Peak identification and quantification by proteomic mass spectrogram decomposition

Pasrawin Taechawattananant[1], Kazuyoshi Yoshii[2,3], Yasushi Ishihama[1*]

1) Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto 606-8501, Japan.
2) Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan.
3) RIKEN Center for Advanced Intelligence Project (AIP), Tokyo 103-0027, Japan.

*Corresponding author:

Yasushi Ishihama, E-mail: yishiham@pharm.kyoto-u.ac.jp

Abstract

Recent advances in liquid chromatography/mass spectrometry (LC/MS) technology have notably improved the sensitivity, resolution, and speed of proteome analysis, resulting in increasing demand for more sophisticated algorithms to interpret highly complex mass spectrograms. We propose a novel application of mass spectrogram decomposition with a group sparsity constraint for joint identification and quantification of peptides and proteins. By incorporating protein-peptide hierarchical relationship knowledge, the isotopic distribution profiles of peptide ions, the learned noise subspace, and predicted retention time initialization into a standard mass spectrogram decomposition approach, we have significantly improved the accuracy of analysis. In benchmarking studies, our proteomic mass spectrogram decomposition showed excellent agreement [3277 peptide ions (94.79%) and 493 proteins (98.21%)] with the results of conventional identification and quantification based on Mascot and Skyline of *E. coli* cell lysate. This is the first application of proteomic mass spectrogram decomposition as a tool for LC/MS-based identification and quantification. Since pre-processing, such as thresholding, is not required, this proteomic mass spectrogram decomposition approach can maximize the efficiency of both protein and peptide identification and quantification.

INTRODUCTION

Liquid chromatography coupled with mass spectrometry (LC/MS) has been widely used to identify and quantify peptides generated by digestion of proteins in shotgun proteomics. Recent advances in LC/MS technologies have permitted the detection of a wide range of signal parameters in a three-dimensional mass spectrogram, including mass-to-charge ratio (*m/z*), retention time, and intensity. However, despite the high resolving power of modern LC/MS instruments, a large part of the mass spectrogram remains unresolved.[1] The common data analysis techniques start with pre-processing steps to simplify mass spectrograms but they inevitably lead to information loss or errors.[2] A new approach based on mathematical and statistical methods is required to interpret mass spectrograms accurately and with higher sensitivity.

Mass spectrograms can be viewed as a non-negative matrix representation in which peaks are continuously added. Target peaks in mass spectrograms are sparse, meaning that a relatively small proportion of interesting signals are present over a large space on the mass spectrogram, and their dynamic range can extend over many orders of magnitude. Multivariate curve resolution-alternating least-squares (MCR-ALS) was recently proposed for application to high-resolution LC/MS data with a simple sparsity constraint and without using a prior threshold.[3] The algorithm is well designed for analyzing sparse data and identifying low-intensity *m/z* signals. However, only eight resolved peak profiles from eight chemicals and four peak profiles from bacterial cell extracts were shown. Also, a pre-processing step of manual selection of chromatographic windows to reduce complexity is required. Moreover, the computation can produce negative intensities, which are inconsistent with the non-negative nature of the mass spectrogram. Various algorithms can analyze matrix data but they are not appropriate to interpret mass spectrograms.

In signal processing, non-negative matrix decomposition with a sparsity constraints is a machine-learning technique that is well applicable for identifying components of matrix input data. Briefly, the mixture of signals in matrix representation is decomposed into two matrices, namely a dictionary and an activation matrices. A dictionary can be predefined and possessed a group structure as shown in the work of Sun et al.[4] They referred their algorithm as supervised Block Kullback-Leibler Divergence Non-negative matrix Factorization (BKL-NMF) and applied to audio data. For LC/MS data, it was reported that application of non-negative matrix factorization enabled identification of 11 out of 18 mixed chemicals.[5] However, the simple design employed in that work has substantial limitations, separating only cleaned data with few common or overlapped peaks and disallowing a large number of chemical

compounds. It thus cannot be applied to complicated proteomic mass spectrograms. So far, no application of non-negative matrix decomposition to identify and quantify peaks of proteome-scale complexity has been reported to the best of our knowledge.

In this study, we propose proteomic mass spectrogram decomposition (protMSD), a new application of matrix decomposition to joint identification and quantification of peptides and proteins using shotgun proteomic samples. The protMSD method incorporates the isotopic *m/z* distribution, the learned noise subspace, and the possible retention times of peptides for reasonable initialization. We add a protein-peptide hierarchical relationship as group information for a group sparsity constraint. We first applied protMSD using a dictionary with all possible theoretical peptides and evaluated the performance with two standard samples. Similar to other complex data, there is more overlapping information in larger-scale proteomic mass spectrograms. Therefore, more prior knowledge is needed for accurate decomposition. Peptide detectability is a main factor to decide which peptides are identified.[6] We also applied protMSD using a dictionary incorporated peptide detectability information and analyzed a proteome-scale mass spectrogram. The protMSD method can increase the number of identifications, improve the accuracy of quantification, and analyze proteomic mass spectrograms without the need for complicated pre-processing, such as thresholding, which carries the risk of information loss during data analysis.

