylation only.

Introduction

- Tandem mass spectrometry is a widely used method to identify and localize phosphorylation sites in proteins.
- Recently developed Hybrid search is now a part of NIST MSPepSearch.
- The Hybrid search, in effect, introduces in-silico modifications into library spectra and, therefore, can identify phosphorylated peptides using non-phosphorylated libraries or can extend the coverage of phosphorylated spectral libraries.

Methods

- The target-decoy approach is used for evaluating False Discovery Rate (FDR) to make comparison between different methods.
- Decoy libraries are generated by the method published in [2].
- To evaluate performance of the newly developed methods, we used a query dataset containing spectra downloaded from PRIDE repository [3].
- The tests are performed with H. sapiens Orbitrap-HCD and H. sapiens Orbitrap-HCD Phospho libraries available from the NIST web-page [4] (referred as non-phospho and phospho libraries below).

Results

MS/MS Hybrid and MS/MS Direct searches

- MSPepSearch has different operating modes including identity (MS/MS Direct) and similarity (MS/MS Hybrid) searches.
- DeltaMass, defined for each hit, is the difference between precursors of the query and matching library spectra.
- The score threshold is larger for the Hybrid Search than for the Direct search according to Figure 3.
- Therefore, despite the Hybrid search is capable of finding hits produced by the Direct search (with zero DeltaMass), it is less effective for this purpose.

Analog of MS/MS Direct search for phosphopeptides

- The straightforward way of extracting phosphopeptides from the output of the Hybrid search is described below.
- Suppose the Hybrid search is used with a non-phospho library.
- Only hits with the DeltaMass close to $M_{mi}(HPO_3)$ (the monoisotopic mass of phosphorylation) are extracted from the first N hits.
- Figure 1 below compares performance of this method for different N.
- We can see that using extra hits does not significantly improve the performance and even reduces it for small FDR.

