Augmenting Protein Database Searching with Peptide Library Searching

Robert J. Chalkley and Peter R. Baker
Mass Spectrometry and Proteomics Resource, University of California San Francisco, USA

Peptide Identification Strategies

Protein Database Searching: Most common strategy for MS/MS peptide identification.

Can consider all potential peptides from all proteins in database, including with post-translational modifications. Large search space, but not informed by knowledge of which peptides are likely to be observed based on previous detection.

Spectral Library Searching: Increasingly popular strategy due to development of DIA approaches for comprehensive peptide detection in complex mixtures. Spectra of peptides identified from protein database searching are used to create a spectral library. New spectra are identified by comparing to spectra in the library.

Peptide Library Searching?: Could use a list of peptides previously identified to restrict those that are considered by conventional protein database search engine.

Library Searching compared to Protein Database Searching

Advantages:

•Speed: Peptide and Spectral Library Searching >> faster than Protein Database Searching

•Sensitivity: Smaller search space gives increased confidence identifications.

Disadvantage:

•Can only identify those peptides in the library: likely to miss some identifications

Peptide Library compared to Spectral Library

Advantages:

•Can combine data acquired on different instruments; different fragmentation types; presence of isotope labels do not matter. Hence, should be able to produce more comprehensive library than a spectral library.

•Can use Protein Database search engine – software familiarity.

•Can add peptides (or modified peptides) of interest: not required that the peptide has been observed before.

Disadvantages:

•Does not use information about which fragments were observed before (or relative intensities).

•Not of use for DIA peptide identification.

Library Design

Peptides are sorted by precursor M+H

DB Peptide	Mods	M+H Calc	-10logP	SwissProt.2016.5.9
LYIIPGIPK		1013.639	78	Q8IU60
TPEHSRK	Phospho@1;Phospho@5	1014.381	71	O75182
DYDDMSPR	Oxidation@5	1014.383	111	P61978
DHSVDSFK	Phospho@3=56	1014.393	138	Q641Q2
DHSVDSFK	Phospho@6=40	1014.393	134	Q641Q2
AASWESQR	Phospho@3=45	1014.404	74	Q15361
GPSFNQER	Phospho@3	1014.404	96	Q9Y520
EMEEGEFK	Oxidation@2	1014.409	66	Q9Y2I7
DVSEELSR	Phospho@3=30	1014.414	101	P40222

Modification site and localization score
Allows filtering by site localization confidence

Pvalue for identification confidence
Allows filtering for reliability

Modifications to Protein Prospector to Allow Peptide Database Searching

Batch-Tag was modified to add an option to allow a library of peptides to be searched. Three options:

•Search Protein Database only
•Search Peptide Library only
•Search both Protein Database and Peptide Library at same time
•Can search multiple libraries, if desired

Peptide Library Construction

•Search Compare (Protein Prospector) was modified to enable the construction of libraries by retaining the best hit for any particular peptide.
•Peptides with ambiguous site localizations can either be permuted or eliminated.
•MS-Viewer program allows combining the libraries from multiple data sets.
•MS-Viewer can filter out redundant identifications across datasets

Data Used For Library Generation and Evaluation of Peptide Library Searching

Peptide library was assembled from Sharma et al.
(PRIDE submission PXD000612)

•Both proteome and phosphoproteome surveyed in HeLa cells.
•50000 phosphopeptides identified
•Data (15 million spectra) was re-searched using Protein Prospector, considering standard mods + Phospho STY
•retained only modifications with unambiguous site localizations.

Resulting library contained 170,000 unique peptides

Dataset used to evaluate library: Scheltema et al 2014
PRIDE submission PXD001203

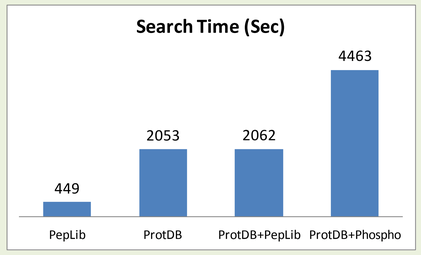
•HeLa cells, not phosphoenriched.
•750000 spectra; i.e. 30x smaller dataset.

