ProteoM enables deep and precise identification of peptides in data-dependent acquisition proteomics

Qiang Zhang[1]

Abstract

Sensitive and reliable identification of proteins and peptides pertains the basis of proteomics. We present a database search tool, proteoM, for the analysis of data-dependent acquisition (DDA) proteomics. Compared to several other search engines, proteoM uses a strategy of intensity tally and attains generally a higher performance by depth and precision in twenty datasets varying from large-scale to single-cell proteomics. It on average identifies globally 20% more tryptic and 82% more semi/non-tryptic peptides, and 19% more phosphopeptides, at higher or comparable precision in twelve large-scale datasets.

Introduction

Confident characterization of peptides and proteins from mass spectrometric (MS) data is essential to biological researches that employ MS-based proteomics techniques. Among a multitude of data interpretation facilities, closed searches of tandem mass (MS/MS) spectra from data-dependent acquisition (DDA)[1] remains the baseline module in proteomics pipelines.[2–10] Here closed searches refer to experimental-to-theoretical matches with a predefined space in the variable modifications of theoretical peptide.

Closed searches of DDA data compare MS/MS spectra to peptide sequences in protein databases.[11] While search algorithms against the DDA data in general work well, qualitative and quantitative findings from software tools that match experimental to theoretical spectra do not always reconcile.[12] Some of the discrepancy can be attributed to the different handling of users' specified fixed and variable modifications by search engines. Others may be due to the scoring algorithms[13–16] that arbitrate peptide spectrum matches (PSM) into the two domains of positive or false identifications, the boosting of machine data with complementary ions,[17] the size of search space, the unexpected side effects of software, etc.

In this work, we develop a tool, proteoM, that supports DDA searches of MS/MS spectra. One key advantage of proteoM lies in a strategy of two-stage searches (Fig. 1). The tool matches separately the main sequences of MS/MS ions (e.g., *b*- and *y*-ions) and the the satellite series (b^0, y^0, b^*, y^* and doubly charged b^2, y^2, b^{20}, y^{20}, b^{2*}, y^{2*}). It then refrains the

[1] Department of Medicine, Washington University in St. Louis, St. Louis, Missouri, USA., zhang.qiang@wustl.edu

information learned from the satellite matches by disallowing them from being counted as independent evidences in peptide scoring. Instead, the satellite intensities are tallied onto the corresponding main fragment-ion intensities, followed by intensity-based enrichment analyses (Fig. 2a). The strategy enables generally deeper proteome discovery at higher or compatible precision, demonstrated by twelve large datasets. ProteoM also generalizes the handling of fixed and variable modifications that are sometime incompatible, with an additional benefit in search space minimization.

Results

Fixed and variable modifications

Interactions of fixed and variable modification terms specified in a database search may lead to additive effects that violate the fundamentals in chemistry. For instance, the accumulated modification from the interaction of fixed peptide N-terminal TMT10plex labeling and variable protein N-terminal acetylation may be undesirable. To address this, proteoM begins by deploying a minimal set of compatible fixed and variable modifications from a users' specification (Supplementary Note 1). In the example of a users' specification of fixed Carbamidomethyl (C), TMT10plex (N-term) and TMT10plex (K) and variable Acetyl (Protein N-term), Gln->pyro-Glu (N-term = Q), Oxidation (M) and Deamidated (N) (Supplementary Table 1), the compilation and minimization by proteoM leads to a total of twelve combinations of fixed and variable modifications (Supplementary Table 2).

It has been commonly observed that an enlarged search space can reduce the numbers of identified PSMs and peptides when controlling false discovery rates (FDR).[18] The combinatorial minimization by proteoM guards against undue expansion of a search space with regard to users' intention in finding peptides that are more likely to present in samples (Supplementary Table 2). In addition, it reduces the users' responsibility in constructing compatible fixed and variable modifications at a given search.

Intensity tally from two-stage searches

A strategy of two-stage search is employed by proteoM. After matching the main sequences of MS/MS ions (e.g., *b*- and *y*-ions), the search engine then matches the satellite series of fragment ions (b^0, y^0, b^*, y^* and doubly charged b^2, y^2, b^{20}, y^{20}, b^{2*}, y^{2*}), tallies their intensities to the corresponding main fragment ions, followed by intensity-dependent enrichment analyses of probability scores against the main fragment ions (Fig. 2a and **Methods**). The findings of PSMs are stored in files `psmC.txt` and `psmQ.txt` where the former contains the complete list of PSMs in the search space at a given set of search arguments and the latter gives the quality subset that passes a user-specified FDR threshold.

High performance in global peptide identification with proteoM

We first compared the performance of proteoM to several other search engines using six datasets of 10-plex global TMT (Supplementary Table 3: WHIM_G). Under the assumption of tryptic enzymatic specificity, all engines perform comparably with similar numbers of

significant PSMs, peptides (without modification), sequences (with applicable modification) and proteins (Fig. 2b). Yet notable difference remains, exemplified by a relatively low percentage of common PSMs across search engines (Fig. 2c, Supplementary Notes 2 and 3). Within the six datasets, proteoM is in general the top performer in term of PSM, peptide and sequence identifications (Supplementary Fig. 1a). It also yielded regularly the fewest numbers of proteins with single-peptide identifications (Supplementary Fig. 1b). Analogous traits can be established with a pancreatic cancer dataset Panc_G1 (Supplementary Fig. 2).