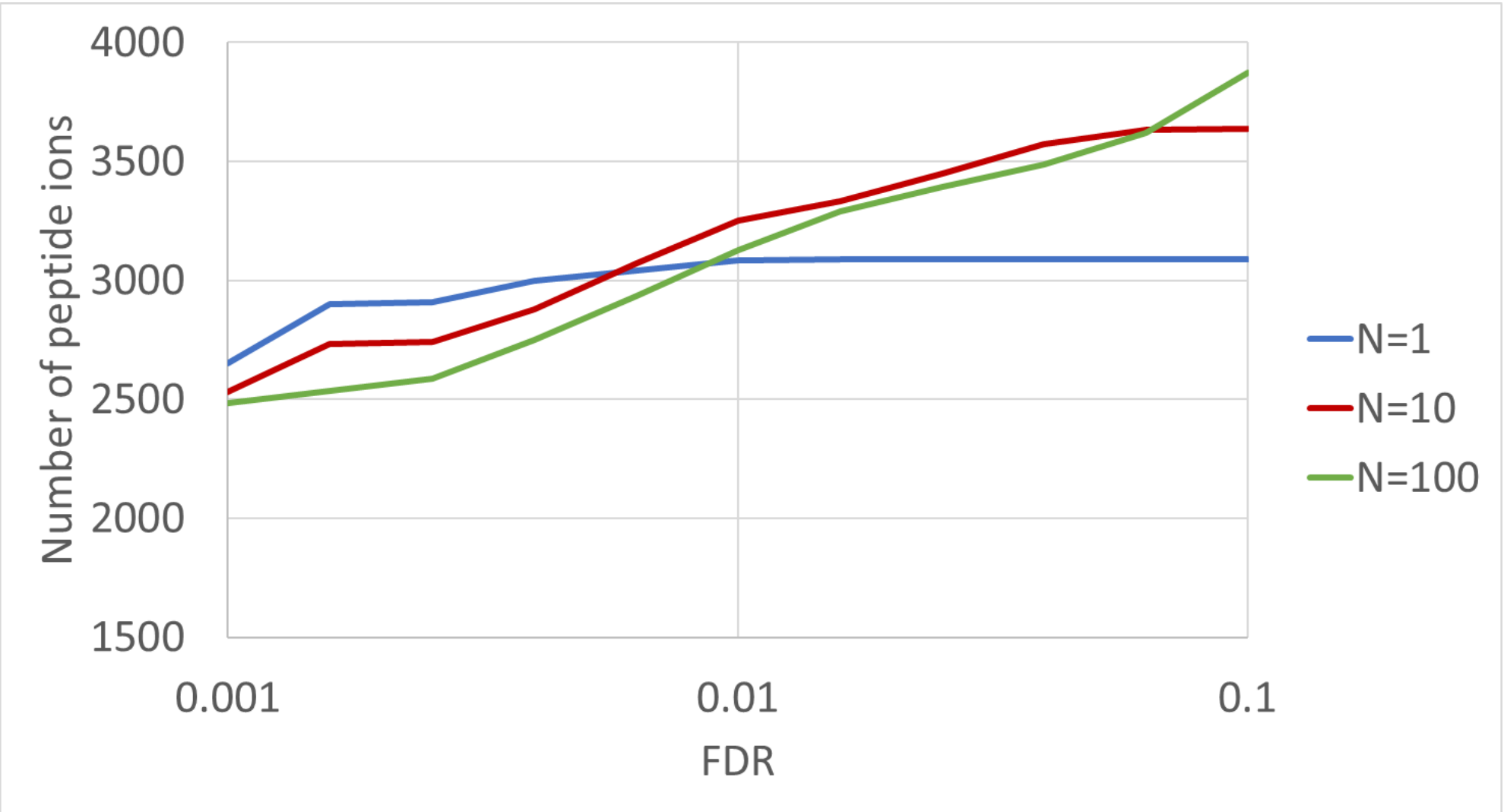

Figure 1. The number of phosphopeptide ions for different number of hits N used in the analog of the MS/MS Direct search

Using a priori information about DeltaMass

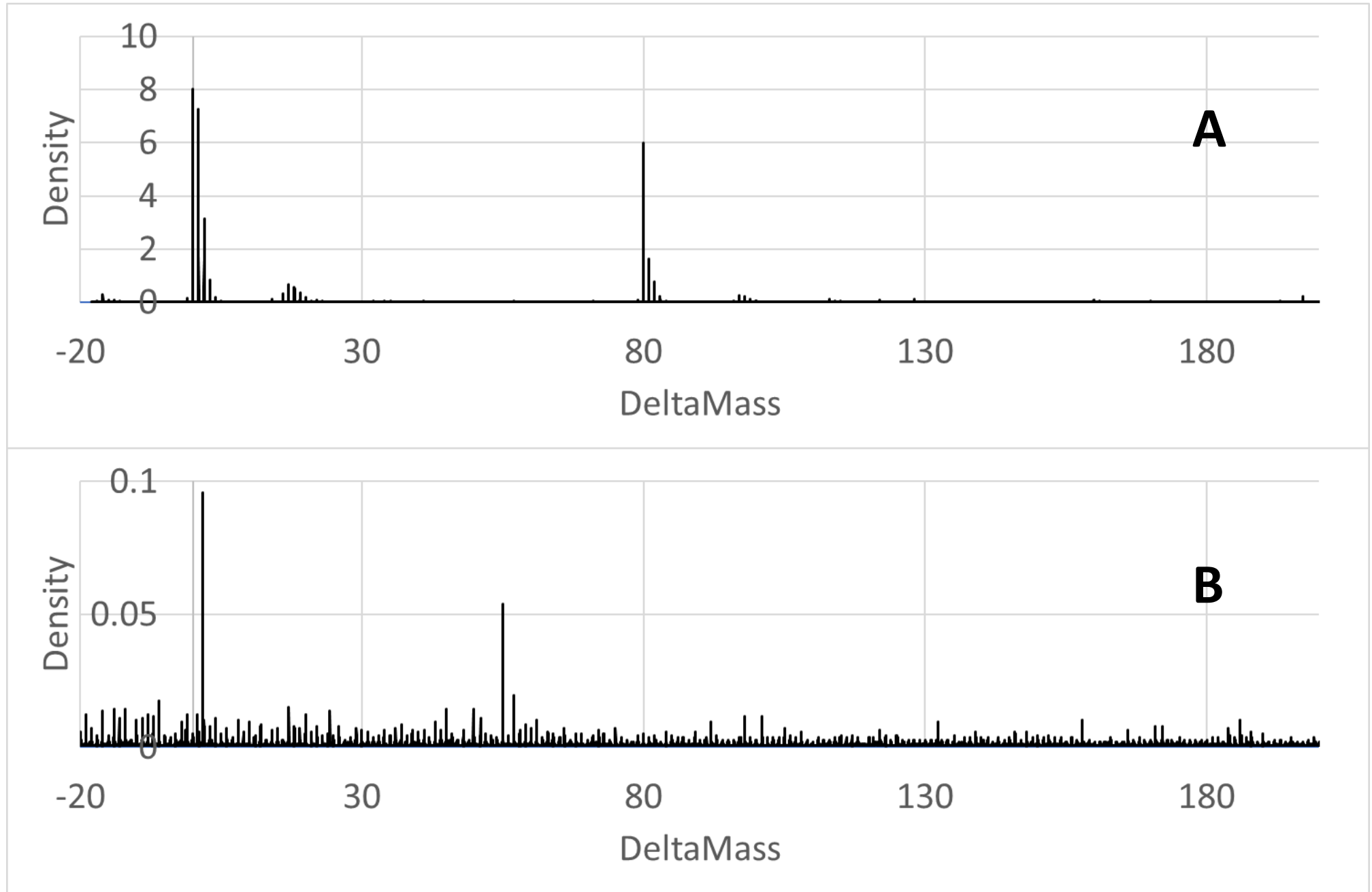

Figure2. Probability density of DeltaMass for high confidence (FDR=0.01) target Hybrid search matches (A) and decoy matches (B).