RESULTS AND DISCUSSION

1. Application to four-protein standard

1.1 Effects of MSD modifications

Application of the standard MSD method[5] to the four-protein mixture separated 1057 peptide ions. We observed a proportional relationship between the observed total ion currents (TICs) and the total standard MSD-activated signals at each time point (Figure 1), with a correlation coefficient of 0.846. On average, 20% of TICs were accounted for. To improve the performance of the algorithm, we developed protMSD to incorporate precursor ion isotopic distributions, noise models, sparsity constraints, and other factors as described in the Theory and Design section, and evaluated the correlation to the observed TICs. With protMSD, we obtained a correlation coefficient of 0.982 and accounted for 70% of the TICs. As expected by the additional incorporation of isotope peaks, we observed a significant improvement in the signal intensity of extracted ion chromatogram. The unaccounted-for signals were from peptides in undefined charged states and modifications, contaminants, and non-electronic noise outside the protMSD dictionary, since TICs included all signal contributions observed during LC/MS.

Next, we compared the peptide retention times approximated by the MSDs with those determined by Mascot[7] and Skyline[8], two conventional software programs used in proteomics. Mascot can perform database searching for identification using MS2 information and Skyline relies on Mascot results to perform quantification using MS1 information. We imported Mascot search results into Skyline MS1-Full Scan Filtering to obtain MS1 parameters. For the peptide ions commonly annotated by MSDs and Mascot/Skyline, the retention times by standard MSD agreed with those determined by Mascot/Skyline, with $y = 1.003x$, $R = 0.902$ (Supplementary Figure S1, left), while protMSD provided better accuracy, with $y = 1.000x$, $R = 0.999$ (Supplementary Figure S1, right). These results indicate that our modifications to standard MSD were effective for annotating peptide ions in the mass spectrogram.

1.2 Peptide identification

Peptide identification by current conventional approaches, including Mascot/Skyline, is based on the use of MS/MS spectra to acquire sequence-related information, meaning that the success of peptide identification depends to a large extent on the quality of the MS/MS spectra, whereas MSD-based approaches require the signal in MS1. The protMSD method identified 297 out of 303 (98.02%) Mascot-identified peptides (Figure 2A, left), and the

remaining six peptides had 52.29 times lower intensity, 20.78 lower Mascot Peptide Score (PeptScore), and 4.12% lower Skyline isotope dot product (idotp) than the peptides identified in common, calculated by median (Supplementary Figure S2). In other words, these six peptides have limited intensities in MS1. Moreover, protMSD identified an additional 299 peptide ions, and provided 100% sequence coverage for the four proteins, whereas the conventional method gave only 94.83% coverage on average (Supplementary Figure S3). Generally, MS1-based methods are more sensitive and can detect more peptides regardless of MS2 information, but MS1 information alone may not be enough for peptide identification if it yields poor profiles, such as distorted isotope patterns.

1.3 Protein and peptide quantification

We compared the peak intensities obtained by protMSD with those obtained by Skyline for commonly identified ions at the peptide level. The comparison at the protein level was done using the sum of the intensities of the common peptide ions in both methods. The quantification results for both methods agreed well with each other. Correlation coefficients of 0.997 and 0.914 were obtained at the protein and peptide levels, respectively (Figure 2A, right). We observed four peptide ion outliers calculated from Tukey's fences, or 1.5 times the interquartile range above the third quartile of the ratio of absolute intensity difference.

We identified two causes of the four outliers, exemplifying the advantage and disadvantage of the algorithms. For the outlier 1, Skyline and protMSD reported uncorrelated intensities of KLVAASQAALGL with +2 charge for 1.41e+05 at 42.1 min and 5.12e+09 at 41.60 min. According to Skyline data visualization, this peptide ion had three significant hits at 41.47, 42.10, and 43.13 min by Mascot, but the maximum MS1 intensity was located at 41.61 min, within the dynamic exclusion time. Since Skyline, which uses an MS2-triggered MS1-peak integrated approach, was able to find a small peak close to the significant hit at 42.10 min, it missed the true maximum intensity at 41.61 min because of the short separation distance. Consequently, Skyline underestimated the intensity from the subordinated peak due to dynamic exclusion (Supplementary Figure S4). In contrast, protMSD is a completely MS1-dependent approach, and can report the maximum MS1 peak intensity as an alternative peak quantification tool. The outliers 2, 3, and 4 resulted from HPYFYAPELLFFAKR with charges of +2, +3, and +4. This mixture contains RHPYFYAPELLFFAK, which shares the same precursor isotopic distribution profile as HPYFYAPELLFFAKR; thus, an MS1-based algorithm such as protMSD cannot distinguish between these two peptides. More features are necessary to quantify peptide isomers uniquely.

To validate the protMSD results, we employed a concatenated target-decoy approach[9] using a shuffle sequence to estimate the false discovery rate (FDR)[10] as N_D/N_T, where N_D is the number of decoys and N_T is the number of target-quantified peptides. We also determined the false quantification rate (FQR), calculated as I_D/I_T, where I_D is the decoy intensity and I_T is the target intensity. protMSD gave values of 0.4% FDR and 0.002% FQR.