We next tested Pearson correlations between the first two WHIM2 replicates in each of the six global datasets. Among the four search engines, proteoM tends to give high correlations with fewer extreme values at low log2FC (Fig. 2d, Supplementary Fig. 3). The low extreme log2FC correspond to the measures with much lower reporter-ion signals when compared to the common reference (the averaged of signals from five WHIM2 and five WHIM16). Note that data stretching by extreme values can amplify correlations (Supplementary Note 4). We quantified additionally the data similarity by measuring the Manhattan distance of log2FC (less affected by reference choices) between replicated samples and found that proteoM is one of the two search engines that give regularly the shortest pairwise distances (Supplementary Fig. 4).

We further assessed the precision of peptide and protein identifications against the global datasets using an entrapment strategy.[19] Different to the estimates of FDR using a reversed database,[20] the entrapment approach introduced an intermediate step by appending a low homology database to a target database. The combined database was then reversed and served as a decoy for FDR controls. Under 1% FDR of proteins, we observed that the entrapment rates progressed comparably across the four search engines (Supplementary Note 5). We also applied an orthogonal entrapment strategy by specifying the variable modification of TMTzero™ (TMT0) to peptide N-terminal and site K in the search. The molecular mass of TMT0 differs to that of TMT10plex by 5.01045 Da and are expected to be absent in general from the WHIM_G dataset. Among the four search engines, proteoM yielded low entrapment rates that progress consistently along the decreasing peptide scores (Supplementary Fig. 5)

Identities of semi/non-tryptic peptides are among the so-termed dark matters in MS data.[21] We further searched the WHIM_G dataset at no enzymatic specificity (NES). We found that, in general, proteoM outperforms other search engines by numbers of PSMs and peptides (Supplementary Fig. 6, Supplementary Table. 4) and precision (Supplementary Fig. 7). For instance with the JHU_G2 subset, it identified twice as many PSMs, peptides and sequences than the search engine that reported the second highest correlations (Fig. 2b). Presumably the production of semi/non-tryptic peptides are often less defined than tryptic ones. In line with the notion, decreased correlations and increased distances between replicated samples were observed for all search engines when comparing the semi/non-tryptic to the tryptic findings (Fig. 2e).

Deep proteome coverage in phosphopeptides

Protein phosphorylation is one of most widely explored post-translational modifications in proteomics. We demonstrate the performance of proteoM in the characterization and quantitation of immobilized affinity chromatography (IMAC) enriched phosphopeptides , using the six datasets of 10-plex TMT under WHIM_P (Supplementary Table 3). We noted that proteoM in most cases is the leading search engine in PSM, peptide and sequence identifications against the six IMAC datasets (Supplementary Fig. 8). We further observed that, in the majority cases, proteoM yielded the shortest similarity distance between replicated samples (Supplementary Fig. 9). In the example of the JHU_P2 dataset, proteoM characterized 19%, 22% and 17% more PSMs, peptides and sequences, respectively, than the search engine that reported the second highest numbers in the three categories, and yet at higher precision (Fig. 3a-b). When compared to the search engine that gave the highest correlation and near equal distance, proteoM gains 20%, 22% and 33% in the corresponding identifications.

It is worthwhile to note the greater discrepancy in the identities of sequences than peptides. For instance, 22.2% of sequences versus 3.2% of peptides were missed from proteoM when compared to the findings from Mascot. The observation is consistent with the documentation of low correlations in the phosphorylation site specificity between search engines.[22] The localization probability by proteoM is modified from the AScore algorithm.[23] Instead of assuming binomial probabilities, proteoM applied alternatively a counting statistics to weigh the localization probability.[24] For instance, at a 2:1 counts of localization-specific observations, the former will be given a 0.67% probability in localization. Analogous at a 1:0 ratio, the former will be 100% responsible for the site specificity. Different to the binomial model in AScore, the localization score with proteoM will be the same at either ratio 1:0 or 2:0.

We further tested the accuracy of phosphopeptide identification with proteoM using dataset Ferries_spikes (Supplementary Table 3). The dataset comprises of 175 known phosphopeptides that are associated to 191 sequences by including the variation in phosphosites. We found that all search engines performed comparably in the numbers of identified phosphopeptides and sequences (Supplementary Table 5). For instance, proteoM and Mascot identified 220 and 212 phosphosequences, respectively, at good proximity to the expected value of 191 or more (from impurities).

General purpose searches with proteoM

ProteoM is a general purpose search engine currently for closed searches (with pre-specified variable modifications) of tandem mass spectra. For instance, it is readily applicable to feature matching of acetylome and ubiquitylome (Supplementary Fig. 10), as well as SILAC experiments (Supplementary Table. 6). In the examples of small datasets from single-cell proteomics, we observed that all search engine performed comparably with proteoM reported the fewest proteins at single-peptide identifications (Supplementary Fig. 11).

Discussion

We have shown that proteoM is capable of in-depth characterization and quantitation of peptides and proteins from DDA data. We further showed that proteoM regularly captures DDA peptides and proteins at high precision. The depth and precision may facilitate biological research by generating novel insights with reliable proteomic discovery. Part of the reason for the popularity of DDA may be ascribed to its compatibility to high-throughput, multiplex TMT workflows. Data independent acquisition (DIA) methods, on the other hand, allow all precursors in a wide isolation window of m/z being fragmented simultaneously for MS2 interrogation.[25] Future works with proteoM may include DIA workflows where intensity tally may aid the separation of signals from background noise and the simplification of more complex DIA spectra.[17,26–29]

ProteoM searches of a set of global data at twenty-five RAW files can be completed typically within 2 hrs (databases: Refseq human mouse and cRAP; number of missed cleavages: 4; precursor mass error tolerance: 20 ppm; product ion mass error tolerance: 20 ppm, fixed and variable modifications in Supplementary Table 2), when using an inexpensive 8-core (Intel i7-11700 CPU, 3.6 GHz) PC at 32 GB RAM. In the example of dataset BI_G1, the matches of *b*- and *y*-ions are typically done in less than 48 mins and the complete process in less than 84 mins when using cached precursor masses. Currently the search rate is limited by the generation of theoretical MS/MS ions that subject to heavy permutations of variable modifications and sites (e.g. searches of phosphopeptides). We conceive further performance improvement upon rewriting the permutation in C++. Another time-consuming step is on the maintenance of the one-to-one correspondence between experimental and theoretical MS2 m/z values, which is used by the intensity tally of product ions for peptide scoring.