- We use the Bayesian approach for correcting Hybrid search scores. For this purpose, distributions of DeltaMass were collected for target and decoy matches.
- According to Figure 2A, some DeltaMasses (including zero and $M_{mi}(HPO_3)$) have much higher chance to be observed.
- Figure 3 demonstrates, that using a priory distribution of DeltaMasses significantly improves performance of the Hybrid search especially for small FDR.

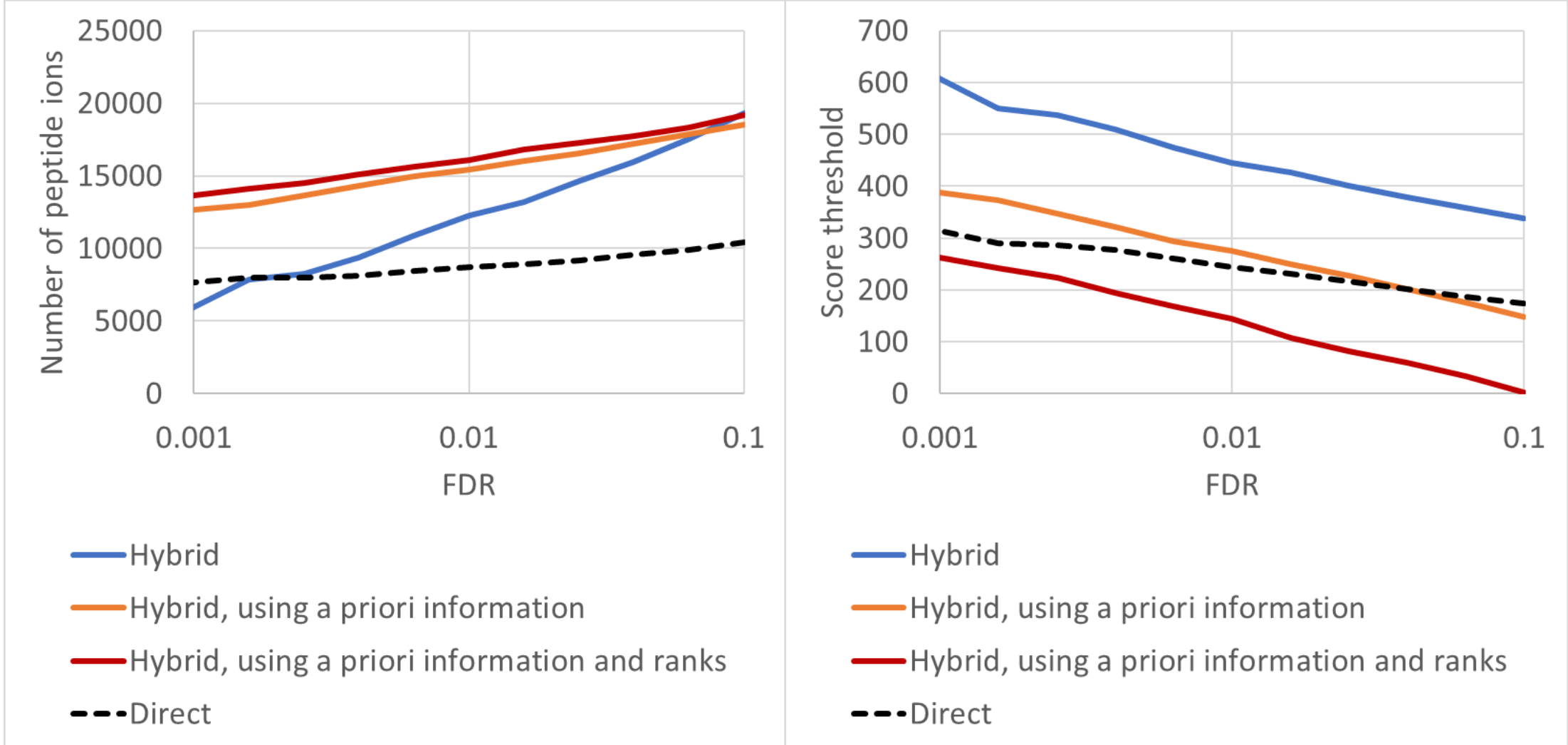

Figure 3. Dependency of the number of peptide ions (not only phosphopeptides) and the score threshold on FDR for the Direct, Hybrid MSPepSearch and its optimizations

Using ranks of MSPepSearch hits

- Rank of a phosphopeptide in the hit list is an additional measure of its reliability and it can be used to improve performance.
- Figure 4 shows that the number of phosphopeptides quickly decreases with rank for the target library but remains about the same for the decoy library.
- If matches of the Hybrid search are penalized based on their rank, then its performance is slightly improved according to Figure 3.