•Both datasets from Hela Cells
•Both datasets produced in same lab.
•Both acquired on QExactive
•Overlap of datasets should be higher than most

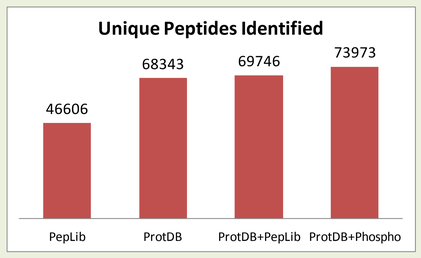
Results

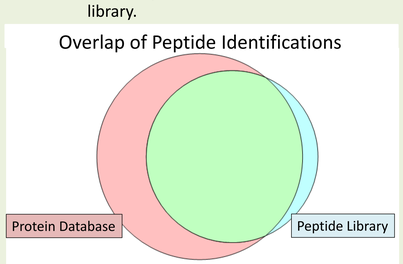
•Protein database search was against target:decoy database of all human entries in SwissProt (≈84K protein entries)
•Peptide library search was against 170K peptide entries.

•Peptide Library search alone is much faster than protein database search.
•Performing peptide library search at same time as protein database search adds essentially no time (about ten seconds to a 34 minute search).
•Considering phosphorylation more than doubles search time.

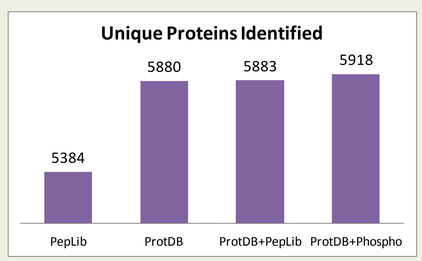


•Searching the protein database led to 40% more IDs than just searching the peptide library
•Library search alone identified 46606 unique peptides
•27% of all entries in the library were identified
•Searching the library in addition to database searching led to 5% more peptide IDs than searching the protein database alone
•Searching considering phosphorylation identified more phosphopeptides than the library search, despite the library containing a lot of phosphorylation sites
•This is largely caused because only peptides where the phospho site could be localized were added to the library.





•9% fewer protein IDs from peptide library search alone.
•Peptide library did not add to number of proteins identified compared to protein database search.
•Considering phosphopeptides did not change the number of proteins identified.



Improving on Peptide Library Performance

If it is to be used on its own:

• the Peptide Library clearly needs to be much bigger

•Many of the missed peptides were due to missed cleavages, or common modifications (e.g. oxidized methionines)
•Should these automatically be permuted and added?

If used in concert with Protein Database searching

•Limited benefit to containing peptides already considered in Protein Database search; i.e. no need for unmodified tryptic peptides.
•Should focus on non-tryptic or modified peptides

•Could populate Peptide Library based on predicted peptides based on published modification sites.

Summary

•Protein Prospector was adapted to allow searching of a Peptide Library as an alternative, or in addition to a Protein Database.
• Results suggest peptide library searching is useful, but more as an addition, rather than alternative to Protein Database searching
•Peptide Library allowed identification of ≈5 % more unique peptide IDs with no appreciable effect on analysis time.
•Fast enough for real-time data analysis
•Many of the peptides missed by Peptide Library searching contained missed cleavages and/or common modifications
•Despite Peptide Library being compiled from data acquired from the same cell type and instrumentation, differences in sample handling were a significant variable
•Optimum benefit of approach may be if the Peptide Library only contains modified and/or non-tryptic peptides that would not routinely be search for in a Protein Database search

Acknowledgements

This work was supported by the Dr Miriam and Sheldon G. Adelson Medical Research Foundation.