2. Application to forty-eight-protein standard

2.1 Protein and peptide identification

UPS1 is a premixed standard containing 48 proteins, and represents a simple proteome with 9319 *in silico* peptide ions from 4045 unique peptides. In benchmarking with the conventional method, protMSD reported 1322 out of 1481 peptide ions or 89.26% of Mascot identification. In terms of accuracy in measuring the retention times of commonly identified peptide ions, the correlation between the two methods was excellent, with y = 0.999x, R = 0.986 (Supplementary Figure S5). The conventional method and protMSD exclusively identified 159 and 2103 peptide ions, respectively (Figure 2B, left). The complexity of the proteomic standard increased both exclusive identifications, presumably because of the larger proportion of low-intensity peptide ions and poor or overlapped MS1 peaks. We observed a 3.45 times lower intensity, 8.59 times lower PeptScore, and 3.37% lower idotp values for the 159 peptides exclusively identified by Mascot, similar to the case of the four-protein standard (Supplementary Figure S6). The conventional method and protMSD identified all proteins present in the sample, and yielded median sequence coverages of 62.02% and 87.29%, respectively (Supplementary Figure S7).

2.2 Protein and peptide quantification

We compared the peak intensities of 1322 commonly identified peptides obtained by the conventional method with those obtained by protMSD. For 1319 peptides, after eliminating three peptide ions with S/N ≤ 10, the quantification results from both methods agreed with each other. Correlation coefficients of 0.976 and 0.909 were obtained at the protein and peptide levels, respectively (Figure 2B, right). We observed 39 outlier peptides, or 2.96% of the total, calculated from Tukey's fences. In statistical validation, this application yielded values of 2.3% FDR and 0.8% FQR at the peptide level.

3. Proteome-scale application

The most frequently used approach to identify peptides is database searching[11], in which all protein sequences provided in the database are *in silico*-digested for reference. Database searching is simple, however it demands high computational power and memory. An alternative to database searching is library searching[12], in which the reference is constructed using a comprehensive collection of confident detections and identifications from previous experiments. With the reference as strong prior knowledge, library searching can be performed relatively quickly and provides better identification.[13–15]

We examined protMSD for analyzing a larger proteome of whole *E. coli* cell lysate with a 70-min gradient using a dictionary of all theoretical peptides. However, the memory on a common desktop was not sufficient. Exploiting the library searching concept, we improved protMSD with a peptide detectability knowledge. The reference library was the combination of the detected peptides analyzed from *E. coli* cell lysate with a 70-min gradient and with a 550-min gradient by Mascot at 1% false discovery rate (FDR) confidence. The dictionary was then constructed to cover all library peptides and extended to include every charge state possible within the *m/z* range. The result from the longer gradient separation provides a larger number of peptides for creating a comprehensive dictionary and offers retention times for initializing an activation matrix. We converted observed retention times to reference retention times by means of linear regression with ±2.5 min window.

3.1 Protein and peptide identification

In benchmarking with the conventional method, protMSD identified 3286 out of 3509 or 93.64% of Mascot-identified peptides (Figure 3A, left). The correlation between the two methods for measuring the retention times of commonly identified peptide ions was high, with $y = 0.999x$, $R = 0.996$ (Supplementary Figure S8). At the protein level, protMSD identified more than twice as many proteins as Mascot and covered 497 out of 502 (99.00%) Mascot identifications (Figure 3A, right). The conventional method and protMSD provided median sequence coverages of 14.80% and 25.12%, respectively (Supplementary Figure S9).

3.2 Protein and peptide quantification

Finally, we quantified the peak intensities of commonly identified peptide ions and proteins. After eliminating peaks with $S/N \leq 10$, correlation coefficients of 0.982 and 0.941 were obtained at the protein and peptide levels, respectively (Figure 3B). We also evaluated the protein quantifications obtained with each method separately and obtained a correlation coefficient of 0.987 (Supplementary Figure S10).

Due to the nature of protMSD, forcing the well-constructed dictionary to include a large number of false peptides weakens the algorithm performance, leading to the problem of FDR overestimation. In order to validate the power of protNMF without the help of any external preprocessing, we also employed a concatenated target-decoy approach. The decoy was generated from highly detected Hela peptide lists and concatenated to the *E. coli* dictionary with the same number of peptide ions. We obtained 5.4% FDR and 0.5% FQR at the peptide level in this application. The low FQR value indicates that false discovery peaks in protMSD are generally small and might be filtered out by appropriate prior knowledge.

4. Computational performance

The computational cost of protMSD depends on the sizes of the matrices and the degree of sophistication in transforming prior knowledge to construct the dictionary or initialize the activation matrix.

For the four-protein standard, standard non-regularized algorithm required an approximate processing time of 4.50 min on the system we used.protMSD required a processing time of 7.30 min. Specifically, protMSD took the extra steps in learning noise by NMF, initializing the H matrix, and updating protMSD with a group sparsity constraint. For the forty-eight-protein standard, protMSD took 45.78 min. Despite the size of the computational tasks, which involve several million elements, the computation time and space required by protMSD are manageable on a common desktop computer, based on code optimization and appropriate choices of libraries for sparse computation, particularly, scipy-sparse.[15] For proteome-scale application using the *E. coli* sample, protMSD with the concept of library searching could reduce the running time to 24.17 minutes for all processing. The memory footprints for all sample processings were less than 2 GB.

