mspms: a R package and graphical interface for the processing and analysis of multiplex substrate profiling by mass spectrometry for proteases (MSP-MS) data.

Charlie Bayne1,2,3, Brianna Hurysz1,2,3, David J. Gonzalez2,3,4, Anthony O'Donoghue2,

Institutions:

1Biomedical Sciences Graduate Program, UC San Diego, La Jolla, California 92093, USA.

2Department of Pharmacology, University of California San Diego, La Jolla, California 92093, USA

3Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California, USA

4Center for Microbiome Innovation, University of California at San Diego, La Jolla, California 92093, USA

Abstract:

Multiplex Substrate Profiling by Mass Spectrometry (MSP-MS) is a powerful method used to determine the substrate specificity of proteases. This method is of interest for many groups interested in the study of proteases and their role as regulators of many biological pathways whether applied to the study of disease states, the development of diagnostic and prognostic tests, generation of tool compounds, or rational design of protease targeting therapeutics. Analysis of the MS based data produced by MSP-MS is a multistep process involving detection and quantification of peptides, normalization, outlier detection, imputation, cleavage sequence identification, statistics, and data visualizations. This process can be challenging, especially for biologists/ mass spectrometrists with limited programming experience. To overcome these issues, we provide the mspms R package alongside a companion graphic user interface hosted at https://gonzalezlab.shinyapps.io/mspms\_shiny/ to facilitate the analysis of MSP-MS data utilizing good software/data analysis practices. Here we show the application of the mspms R package to determine the sequences specificity of 6 well-studied cathepsins using the MSP-MS method. We show that a mspms based analysis is able to uncover expected sequence specificities, facilitates reproducible research, and has expanded capabilities relative to previous MSP-MS analysis tools.

Introduction:

Multiplex substrate profiling by mass spectrometry for proteases (MSP-MS) is a method for determining protease substrate specificity1. This method works by utilizing a rationally designed library of peptide sequences, incubating them with a protease directly or a protease-containing biological sample, treating with any desired experimental conditions (e.g. inhibitor addition, buffer pH variations, etc.), and then detecting and quantifying the cleaved peptides through mass spectrometry 2.

MSP-MS is applicable to a wide group of researchers as it only requires a synthetically created peptide library, mass spectrometer, and samples containing protease(s) to perform. While relatively straight forward conceptually, MSP-MS produces high dimensional data that can be very challenging for biologists or mass spectrometrists to analyze. Adequate interpretation of the MSP-MS data requires several steps. The data must be median normalized, outliers detected and removed, missing values imputed, cleavage motifs relative to the peptide library recognized, and then downstream statistics and visualizations must be performed.

A central component of any experimental method involving complex data is a robust data analysis pipeline. Poorly documented analysis code poses a number of challenges. It has been a prevalent problem in the biological research world, leading to the retraction of several high-profile papers in recent years3. There is also a logistical problem, as it limits the portability/ reproducibility of analysis lab to lab since it requires specific knowledge from the programmer to successfully utilize. This is commonly a problem in academic labs, where a researcher may develop code widely used for analysis within the lab before leaving for a new opportunity. Often, this results in an opaque codebase that may be treated as a black box and can ultimately result in erroneous data analysis, especially as the original code gets distorted over time through decentralized use by various researchers. Solutions for this exist in the domain of software engineering 4, but often times they are not successfully implemented in biology centric labs, due to the background knowledge required and the amount of time it would take to develop.

To overcome these issues, we developed mspms: a well-documented, portable, and reproducible R package for the analysis of MSP-MS data. This package brings good software/data analysis practices to MSP-MS data analysis. Recognizing there may be a wide community of researchers interested in the MSP-MS method who find the R programming environment to be inaccessible, we also developed mspms-shiny, a graphical user interface to the core features of the mspms package. Mspms-shiny is available on the web at https://gonzalezlab.shinyapps.io/mspms\_shiny/, or by downloading and running locally on the researcher’s computer of choice. These user-friendly tools in combination with the powerful MSP-MS method offer a robust resource that is widely applicable for the protease research community.