Methods

Specifications of fixed and variable modifications

The Unimod[30] definition of positions and sites were adopted by proteoM for specifying fixed and variable modifications. The value of a position is in one of "Anywhere", "Protein N-term", "Protein C-term", "Any N-term" or "Any C-term". The last two position labels can be shorthanded as "N-term" and "C-term". A site is a one-letter representation of the twenty amino-acid residues, as well as the terminal sites of "N-term" and "C-term". The general format in specifying a fixed or variable modification is `title (position = site)` where title is a unique character string without space. At a position of "Anywhere", the modification can be shorthanded as `title (site)`, for example, `TMT10plex (K)`. For a terminal modification at any site, it can be abbreviated as `title (position)`, for example, `Acetyl (Protein N-term)` and `TMT10plex (N-term)`. There are circumstances that both position and site are needed for specifying a modification, for instance, `Gln->pyro-Glu (N-term = Q)`. More examples are available in the help document of proteoM utility of `parse_unimod`.

Workflow of proteoM

Software environment and input data. ProteoM was developed under the free software environment R[31] and Posit/RStudio[32]. It processes peak lists from MSConvert[33] for searches of experimental mass spectra. When converting RAW MS files, the peak lists can be at either Mascot generic format (MGF) or mzML format. The option of `TPP compatibility` needs to be checked (as defaulted) at either format for downstream parsing of the title lines in the peak lists with proteoQ (http://github.com/qzhang503/proteoQ). With mzML outputs, the default of `Use zlib compression` needs to be *unchecked* (opposite to the default) for proper decoding of binary data to MS2 m/z and intensity values.

Compilation of fixed and variable modifications. The same site, including "N-term" or "C-term" of a peptide, at multiple variable modifications are allowed (for example variable TMT10plex modification and acetylation both at site K). However, the same site being specified under both fixed and variable modifications are considered initially incompatible in that the addition of the variable mass to the fixed mass may be chemically unsound (for example fixed TMT10plex modification and variable acetylation of K) . To handle this, proteoM first coerces the incompatible fixed modifications, F_x, to the category of variable modifications. Sets of combinatorial variable modifications are then computed. Among them, combinations containing multiple "N-term" or "C-term" modifications are removed (for example TMT10plex modification at "N-term" and acetylation at "Protein N-term"). The program next looks for a coerced site at position "Anywhere" and removes the combinations without the site (for example removals of combinations lacking K with the fixed-to-variable coercion of TMT10plex modification of K). The same rule applies to coerced terminal modifications. In case of multiple coercions of "Anywhere" sites, combinations that do not contain *all* of the coerced sites are removed. Finally for each combination, the coerced F_x will be reverted back to fixed modifications if there are no site violations (for example F_x is TMT10plex modification at "N-term" and there is no other variable "N-term" modifications). Note that to obtain an additive effect, a new Unimod entry containing multiple modifications to the same site can be constructed (for example TMT10plex modification at "N-term" and the formation of pyro-glutamic acid from N-terminal glutamine). The proteoM utility `add_unimod` can be used for the purpose.

Amino-acid look-ups. Look-ups of amino-acid residues are stored as named vectors for each combination of fixed and variable modifications (for example, twelve modules of look-ups when instantiating the modifications specified in Supplementary Table 2). Sites and variable modifications are in names and the corresponding mono-isotopic masses are in values. For example with the specification of variable "Oxidation (M)", both unmodified "M" and "Oxidation (M)" are available in the look-up at masses of 131.0405 and 147.0354, respectively. Masses of fixed modifications are applied directly to the corresponding sites. Titles, positions and sites of modifications are linked to the look-ups as R-language attributes. Additional attributes includes sites of coercion, sites with neutral losses, etc. The attributes were calculated once and can be re-used when needed. For instance when calculating the theoretical MS2 ions, an attribute of "nl+" prompts the calling of utilities that consider the permutation of neutral losses".

Precursor masses. Protein entries in FASTA databases were digested *in silico* into peptide sequences according to a full enzymatic specificity, for example tryptic cleavages at C-terminal K or R, at no missed cleavage. Masses of the bare sequences are first obtained by summing over the masses of amino-acid residues therein (temporarily ignore the masses at sites "N-term", "C-term" and the mass deltas by "Anywhere" variable modifications). Rolling sums are then applied to obtain the masses of longer sequences by concatenating adjacent pieces of peptide sequences according to the number of allowed missed cleavages, followed by the capping of "N-term" and "C-term" masses. The full sequences are then filtered by the range of lengths specified in the search. Analogous rolling sums are employed to obtain the masses of peptide sequences over a range of length for searches at NES. For searches with semi-enzyme specificity, the semi-enzyme sequences and their masses are generated by sequential trimmings of residues from the N- and/or the C-terminals of full sequences. Next the sequences are dispatched according to the attributes of amino-acid look-ups (Supplementary Fig. 12). For instance, when against a look-up with an attribute of "Oxidation (M)", only sequences with residue M are retained.