EXPERIMENTAL SECTION

THEORY AND DESIGN

The work of Rapin et al.[5] applied standard non-regularized MSD, which was originally described as non-negative matrix factorization (NMF)[16], to decompose LC/MS data containing non-negative signals in a matrix V into two non-negative matrices, $W \in \mathbb{R}^{m\times k}$ and $H \in \mathbb{R}^{k\times t}$, whose product approximates the LC/MS data.

$$V \simeq WH$$

W is a reference mass spectra profile which can be called as a dictionary, and H is a retention time profile which can be called an activation matrix, in which the activation at frame H_j represents the identification and the quantification results. The work of Sun et al.[4] shows that either both matrix results can be iteratively calculated or one of matrix results can be predefined, which are referred to as semi-supervised or supervised NMF, respectively.

If a sparse dictionary W is predefined, standard MSD algorithm will be calculated only H by minimizing the difference between V and WH. The difference $D(V||WH)$ can be quantified by a cost function D, such as the KL divergence. Note that even when W is given, H should be estimated with an iterative optimization algorithm as in NMF, because V / W is not mathematically defined.

$$\underset{W\geq 0, H\geq 0}{\text{minimize}}\, D(V||WH)$$

In this work, we first examined the potential of standard MSD using KL divergence for representing an observed proteomics mass spectrogram as a matrix V (*m/z* × retention time) by employing a predefined dictionary of peptide signals with reference mass spectra of monoisotopic *m/z* of all theoretical peptide ions as matrix W (*m/z* × theoretical peptide ion). The resulting matrix H (theoretical peptide ion × retention time) containing the activation of each theoretical peptide ion over the retention time was initialized randomly.

To tailor the decomposition for proteomic mass spectrogram, we then designed protMSD (Figure 4). We first incorporated two subspaces into the dictionary: the peptide signal subspace W_S and the noise subspace W_N. This subspace partitioning prevents noise from interfering with interesting signal approximations.[17] W_S was constructed by using reference mass spectra with isotopic *m/z* of theoretical peptides. Since mass spectrograms are commonly subjected to considerable but fairly constant background noise, a few constant

noise spectra for W_N can be learned from NMF with a small k numbers using a blank region of mass spectrogram. For H, appropriate initialization enhances the accuracy and reduces the complexity of the algorithm. Instead of using random initialization, we initialized H using predicted retention times.

Since proteomic mass spectrograms have a unique sparsity structure arising from the protein and peptide hierarchy, protMSD can exploit a group sparsity constraint Ω.[18] We imposed Ω on each protein using $\log/L1$ from the work of Sun et al due to its simplicity and good performance. For a weight factor λ, instead of tuning as Sun, the sparseness measure was introduced and applied for controlling the level of sparsity.[19,20] Also, for proteomic data, the number of peptides in the group of each protein are vary and can bias the algorithm result. λ was calculated by the sparseness measure, normalized by the number of theoretical peptide ions per protein, and incorporated into Ω. protMSD aims to solve:

$$\underset{W\geq 0, H\geq 0}{\text{minimize}}\, D(V||WH) + \Omega(H_S)$$

The iterative algorithm can be derived using majorization-minimization. protMSD is guaranteed to converge to a local optimum. The technical details are described in Supplementary Methods.

WORKFLOW

The protMSD workflow comprises the following four steps.

Step 1. Matrix construction

[1.1] The observed mass spectrogram V:

Either an mzML or a text file of a raw file converted by ProteoWizard MSconvert[21] can be processed by protMSD. The intensity values are normalized by mean and assigned to V at appropriate *m*/*z* and retention time bins. Thresholding is not performed, so that low-abundance peptide ions are included in the computation.

[1.2] The dictionary W:

All theoretical peptide ions from *in silico* tryptic digestion of expected proteins or from the experimentally collected reference library are featured with six isotopic peaks, calculated, normalized, and assigned to columns of W_S. The learned noise patterns from a blank region of the mass spectrogram, i.e., the last 10 min of the LC/MS run, are computed by separate NMF and concatenated as W_N.

[1.3] The activation H:

The activation of each peptide is restricted to zero at retention time frames in which the peptide peaks are highly unlikely to be detected. In case of standard non-regularized algorithm, random initialization is used. In the case of protMSD with a database-based dictionary, two initializations are combined using the predicted retention times and the similarities to the theoretical isotopic distribution. In the case of protMSD with a library-based dictionary, only the predicted retention times are used. For H_N, since noise time frames are unpredictable, the noise subspace is fully initialized.

Step 2. Sparsity constraint weight calculation

Group sparsity constraint is calculated by use of a sparseness measure and normalized by the number of peptides per protein.

Step 3. protMSD update

H is updated iteratively until convergence with a tolerance value of 1e-4.

Step 4. Post-processing for interpretation

The resulting H is multiplied by the mean to reverse the normalization in 1.1. The most confident peak in H_k, which has an activated intensity with a signal-to-noise ratio (S/N) > 3, is reported as a peptide candidate for identification and the criterion of S/N > 10 is used for quantification. Protein identification is reported if there are at least two identified peptides, and quantification is calculated based on the sum of activated intensities of the reported peptides.