Here we describe the core features of the package and demonstrate their effectiveness through the analysis of previously published MSP-MS data from our group profiling several well-studied cathepsin proteases. We profile the serine protease cathepsin A, the aspartic protease cathepsin D (bovine), and the cysteine proteases cathepsins B (human and bovine), C, and L.

Cathepsins are the most abundant lysosomal proteases that possess a range of substrate specificities5. Of the cathepsins profiled here, cathepsin A possesses carboxypeptidase activity6, cathepsin B: dipeptidyl carboxypeptidase activity7, cathepsin C: dipeptidyl aminopeptidase activity8, cathepsin D: endoprotease activity with preference for hydrophobic residues on both sides of the scissile bond (P1 and P1’)9,10, and cathepsin L: endopeptidase activity with preference for aromatic residues in P2 and hydrophobic residues in P3 position11. As the specificity of these cathepsins have been well- documented in the literature, these *apriori* known substrate specificities are an ideal data source to benchmark the results of a mspms facilitated analysis.

Materials and Methods

**Package Structure**

mspms contains 4 categories of functions: those that are responsible for preprocessing the data, normalizing/processing data, performing statistics, and visualizing the data.

**Data Preprocessing**

These functions intend to extend mspms analysis to the largest number of circumstances possible. Many of these functions are developed based on issues raised by specific research groups on GitHub to address their specific needs.

*Preprocessing*

Mspms is designed to be downstream of computationally intensive solutions for detecting and quantifying peptides. These solutions generally produce outputs that contain the same information but are formatted differently. To make mspms compatible with all of these different software solutions we provide functions that parse them into a consistent format that is then capable of being operated on by downstream functions. Prepare\_peaks() and prepare\_pd() function on exported data from PEAKS12 and Proteome Discoverer13 respectively.

*All possible cleavages of peptide library*

It is necessary to determine all of the possible cleavages of a defined length in a peptide library for downstream analyses such as recognition motif visualization. We provide the calculate\_all\_cleavages() function for this use case. This function takes a vector of given peptide sequences, and then splits them up into all possible cleavages of a specified length.

**Data Processing/ Normalization**

*Data Normalization*

Median normalization is performed as implemented in the NormalyzerDE R package14. The intensity from a peptide in a sample is first divided by the median of all peptides in the sample. This value is then multiplied by the mean of median of sum of intensities of all peptides in all samples.

*Outlier Removal*

Outliers are detected using a dixon test as implemented in the outliers R package 15. Values determined to be outliers (p < 0.05) are converted to NA.

*Data Imputation*

Values for missing data (peptides detected as 0) are imputed. First, a univariate distribution is fit on the lowest range of values as determined by noise \* length using the MASS R package 16. Then values are imputed using these parameters utilizing the truncated normal distribution as implemented in the truncnorm R package 17. Outlier values are not imputed.

*Cleavage Motif Determination*

The experimentally detected peptides are assigned to the corresponding full length library peptide using the join\_with\_library() function. Cleavage motifs are then generated for each detected peptide using the add\_cleavage() function. This function reports amino acids on each side of the cleavage site, with the option to set a user defined distance from the scissile bond. If there is not an AA at that position of the peptide library because it is past the terminal end, that position of the cleavage motif is represented as a X. This is performed for detected peptides that were cleaved at the N terminus, the C terminus, and both the N and C terminus.

*Consolidating Data*

The prepare\_for\_stats() function converts data from a wide to long format, and appends all of the metadata contained in the design matrix. The polish() function then removes peptides that have been cleaved on both the N and C terminus, and then combines all cleavage motifs into one column of the data frame to facilitate downstream analysis.

**Standard normalization workflow**

We provide mspms() as a convenience function to run all aspects of a standard mspms data normalization in one function. This function takes preprocessed data, the design matrix, the number of cleavages, and the peptide library used in the experiment as arguments and combines data normalization, outlier removal, data imputation, cleavage motif determination, and data consolidation.