The base module of the look-ups corresponds to the combination of fixed and variable modifications that were specified originally in a search. Subjected to fixed-to-variable coercions at sites "N-term", "C-term" and "Anywhere", the masses in the base module are compensated accordingly and concluded. The masses of various variable terminals are then capped onto the sequences for the remaining modules (there is only one N- and C-terminal in a peptide sequence and simple summation makes no difference whether the terminal sites are coerced or not). Note that coerced "Anywhere" modifications are reverted back to fixed modifications when possible (*Compilation of fixed and variable modifications*). The precursor masses are thus compensated by the back-coereced "Anywhere" masses accordingly. Finally, combinatorial variable masses of sequences, including neutral losses, are computed to form the space of precursor masses by modules.

The modualized precursor masses are further binned according to the tolerance in mass error. The bin width is half of the size of the tolerance. For instance, at a default precursor mass tolerance of 20 ppm, the theoretical masses are binned at every 10 ppm. Later when matching tandem spectra, +/-20 ppm is the maximum mass error (the maximum distance between two adjacent bins) that can be found between experimentals and theoreticals. Experimental spectra are binned accordingly. Namely, the minimum precursor mass that have been specified in a search is used as the common starting point in the binning. This ensures a one-to-one correspondence of bin indexes between experimentals and theoreticals. Finally, the binned theoretical and experimental precursor masses in the form of neutral masses are stored on disk (and can be reused in a new search at matched search parameters). The decoy sequences are constructed by reversing target peptide sequences with the fixation of both the N- and the C-terminal residues (e.g. PEPTIDE to PDITPEE).

Tandem spectra matches. Experimental tandem spectra are matched to theoreticals by modules of fixed and variable modifications. Under each module, the experimental and theoretical spaces are first subset mutually by bins of overlapping precursor masses. Tandem spectra are then generated for the theoretical precursors and the searches traverse through the bins. To avoid combinatorial explosion, the permutation of variable

modifications and their sites are modified from the recursion version in R package gtools with "for" loops.

When matching the primary MS2 species (e.g., b and y ions), both experimental and theoretical MS2 values, x, are converted to indexes $\lceil \ln(x/x_0)/\ln(1+d) \rceil$ where x_0 is the minimum MS2 m/z supplied in a search and d is a bin width, for example 10 ppm. The experimental and theoretical values with mass difference smaller than d can be off by +/-1 in bin indexes during the numeric-to-integer conversions. Additional matches with 1-increment or decrease of theoretical indexes are carried out. The one-to-one correspondence between experimental and theoretical m/z values, as well as experimental MS2 intensities, are necessarily maintained. These allow a later enrichment analyses of peptide scores based on the number of matches at descending MS2 ion intensity. Disparity between experimental and theoretical values can occasionally occur due to the numeric-to-integer conversions (for example, multiple experimental indexes match to the same theoretical index). In rare events of this kind, computationally more expensive outer products are performed for matches between the original experimental and theoretical m/z values.

Peptide scores. Additional searches are performed against tandem spectra with a minimum of n matched primary b and/or y ions ($n = 6$ by default). Outer products are used when matching secondary ions b^*, b^0, y^*, y^0, b^2, b^{*2}, b^{02}, y^2, y^{*2} and y^{02}. Experimental intensities of the matched secondary ions are ascribed to those of the corresponding primary ions, for example,

$$I'_{b_i} = I_{b_i} + I_{b_i^*} + I_{b_i^0} + I_{b_i^2} + I_{b_i^{*2}} + I_{b_i^{02}}, \quad i = 1,2,3, \ldots$$

where I stands for experimental intensity and I'_{b_i} is the boosted experimental intensity upon summing I_{b_i}, $I_{b_i^*}$ etc. Theoretical m/z are next ordered decreasingly according to the boosted intensity. Hypergeometric probability scores are assessed progressively against the numbers of matched primary ions versus the numbers of MS2 features. The best (minimum) probability, P, is applied and the corresponding entrenchment scores is $-10 \cdot \log_{10}(P)$. The FDR is estimated by a target-decoy approach against the subclass of search results that yields the maximum number of matches. The estimated FDR are then transferred to all results.[34]

Protein groups. Identities of proteins and peptides are arranged in a sparse logical matrix (R package Matrix) with column and row associations to proteins and peptides, respectively. A logical value of 1 in the matrix stands for the presence of a peptide under a protein. The matrix is next divided rowly by unique (r1) and shared (r2) peptides (Figure 4). The submatrix r2 is further separated columnly by proteins with (c1) or without (c2) peptide counts. The same column separation applies to the submatrix of unique peptides (r1). Note that values in the submatrix intersecting r2 and c2 are all logical zeroes and each of the c2 proteins forms their own group. Further note that unique peptides have no effect on the grouping of proteins by distance.

It is sufficient to perform protein-peptide grouping against the submatrix intersecting c1 and r2, followed by additional increments in cardinal numbers for proteins with single

peptide identifications. Pairwise distance of proteins are first assessed (R package proxyC) by the logical condition `!any(A & B)` (not any "both A and B are 1"). It returns 0 if protein A and B share peptides and 1 if not. The dichotomy distance matrix is next clustered hierarchically with the agglomeration method of single link. The cluster is then cut at an arbitrary height greater than 0 (heights are 0 between proteins that share peptides). Finally, essential proteins are determined using a greedy set cover algorithm.[35]

Search speed. The searches of a set of global data (at twenty-five RAW files) can be completed typically within 2 hrs (databases: Refseq human mouse and cRAP; number of missed cleavages: 4; precursor mass error tolerance: 20 ppm; product ion mass error tolerance: 20 ppm, fixed and variable modifications: Supplementary Table 2), when using an 8-core (Intel i7-7820 CPU, 3.6 GHz) PC at 32 GB RAM. In the example of dataset BI_G1, the matches of *b*- and *y*-ions are typically done in about 50 mins and the complete process in about 105 mins or within 100 mins when using cached precursor masses. Currently the search rate is limited by the generation of theoretical MS/MS ions that subject to heavy permutations of variable modifications and sites (e.g. searches of phosphopeptides). We conceive further performance improvement upon rewriting the permutation in C++. Another time-consuming step is on the maintenance of the one-to-one correspondence between experimental and theoretical MS2 m/z values, which is used by the intensity tally of product ions for peptide scoring.