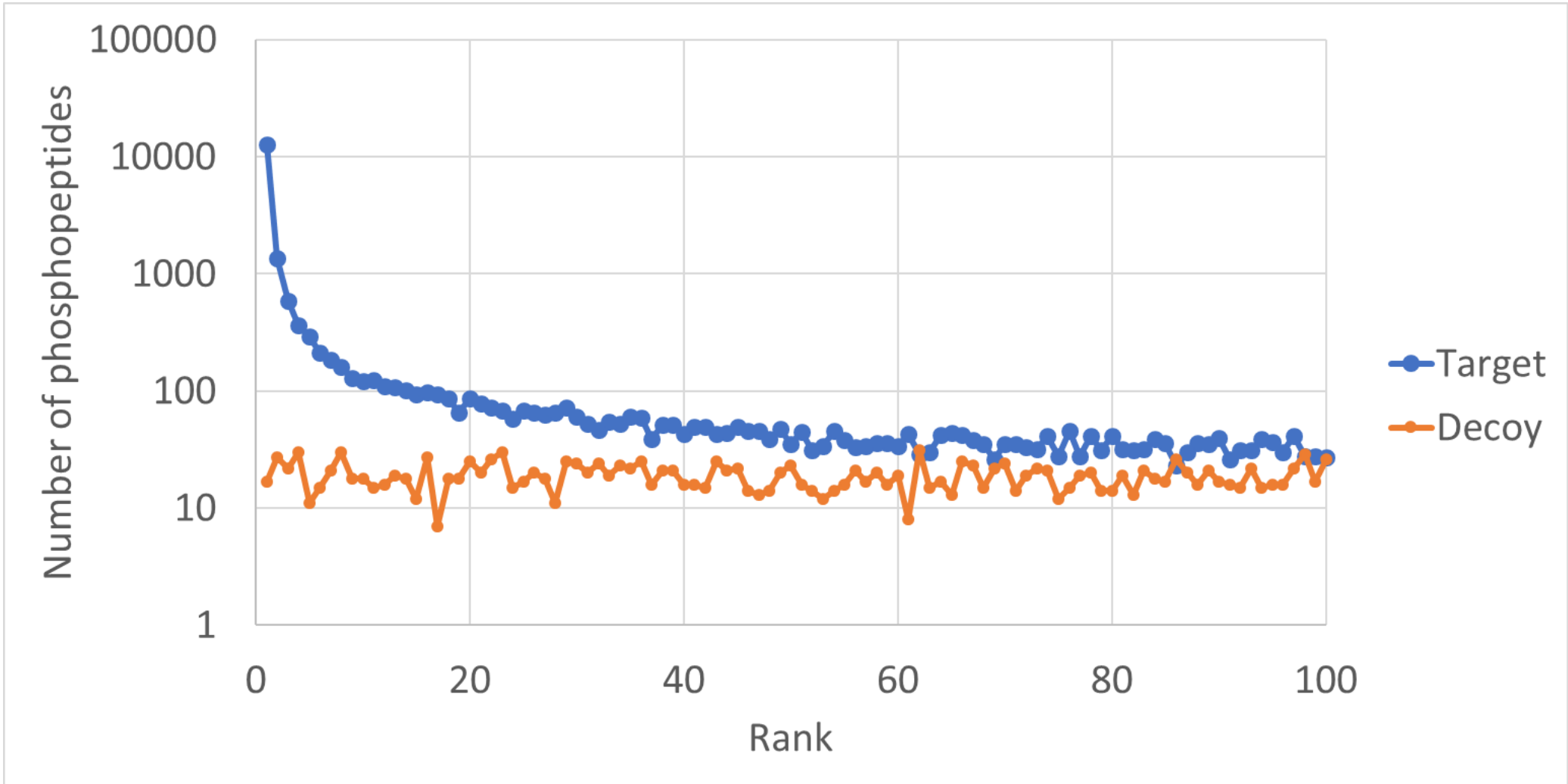

Figure 4. Distribution of phosphopeptides over ranks in the hit list

Fragmentation of phosphopeptides

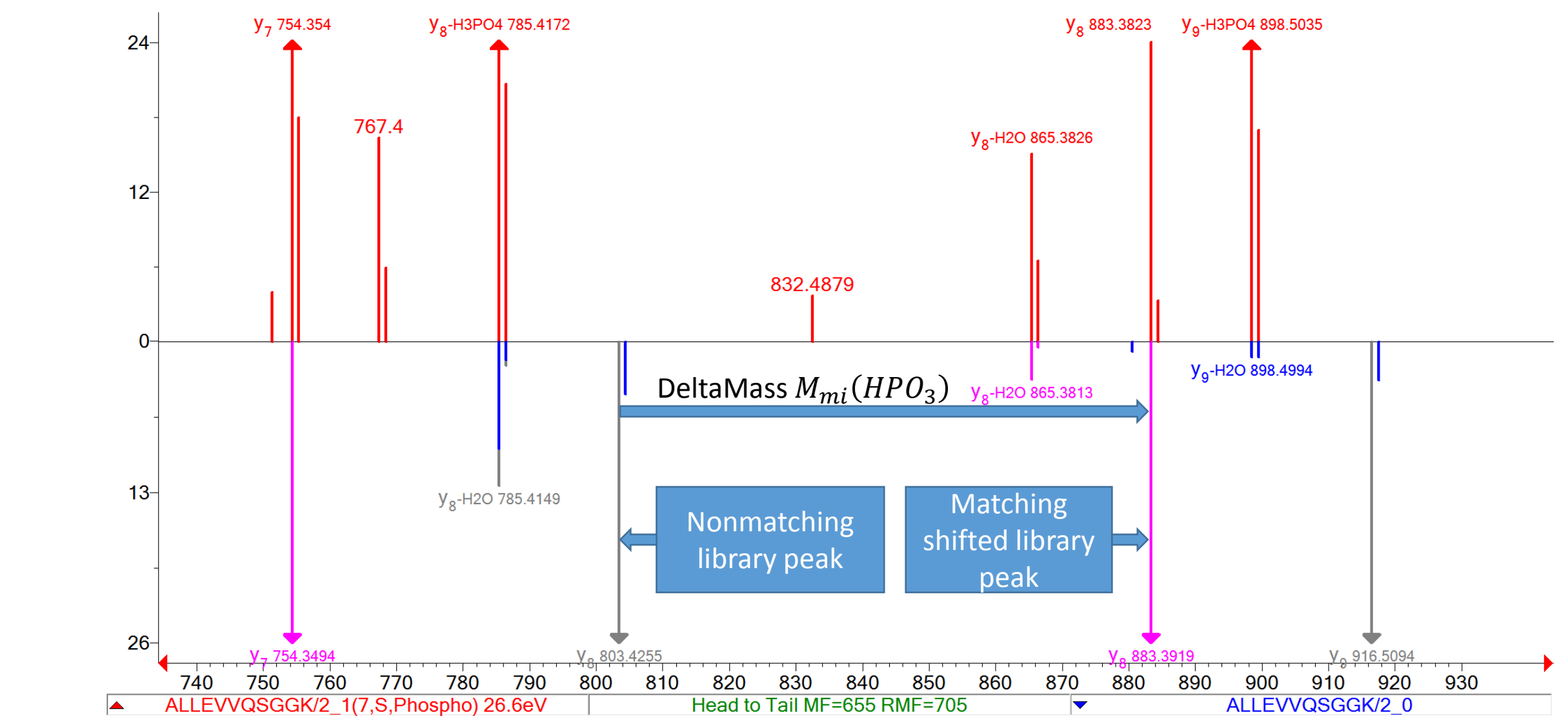

Figure 5. Matching a phosphopeptide spectrum to a non-phospho library spectrum by the Hybrid search which matches the query spectrum to the original and shifted by the DeltaMass library spectra

- Phosphopeptides have certain fragmentation features, in particular, occurrence of large y-ion peaks with the loss of H_3PO_4 is typical (see Figure 5).
- The Hybrid search can be adjusted to make sure that such peaks would contribute to the dot-product accordingly (see the optimized algorithm).

Finding fragmentation patterns

- The fragmentation patterns of phosphopeptides useful for improving the search algorithm, like the presence of large $y - H_3PO_4$ peaks, can be discovered automatically by the method described below.
- Only high-scored matches with DeltaMass close to $M_{mi}(HPO_3)$ are selected.
- For every y-ion library peak with charge 1, distribution of query spectra peaks located within the window $\pm 200mz$ around the y-ion is collected.
- This distribution, shown on Figure 6, helps identify correlations between peaks that can discriminate phosphopeptides in the Hybrid search.