Further details of matrix construction, weight calculation formula, protMSD update algorithm, and post-processing technique are provided in Supplementary Methods.

Mass spectrograms with three different levels of complexity were evaluated in this study. Two standard protein mixtures were prepared from four proteins and a universal proteomic standard (UPS1) containing 48 proteins. For proteome-scale analysis, *E. coli* cell lysate data were used.[22] Further details of the experimental setup for sample preparation as well as for LC/MS are given in Supplementary Methods.

All computations were performed on a desktop computer with a 3.30 GHz E3-1226 v3 4-core processor and 32 GB main memory. The raw MS data were deposited at the ProteomeXchange Consortium via jPOST partner repository[19] with identifier JPST000663 for two standard proteins and JPST000765 for *E. coli* lysate. protMSD is available in Python 2.7 at GitHub https://github.com/pasrawin/ProteomicMSD and https://github.com/pasrawin/LibraryProteomicMSD.

CONCLUSIONS

We have applied and modified the matrix decomposition algorithm to efficiently analyze three-dimensional mass spectrograms obtained from proteomic studies. Our protMSD incorporates isotopic distribution, learns noise, exploits a protein-peptide hierarchical relationship by using a group sparsity constraint, and configures a reasonable initialization by using possible retention times. The proposed update rule with the constraint of choice always remains non-negative and guarantees convergence to a local optimum. Therefore, with good initialization, protMSD converges to the good identification result.

In the case of the four-protein standard, protMSD gave a better-resolved chromatogram with a higher accuracy of peptide retention times than did standard non-regularized algorithm. In the cases of the forty-eight-protein standard and *E. coli* samples, protMSD provided the excellent correlations with the conventional Mascot/Skyline methodology for common identification and quantification without the need for thresholding. In additional, protMSD increased the number of identified peptides and enabled quantification of more than a thousand proteins. Our results show that protMSD with the library-based dictionary can achieve rapid and large-scale proteome analysis. In this study, we have focused on protein identification and quantification, but the same approach should be readily applicable for various other purposes related to the interpretation of mass spectrograms.

The present results indicate that the matrix decomposition algorithm-based approach is very effective for mass spectrometry-based proteome analysis. Further improvements should also be possible by incorporating other significant features, such as product ions from proteomic mass spectrograms. Work along this line is already underway in our laboratory.

Acknowledgements

This work was supported by the CREST Strategic Basic Research Program from Japan Science and Technology Agency (JST), No. 18070870, and the Database Integration Coordination Program from the National Bioscience Database Center, JST No.18063028.

Supporting Information Available

The Supporting Information is available free of charge on the ACS Publications website.

- Supplementary Figures: S1−S10; Supplementary Methods: protMSD techniques and experiment setup (PDF)

Figure Legends

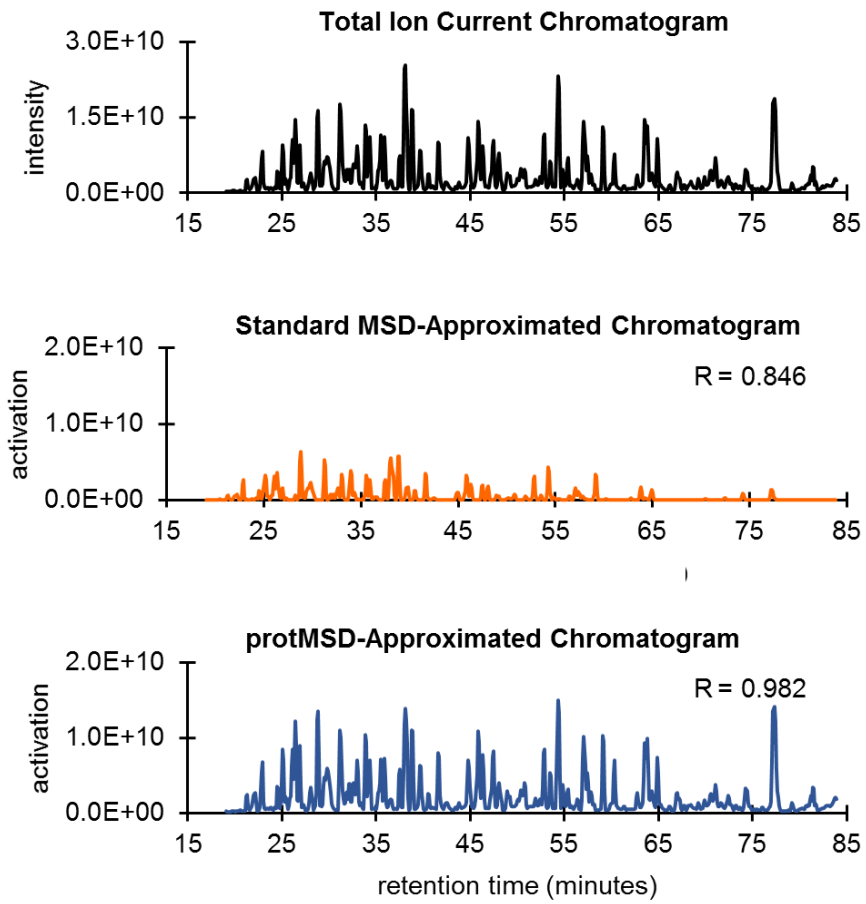


Figure 1. Effects of protMSD on chromatogram extraction.