Database searches

MS data were converted to peak lists using MSConvert (v3.0.22279). The MS/MS spectra were analyzed using Mascot (v2.8.0.1), MaxQuant (v2.0.2.0), MSFragger (v18.0) and proteoM (v1.2.0.0). The top-100 most abundance product ions were used with MSFragger and proteoM. The allowance in mass error was set to +/-20 ppm for both precursor and product ions. A maximum of 4 missed cleavages was allowed for searches assuming the digestion enzyme of trypsin/P. The range of peptide lengths was between 7 and 40 residues and the range of precursor masses between 700 and 4500 Da (Supplementary Fig. 13). In multiplex TMT studies, the tolerance in reporter-ion mass error was set to +/-10 ppm with Mascot and proteoM, 0.003 Da with MaxQuant and +/-20 ppm with MSFragger. A maximum of five variably modified sites per peptide sequence was allowed with a maximum of three positions at the same modification and 64 permutations in positions.

Datasets WHIM_G and WHIM_P. The global WHIM_G dataset was set up to search against a Refseq database of human mouse proteins (ver Jul. 2018; 56,789 entries) and common contaminant proteins (cRAP, v1.0 Jan. 2012; 116 entries). The Refseq database was changed to Uniprot human mouse (ver Oct. 2020; 37,522 entries) when searching against the phosphopeptide WHIM_P dataset. Cysteine carbamidomethylation was specified as a fixed modification. Methionine oxidation, asparagine deamidation, N-terminal glutamine to glutamic acid conversion and protein N-teriminal acetylation were specified as variable modifications. Phosphorylation of S, T and Y were specified additionally as variable modifications when searching against the WHIM_P dataset. TMT10plex modification of peptide N-terminals and K were specified under the quantitation method with Mascot and MaxQuant. The TMT10plex modification of K was set as a fixed modification with both MSFragger and proteoM. The TMT10plex modification of peptide N-terminals was set as a

variable modification with MSFragger and fixed with proteoM with applicable coercion to a variable modification in the latter. The specifications of variable asparagine deamidation and N-terminal glutamine to glutamic acid conversion were removed in the searches of WHIM_G at no enzymatic specificity. Results were filtered at 1% FDR at proteins (also met at the levels of PSM and peptide) with the exception of Mascot results being reported at 1% engine-specified PSM FDR.

Additional datasets. The Uniprot database was subset by species human (20,512 entries) when searching against the datasets of lung (acetylome Lung_A1_A10 and ubiquitylome Lung_U) and pancreatic (Panc_G1 and Panc_P1) carcinomas, as well as datasets Ferris_spike, single-cell Hela at 200pg and 2 ng inputs. Lysine acetylation (K-Ac) and carbamylation were both set to variable modifications in acetylome searches. Lysine ubiquitylation (K-GG) with and without TMT10plex were set to variable modifications in ubiquitylome searches. The search parameters against the WHIM_P were used for the searches of dataset Ferries_spike with the exclusion of TMT10plex modifications. The Dong_SILAC dataset was searched against a Uniprot database of E-coli (Jul. 2022; 4,564 entries) and the cRAP. The same fixed and variable modifications specified in the global WHIM_G searches were used except for TMT10plex.

Entrapment analysis by species was performed against the combined dataset of Refseq human mouse, Uniprot arabidopsis thaliana (Sept. 2022; 16,312 entries) and cRAP with the identical sets of fixed and variable modifications specified in the WHIM_G searches. Entrapment analysis by modifications was performed with the same settings in the WHIM_G searches, with the additional variable TMT0 modification to both K and peptide N-terminal.

Data analysis

Results from Mascot and MaxQuant were filtered by a minimum of 6 matching product ions (the same applied to MSFragger and proteoM via search arguments) . Note that the tolerances in precursor and product-ion mass errors can be interpreted differently by search engines. Results of Mascot, MSFragger and proteoM were further filtered by +/- 10 ppm in precursor mass errors. With quantitative multiplex TMT data, PSMs without peptide N-terminal modifications were also removed from MSFragger (filtered by “N-term” under column “Assigned Modifications” in the psm.tsv output) and proteoM (automated with the latter when compiling fixed and variable modification sets). PSMs with all missing precursor intensity across samples were removed from quantitative analysis. Protein identifications were summarized to gene products.

User-interface utilities

The main utility of proteoM is `matchMS` for database searches. Additional user-interface utilities include `table_unimod` that tabulates titles, sites, positions and monoisotopic masses of all Unimod modifications, `find_unimod` that extracts the information of a specific Unimod entry with additional specifications in neutral losses, `parse_unimod` that checks the grammar of a user-specified modification, `load_fasta2` that loads FATSTA databases, `calc_unimod_compmass` that calculates the monoisotopic mass of a modification by a

chemical composition, `add_unimod`, `remove_unimod` and `remove_unimod_title` that add or remove Unimod entries. Utility `mapMS2ions` maps MS2 ions between theoreticals and experimentals and `make_mztab` prepares mzTab files for publication compliance. Utilities `calc_monopeptide` and `calc_ms2ionseries` calculate the possible precursor masses and MS2 product ion series for a peptide sequence at a given set of fixed and variable modifications.