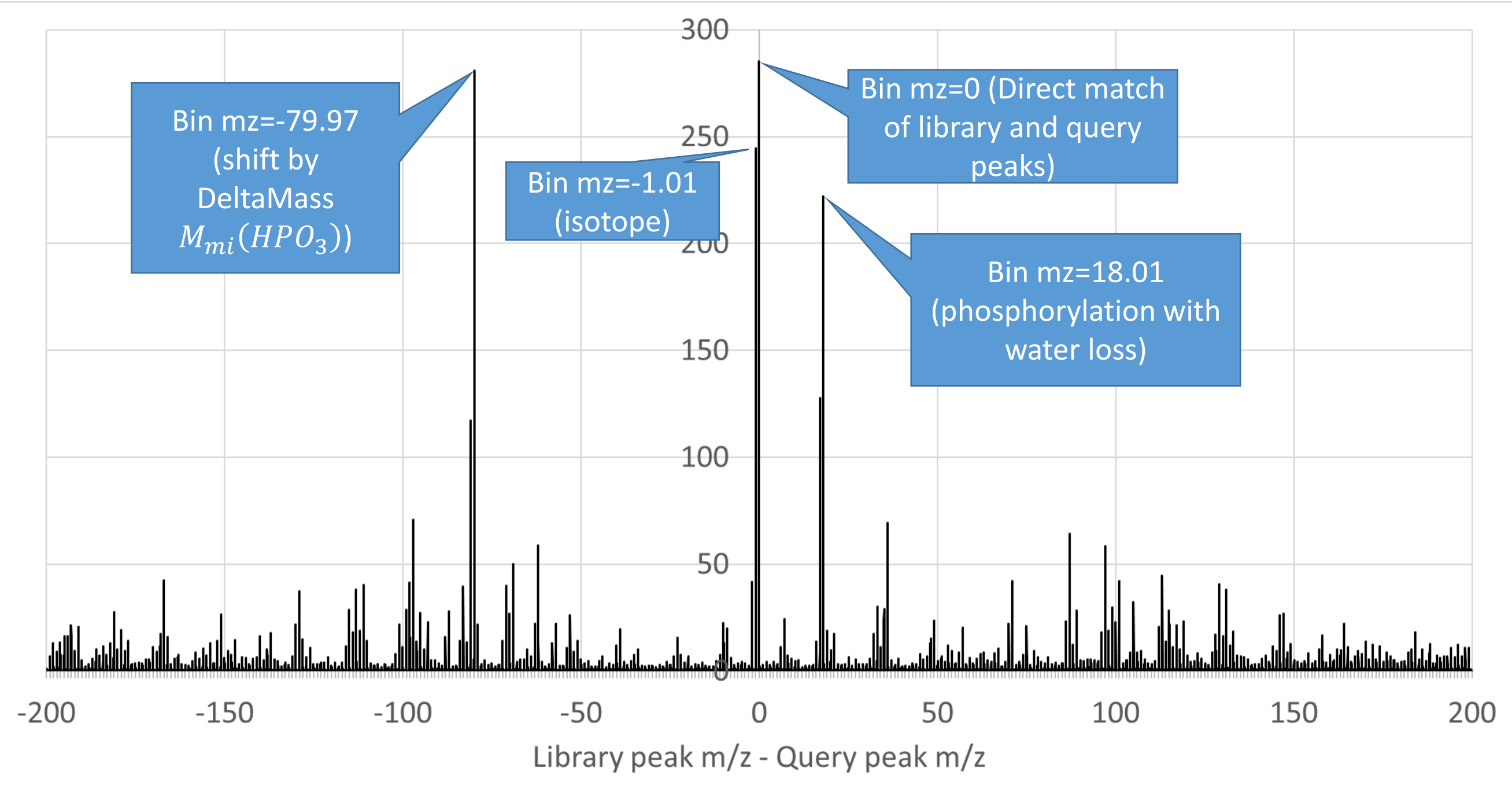

Figure 6. Distribution of peaks of search spectra around y-ions of the matching library spectra

Optimized algorithm for searching phosphopeptides

1. Check if the DeltaMass is close to $M_{mi}(HPO_3)$.
2. Check if the library peptide contains amino acids S, T, Y.
3. Mark peaks of the query spectrum matching some y-ion of the library spectrum after the shift by $M_{mi}(H_2O)$.
4. Shift the marked peaks (only corresponded to fragment ions containing S, T, Y) by $M_{mi}(H_2O)$ and perform the Hybrid search once again to calculate the match score.
5. Adjust the score of each hit according to its rank in the hit list.

Phospho library

Library	Peptide ions
Non-phospho library	489,921
Phospho library	66,922
Intersection	23,308

- Only about 35% of peptide ions (without counting phosphorylations) are common for the libraries.
- The best result is achieved if both libraries are used together (see Figure 8).

Generating libraries by adding phosphorylation in-silico

- We generated an in-silico phospho library by introducing phosphorylations into the non-phospho library.
- The library works with the Direct search which is much faster than the Hybrid search.
- According to Figure 7, this approach allows finding phosphopeptides with efficiency comparable with other methods.

Comparison of methods

- To compare methods, the MS-GF+ [5] and phospho library results are filtered in the way that they
  - A. include only peptide ions from the non-phospho library
  - B. have one phosphorylation per a peptide.
- According to Figure 7, the optimized Hybrid search is superior to other approaches and returns extra IDs.