The observed total ion current chromatogram (top). Standard MSD-approximated chromatogram (middle). protMSD-approximated chromatogram (bottom). Correlation coefficients were calculated between the observed total ion currents and total activated intensities of each algorithm at the same retention time points.

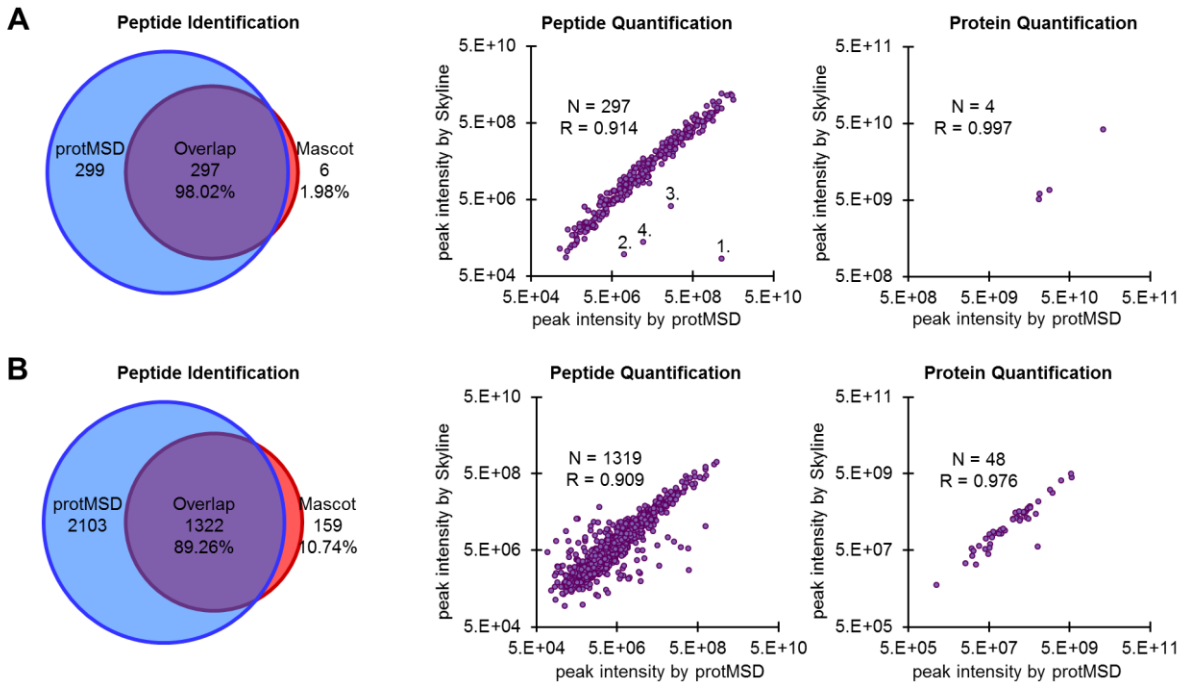


Figure 2. protMSD for the four-protein standard (A) and for the forty-eight-protein standard (B).

(A) The Venn diagram represents the peptide ions identified using protMSD and Mascot (left). The peak intensities of peptide ions commonly identified by protMSD (x-axis) and Skyline (y-axis) show correlation coefficients of 0.914 with four outliers at the peptide level, and 0.997 at the protein level (right).

(B) The Venn diagram represents the peptides identified using protMSD and Mascot (left). The peak intensities of peptide ions commonly identified by protMSD (*x*-axis) and Skyline (*y*-axis) show correlation coefficients of 0.909 at the peptide level, and 0.976 at the protein level (right).

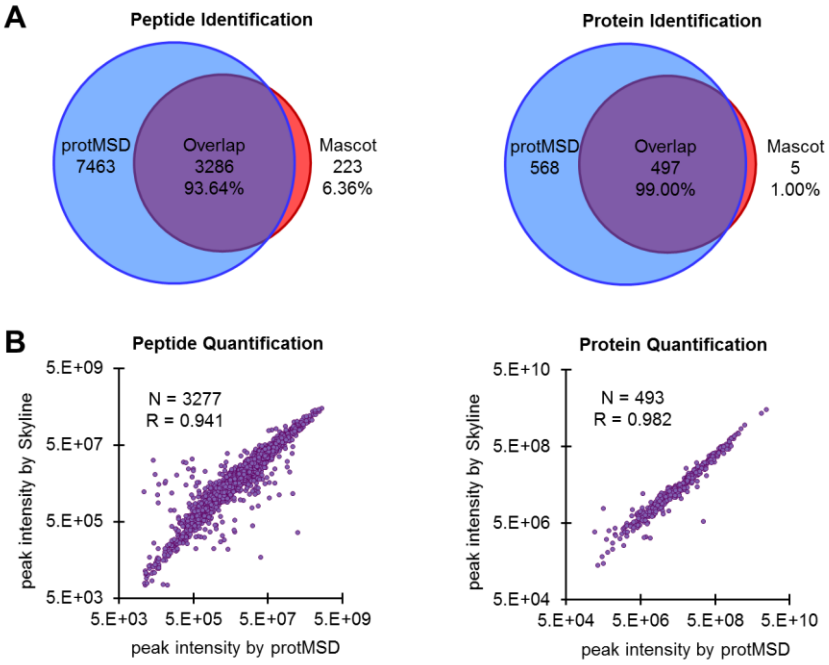


Figure 3. protMSD for proteome-scale application.