Acknowledgements

QZ thanks Drs. Hao Chi, Dirk Eddelbuettel, Alexey Nesvizhskii, Clay Semenkovich, Reid Townsend and Kohei Watanabe for discussion.

References

1. Bantscheff, M., Lemeer, S., Savitski, M. M. & Kuster, B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal. Bioanal. Chem.* **404**, 939–965 (2012).
2. 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.* **5**, 976–989 (1994).
3. Perkins, David N, Pappin, Darryl JC, Creasy, David M, & Cottrell, John S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
4. Ma, B. *et al.* PEAKS: powerful software for peptidede novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **17**, 2337–2342 (2003).
5. Cox, J. *et al.* Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
6. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: An open-source MS/MS sequence database search tool. *Proteomics* **13**, 22–24 (2013).
7. Wenger, C. D. & Coon, J. J. A Proteomics Search Algorithm Specifically Designed for High-Resolution Tandem Mass Spectra. *J. Proteome Res.* **12**, 1377–1386 (2013).

8. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5**, 5277 (2014).
9. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat. Methods* **14**, 513–520 (2017).
10. Chi, H. *et al.* Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nat. Biotechnol.* **36**, 1059–1061 (2018).
11. Eng, J. K., Searle, B. C., Clauser, K. R. & Tabb, D. L. A Face in the Crowd: Recognizing Peptides Through Database Search. *Mol. Cell. Proteom.* **10**, R111.009522 (2011).
12. Mertins, P. *et al.* Reproducible workflow for multiplexed deep-scale proteome and phosphoproteome analysis of tumor tissues by liquid chromatography–mass spectrometry. *Nat. Protoc.* **13**, 1632–1661 (2018).
13. Mann, M. & Wilm, M. Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. *Anal. Chem.* **66**, 4390–4399 (1994).
14. Olsen, J. V. & Mann, M. Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 13417–13422 (2004).
15. Reiter, L. *et al.* Protein Identification False Discovery Rates for Very Large Proteomics Data Sets Generated by Tandem Mass Spectrometry. *Mol. Cell. Proteom.* **8**, 2405–2417 (2009).
16. Ma, K., Vitek, O. & Nesvizhskii, A. I. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC Bioinform.* **13**, S1 (2012).
17. Tsou, C.-C. *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **12**, 258–264 (2015).
18. Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteom.* **73**, 2092–2123 (2010).

19. Granholm, V., Navarro, J. F., Noble, W. S. & Käll, L. Determining the calibration of confidence estimation procedures for unique peptides in shotgun proteomics. *J. Proteom.* **80**, 123–131 (2013).
20. Elias, J. E. & Gygi, S. P. Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics. in *Proteome Bioinformatics* (eds. Hubbard, S. J. & Jones, A. R.) vol. 604 55–71 (Humana Press, 2010).
21. Skinner, O. S. & Kelleher, N. L. Illuminating the dark matter of shotgun proteomics. *Nat Biotechnol.* **33**, 717–718 (2015).
22. Savitski, M. M. *et al.* Confident Phosphorylation Site Localization Using the Mascot Delta Score. *Mol. Cell. Proteom.* **10**, S1–S12 (2011).
23. Beausoleil, S. A., Villén, J., Gerber, S. A., Rush, J. & Gygi, S. P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol.* **24**, 1285–1292 (2006).
24. Satpathy, S. *et al.* A proteogenomic portrait of lung squamous cell carcinoma. *Cell* **184**, 4348-4371.e40 (2021).
25. Ludwig, C. *et al.* Data-independent acquisition-based SWATH - MS for quantitative proteomics: a tutorial. *Mol. Syst. Biol.* **14**, (2018).
26. Prakash, A. *et al.* Hybrid Data Acquisition and Processing Strategies with Increased Throughput and Selectivity: pSMART Analysis for Global Qualitative and Quantitative Analysis. *J. Proteome Res.* **13**, 5415–5430 (2014).
27. Guan, S., Taylor, P. P., Han, Z., Moran, M. F. & Ma, B. Data Dependent–Independent Acquisition (DDIA) Proteomics. *J. Proteome Res.* **19**, 3230–3237 (2020).
28. Yang, Y. *et al.* In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nat. Commun.* **11**, 146 (2020).

29. Sinitcyn, P. *et al.* MaxDIA enables library-based and library-free data-independent acquisition proteomics. *Nat. Biotechnol.* **39**, 1563–1573 (2021).
30. Creasy, David M & Cottrell, John S. Unimod: Protein modifications for mass spectrometry. *Proteomics* **4**, 1534–1536 (2004).
31. R Core Team. R: A Language and Environment for Statistical Computing.
32. RStudio Team. RStudio: Integrated Development Environment for R. (2020).
33. Adusumilli, R. & Mallick, P. Data Conversion with ProteoWizard msConvert. in *Proteomics* (eds. Comai, L., Katz, J. E. & Mallick, P.) vol. 1550 339–368 (Springer New York, 2017).
34. Fu, Y. & Qian, X. Transferred Subgroup False Discovery Rate for Rare Post-translational Modifications Detected by Mass Spectrometry. *Mole. Cell. Proteom.* **13**, 1359–1368 (2014).
35. Koskinen, V. R., Emery, P. A., Creasy, D. M. & Cottrell, J. S. Hierarchical Clustering of Shotgun Proteomics Data. *Mol. Cell. Proteom.* **10**, M110.003822 (2011).