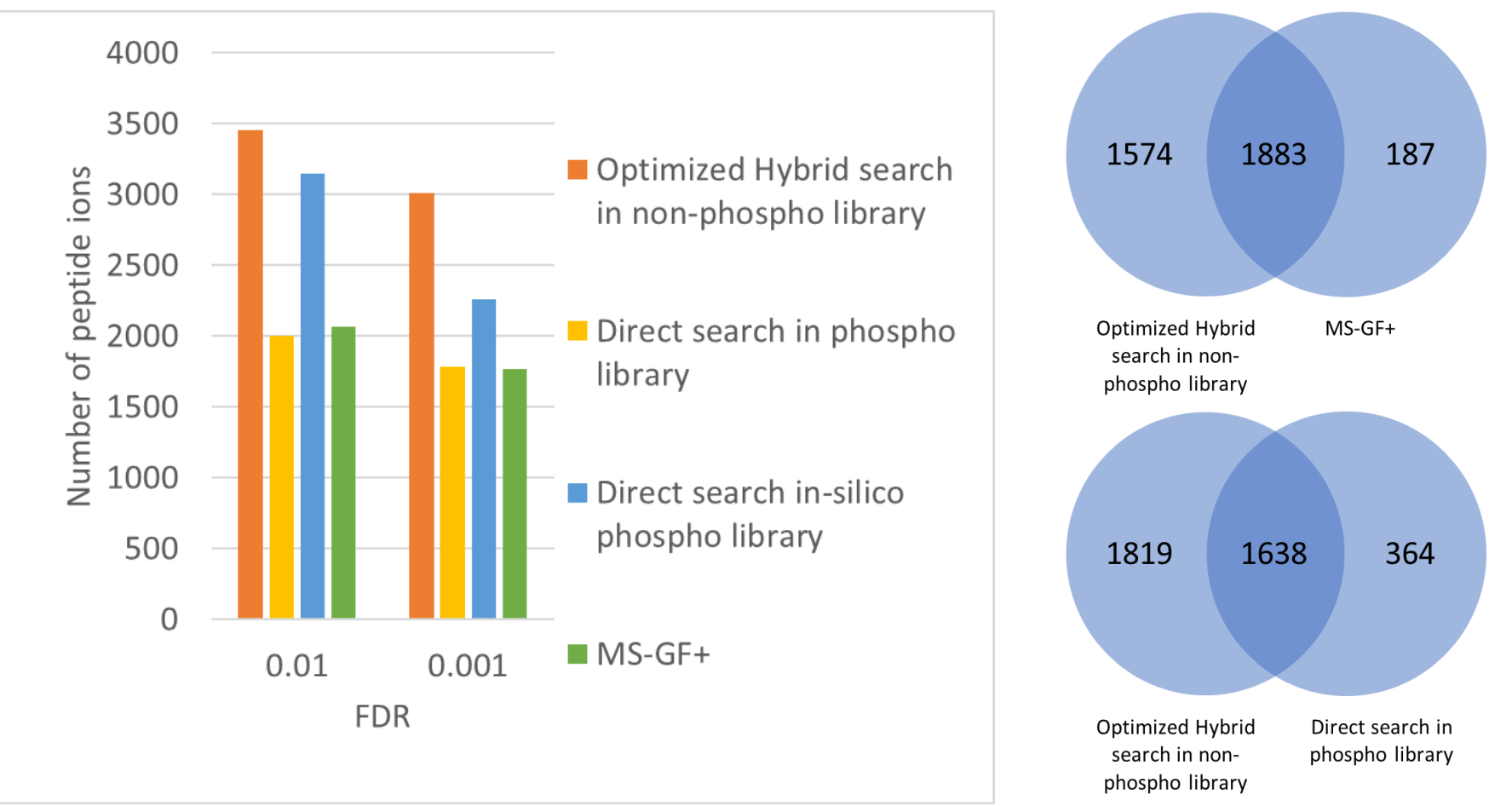

Figure 7. Comparison of different methods for finding phosphopeptides. The right part of the figure shows overlapping of the results for FDR=0.01.

Unfiltered results

- According to Figure 8, performance of the combined search in the phospho and non-phospho libraries is comparable with MS-GF+.
- The largest number of identifications is achieved by combination of the MS-GF+ and MSPepSearch results.