(A) The Venn diagrams represent the *E. coli* peptide ions and proteins identified using protMSD and Mascot.

(B) The peak intensities of peptide ions commonly identified by protMSD (x-axis) and Skyline (y-axis) show correlation coefficients of 0.941 at the peptide level and 0.982 at the protein level.

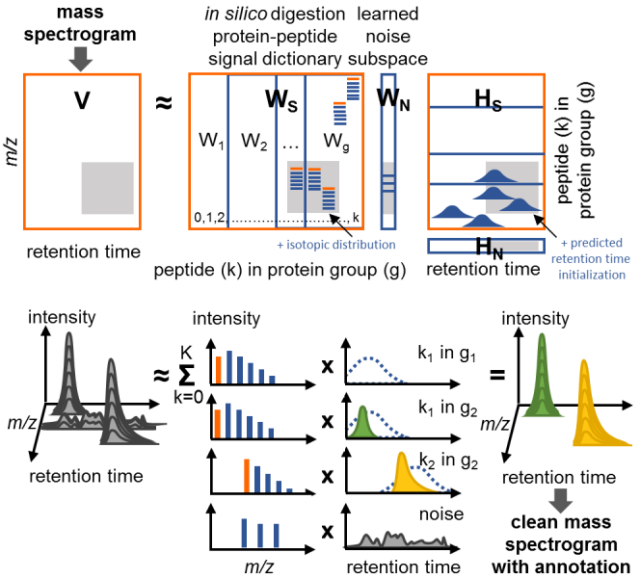
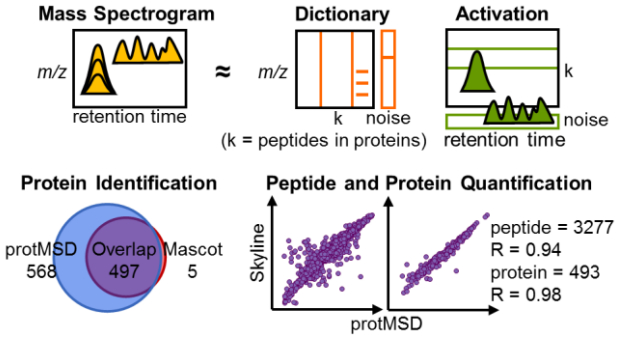


Figure 4. Overview of protMSD.

In the standard non-regularized algorithm, three matrices are designed to correspond to an observed mass spectrogram V (*m/z* × retention time), a given dictionary of theoretical peptide monoisotopic *m/z* W (*m/z* × theoretical peptide ion), and an activation matrix H (theoretical peptide ion × retention time) for updating the results (shown in orange). For protMSD, the dictionary incorporates two subspaces: the peptide subspace W_S and the noise subspace W_N. The W_S column includes monoisotopic *m/z* and subsequent isotopic *m/z*, and W_N contains learned noise patterns. For the activation matrix at H_S, instead of random initialization, the activations are restricted to zero outside possible retention time frames (shown in blue) (top).

Specifically, the observed mass spectrogram V can be viewed as the linear sum of the isotopic distribution profile of each peptide k in the dictionary W_S and the noise profile in W_N scaled by retention time-varying activations. The group sparsity constraint uses protein and peptide selection to identify a few highly related patterns to approximate the mass spectrogram (bottom).

For Table of Contents Only



References

(1) Michalski, A.; Cox, J.; Mann, M. More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority Is Inaccessible to Data-Dependent LC-MS/MS. *J. Proteome Res.* **2011**, *10* (4), 1785–1793. https://doi.org/10.1021/pr101060v.

(2) Tsai, T.-H.; Wang, M.; Ressom, H. W. Preprocessing and Analysis of LC-MS-Based Proteomic Data. *Methods Mol Biol* **2016**, *1362*, 63–76. https://doi.org/10.1007/978-1-4939-3106-4_3.

(3) Cook, D. W.; Rutan, S. C. Analysis of Liquid Chromatography-Mass Spectrometry Data with an Elastic Net Multivariate Curve Resolution Strategy for Sparse Spectral Recovery. *Anal. Chem.* **2017**, *89* (16), 8405–8412. https://doi.org/10.1021/acs.analchem.7b01832.

(4) Sun, D. L.; Mysore, G. J. Universal Speech Models for Speaker Independent Single Channel Source Separation. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*; 2013; pp 141–145. https://doi.org/10.1109/ICASSP.2013.6637625.

(5) Rapin, J.; Souloumiac, A.; Bobin, J.; Larue, A.; Junot, C.; Ouethrani, M.; Starck, J.-L. Application of Non-Negative Matrix Factorization to LC/MS Data, 2015.