Figure 1. The proteoM workflow. User-supplied fixed and variable modifications are compiled into sets of fixed and variable modifications (Supplementary Note 1). The peptide candidates from *in silico* digestion are dispatched according to the modification sets (Supplementary Fig. 12). For instance, only sequences with amino-acid residue M will be retained if methionine oxidation is part of the variable modifications in a given set. The sequences are then binned by their precursor masses and searched against experimental MS data. The MS/MS features are first matched by primary product ions, for example, *b*- and *y*-ions. The secondary ions, *b**, *y** (-NH_3) etc., are then searched for and the signals are ascribed to the corresponding primary ions. The, and only the, primary ions with the boosted intensity are used in peptide scoring. The codes are available under R package https://github.com/qzhang503/proteoM/.

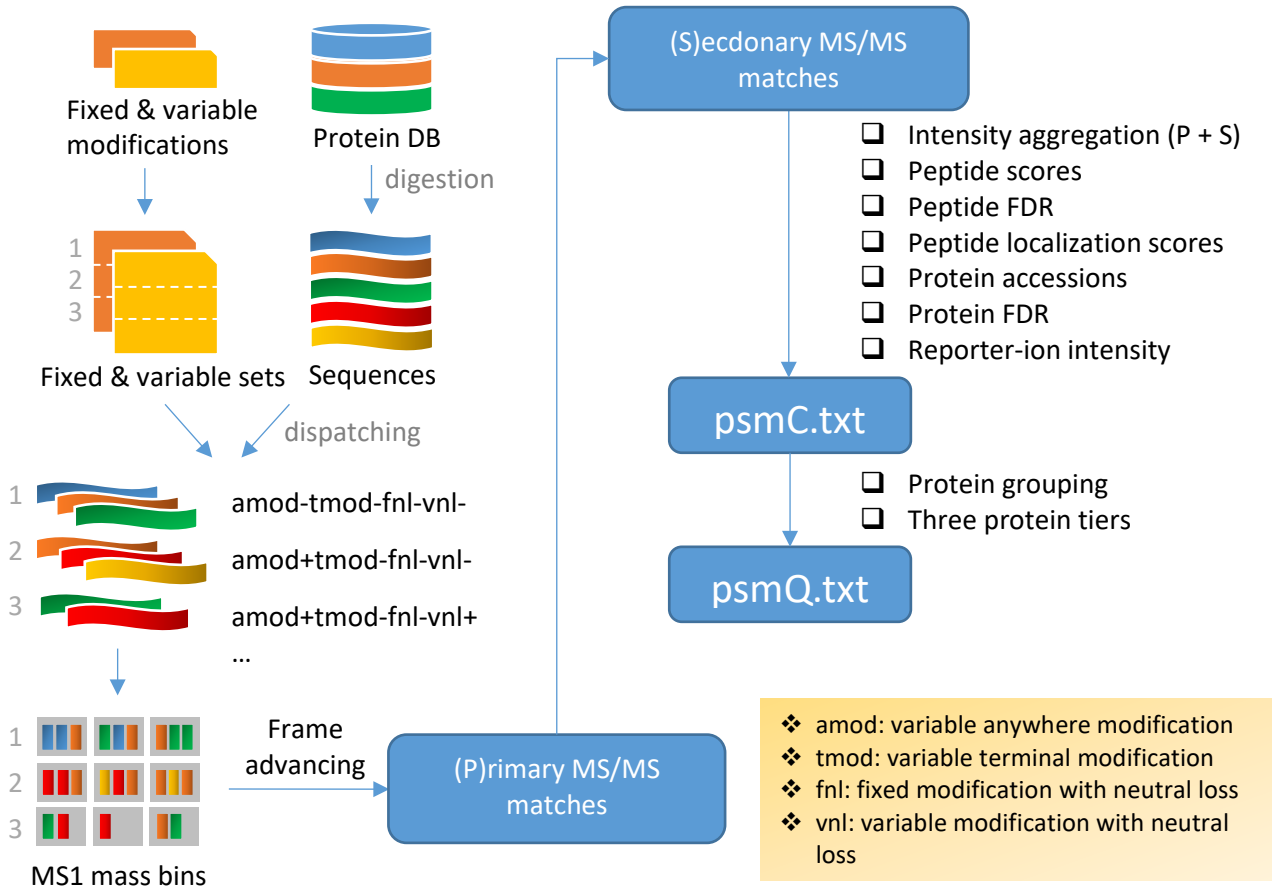
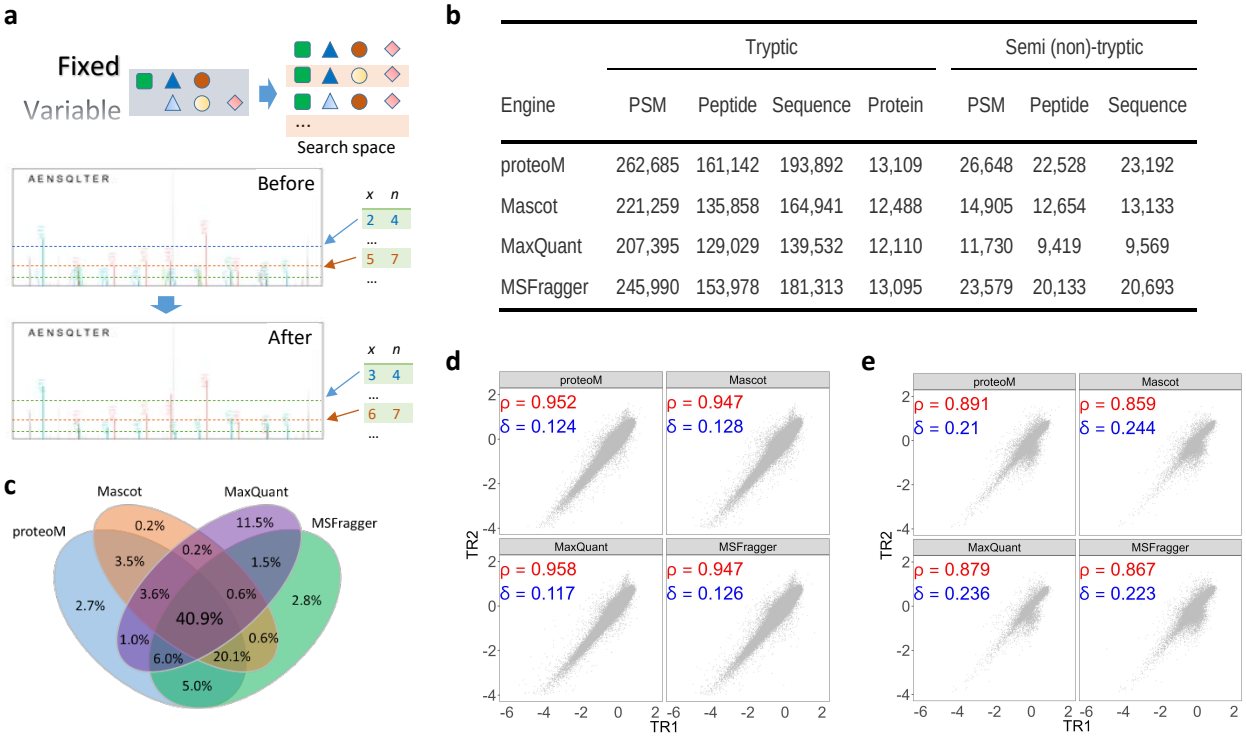