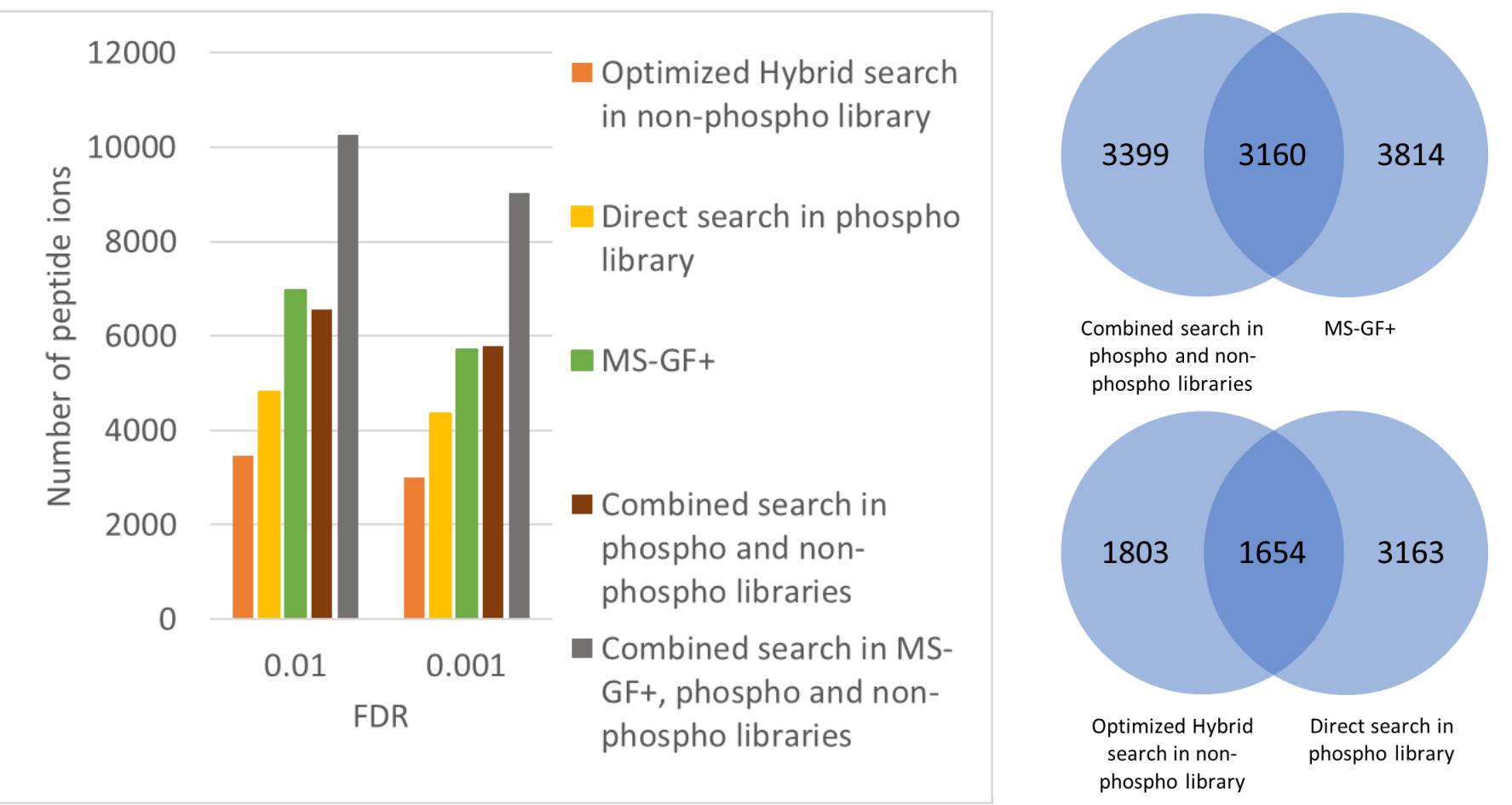

Figure 8. All phosphopeptides are counted regardless of the number of phosphorylations or presence in the non-phospho library.

Conclusions

- MSPepSearch Hybrid search is capable of finding phosphopeptides with any non-phospho library.
- Improvements of the Hybrid search can be achieved by
  - ❖ using a priory information about distribution of DeltaMasses
  - ❖ penalizing matches according to their rank in the hit list
  - ❖ utilizing unique fragmentation patterns
- The developed algorithm can identify phosphopeptides not found either by MS-GF+ or by the Direct search in the available phospho-library.

References.

1. Meghan C. Burke et al. “The Hybrid Search: A Mass Spectral Library Search Method for Discovery of Modifications in Proteomics.” J Proteome Res. 2017 May 5;16(5):1924-1935.
2. Zheng Zhang et al. “Reverse and Random Decoy Methods for False Discovery Rate Estimation in High Mass Accuracy Peptide Spectral Library Searches.” Journal of Proteome Research 2018 17 (2): 846-857.
3. ftp://ftp.pride.ebi.ac.uk/pride/data/archive/2017/05/PXD004415/
4. NIST MS Search libraries of peptide spectra are available at chemdata.nist.gov
5. Sangtae Kim, Pavel A. Pevzner. “MS-GF+: Universal Database Search Tool for Mass Spectrometry.” Nat Commun. 2014 Oct 31;5:5277. doi: 10.1038/ncomms6277.