(6) Li, Y. F.; Arnold, R. J.; Tang, H.; Radivojac, P. The Importance of Peptide Detectability for Protein Identification, Quantification, and Experiment Design in MS/MS Proteomics. *J Proteome Res* **2010**, *9* (12), 6288–6297. https://doi.org/10.1021/pr1005586.

(7) Anderson, N. L.; Anderson, N. G. Proteome and Proteomics: New Technologies, New Concepts, and New Words. *Electrophoresis* **1998**, *19* (11), 1853–1861. https://doi.org/10.1002/elps.1150191103.

(8) MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J. Skyline: An Open Source Document Editor for Creating and Analyzing Targeted Proteomics Experiments. *Bioinformatics* **2010**, *26* (7), 966–968. https://doi.org/10.1093/bioinformatics/btq054.

(9) Elias, J. E.; Gygi, S. P. Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics. *Methods Mol Biol* **2010**, *604*, 55–71. https://doi.org/10.1007/978-1-60761-444-9_5.

(10) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases. *J. Proteome Res.* **2008**, *7* (1), 29–34. https://doi.org/10.1021/pr700600n.

(11) 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.* **1994**, *5* (11), 976–989. https://doi.org/10.1016/1044-0305(94)80016-2.

(12) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. Development and Validation of a Spectral Library Searching Method for Peptide Identification from MS/MS. *Proteomics* **2007**, *7* (5), 655–667. https://doi.org/10.1002/pmic.200600625.

(13) Craig, R.; Cortens, J. C.; Fenyo, D.; Beavis, R. C. Using Annotated Peptide Mass Spectrum Libraries for Protein Identification. *J. Proteome Res.* **2006**, *5* (8), 1843–1849. https://doi.org/10.1021/pr0602085.

(14) Frewen, B. E.; Merrihew, G. E.; Wu, C. C.; Noble, W. S.; MacCoss, M. J. Analysis of Peptide MS/MS Spectra from Large-Scale Proteomics Experiments Using Spectrum Libraries. *Anal. Chem.* **2006**, *78* (16), 5678–5684. https://doi.org/10.1021/ac060279n.

(15) Deutsch, E. W.; Perez-Riverol, Y.; Chalkley, R. J.; Wilhelm, M.; Tate, S.; Sachsenberg, T.; Walzer, M.; Käll, L.; Delanghe, B.; Böcker, S.; Schymanski, E. L.; Wilmes, P.; Dorfer, V.; Kuster, B.; Volders, P.-J.; Jehmlich, N.; Vissers, J. P. C.; Wolan, D. W.; Wang, A. Y.; Mendoza, L.; Shofstahl, J.; Dowsey, A. W.; Griss, J.; Salek, R. M.; Neumann, S.; Binz, P.-A.; Lam, H.; Vizcaíno, J. A.; Bandeira, N.; Röst, H. Expanding the Use of Spectral Libraries in Proteomics. *J. Proteome Res.* **2018**, *17* (12), 4051–4060. https://doi.org/10.1021/acs.jproteome.8b00485.

(16) Oliphant, T. E. Python for Scientific Computing. *Computing in Science Engineering* **2007**, *9* (3), 10–20. https://doi.org/10.1109/MCSE.2007.58.

(17) Lee, D. D.; Seung, H. S. Algorithms for Non-Negative Matrix Factorization. *Adv. Neural Inf. Process. Syst.* 556–562.

(18) Wilson, K. W.; Raj, B.; Smaragdis, P.; Divakaran, A. Speech Denoising Using Nonnegative Matrix Factorization with Priors. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*; 2008; pp 4029–4032. https://doi.org/10.1109/ICASSP.2008.4518538.

(19) Kim, J.; Monteiro, R. D. C.; Park, H. Group Sparsity in Nonnegative Matrix Factorization. In *Proceedings of the 2012 SIAM International Conference on Data Mining*; Society for Industrial and Applied Mathematics, 2012; pp 851–862. https://doi.org/10.1137/1.9781611972825.73.

(20) Hoyer, P. O. Non-Negative Matrix Factorization with Sparseness Constraints. *J. Mach. Learn. Res.* **2004**, *5*, 1457–1469.

(21) Qian, Y.; Jia, S.; Zhou, J.; Robles-Kelly, A. Hyperspectral Unmixing via L1/2 Sparsity-Constrained Nonnegative Matrix Factorization. *IEEE Transactions on Geoscience and Remote Sensing* **2011**, *49* (11), 4282–4297. https://doi.org/10.1109/TGRS.2011.2144605.

(22) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M.-Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P. A Cross-Platform Toolkit for Mass Spectrometry and Proteomics. *Nat. Biotechnol.* **2012**, *30* (10), 918–920. https://doi.org/10.1038/nbt.2377.

(23) Iwasaki, M.; Miwa, S.; Ikegami, T.; Tomita, M.; Tanaka, N.; Ishihama, Y. One-Dimensional Capillary Liquid Chromatographic Separation Coupled with Tandem Mass Spectrometry Unveils the Escherichia Coli Proteome on a Microarray Scale. *Anal. Chem.* **2010**, *82* (7), 2616–2620. https://doi.org/10.1021/ac100343q.