Figure 2. Performance of proteoM on the global dataset JHU_G2.

(**a**) The workflow. (**b**) Summary in peptide and protein identifications. Peptide: without modification information in the literal of a peptide representation; Sequence, with applicable modification information in the literal. (**c**) Significant PSMs. (**d-e**) Pearson correlation (ρ) and Manhattan distance (δ) of log2FC. (d) Tryptic; (e) semi/non-tryptic.

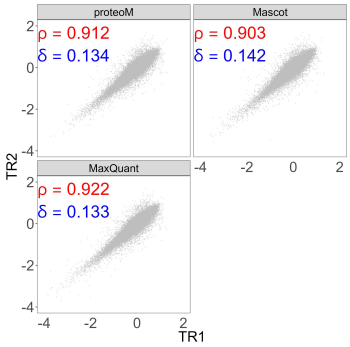


	Tryptic				Semi (non)-tryptic		
Engine	PSM	Peptide	Sequence	Protein	PSM	Peptide	Sequence
proteoM	262,685	161,142	193,892	13,109	26,648	22,528	23,192
Mascot	221,259	135,858	164,941	12,488	14,905	12,654	13,133
MaxQuant	207,395	129,029	139,532	12,110	11,730	9,419	9,569
MSFragger	245,990	153,978	181,313	13,095	23,579	20,133	20,693

Figure 3. Performance of proteoM on the phosphopeptide dataset JHU_P2.

(**a**) Pearson correlation (ρ) and Manhattan distance (δ) of log2FC. (**b**) Summary in PSM and peptide identifications.

a



b

				PSM		Peptide		Sequence	
Engine	PSM	Peptide	Sequence	! Q%	! C%	! Q%	! C%	! Q%	! C%
proteoM	70,482	33,071	52,257	0	0	0	0	0	0
Mascot	59,181	27,114	44,779	4.87	0.68	3.16	0.38	22.16	11.10
MaxQuant	58,735	27,130	39,354	26.08	10.00	9.70	2.34	28.91	14.00

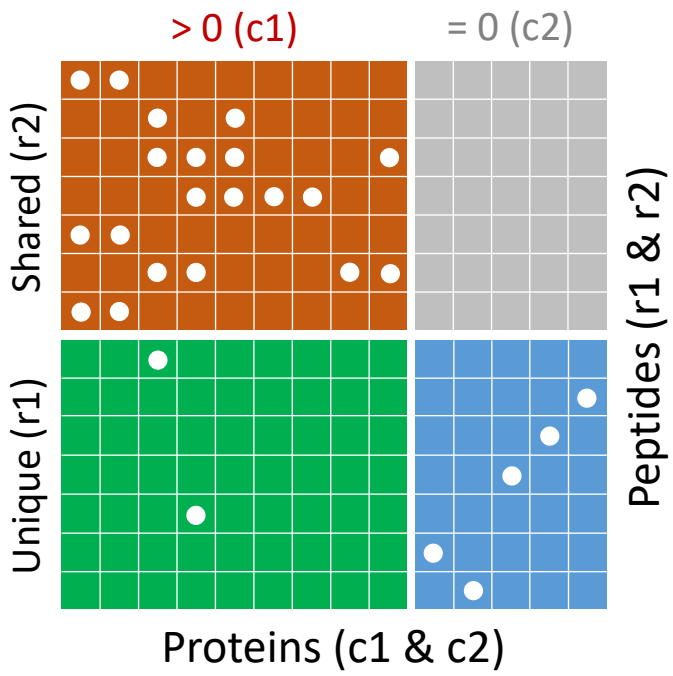



Figure 4. A schematic of peptide-to-protein grouping. Proteins and peptides entries are represented by columns and rows, respectively. The peptide-to-protein grouping is only affected by the intersection of protein columns with shared peptides and peptide rows that are shared by proteins (c1 × r2).