**Statistics**

There are many statistics that could be performed on MSP-MS data depending on the experimental design and intent. Most are best left to the researcher to implement on their own. We provide convenient implementations of a small subset of statistical tests that we have found to be generally useful across all MSP-MS experiments.

*Anova*

An anova is performed for each peptide grouped by experimental condition to test for an effect of time as implemented in the Rstatix package. P values are subsequently FDR corrected.

*Log2 Fold Change*

The log2 fold change for each condition relative to time 0 is calculated by the mspms\_log2fc() function.

*T tests*

Pairwise t-tests are performed as implemented in the Rstatix package 18 and are subsequently FDR corrected. A number of functions are provided for different use cases.

* mspms\_t\_tests(): calculates t-tests on data relative to time 0.
* log2fc\_t\_test(): perfoms t-tests and log2 fold change on data relative to time 0
* log2fct\_time(): calculates the log2 fold change and t-tests for each peptide keeping the condition constant and comparing user specified times.
* log2fct\_condition(): calculates the log2 fold change and t-tests for each peptide keeping the time constant and comparing user specified conditions.

*Count of cleavage position*

We provide the count\_cleavages\_per\_pos() function to summarize the number of peptides in a data frame by the position in the peptide library they were cleaved at.

*Icelogo*

We provide a series of functions to process data in order to create an Icelogo19.

* calc\_AA\_count\_of\_motif(): given a vector of inputs, this function returns a matrix of the counts of each amino acid at each position.
* calc\_AA\_prop\_of\_motif(): Given a matrix containing the count of amino acids at each position, this function calculates the proportions.
* calc\_AA\_motif\_zscore(): Calculates the zscores given the count and proportion of the amino acids at each position of a set of cleavages in reference to a background set.
* calc\_sig\_zscores(): given a matrix of zscores, determines which are significant given a user defined p value.
* calc\_AA\_percent\_difference(): calculates the percent difference of a matrix containing the proportions of amino acids at each position relative to a background matrix.
* calc\_AA\_fold\_change(): calculates the fold change of a matrix containing the proportions of amino acids at each position relative to a background matrix.
* prepare\_sig\_p\_diff(): filters matrix of amino acid counts by position to only contain those that are significant.
* prepare\_pd(): prepares a matrix with significant amino acid percent differences by position for plotting.
* prepare\_fc(): prepares a matrix with significant amino acid fold changes by position for plotting.
* prepare\_icelogo\_data(): combines the functions described above to generate an icelogo matrix of significantly enriched amino acids at each position of a cleavage motif.
* extract\_re(): extracts regular expressions specifying amino acids above a defined threshold cutoff of each position of a cleavage motif from an icelogo matrix.

**Data Visualization**

Similar to our approach with statistics, we provide functions for plotting a subset of graphs that we have found to be generally useful for MSP-MS experiments.

*Heatmap/* *Hierarchical Clustering*

We provide plot\_heatmap() to visualize a heatmap of the normalized intensity, where the rows and columns have been subjected to hierarchical clustering and the color of the cells represent the values the data as implemented in the heatmaply R package20. Euclidian distance and the complete agglomeration method are used to perform the hierarchical clustering.

*Principal Component Analysis*

We provide plot\_pca() as a convenience function to visualize the first two principal components of the data colored by time, and label shapes showing the experimental condition using the ggplot2 package21. Only complete cases are considered, the peptides determined to have outliers are omitted.

*Visualization of Sequence Specificity by Icelogo plots*

We provide plot\_icelogo() to create an Icelogo relative to the background of all possible cleavages in the experiment. A vector of user supplied sequence motifs are compared to the background of all possible combinations in the peptide library used for the experiment. The ggseqlogo R package 22 is then used to plot the underlying data.

The plot\_all\_icelogos() function generates icelogos representing enriched sequences relative to time 0 for each condition. This function compares cleavage sequences corresponding to peptides that are significant (p.adj < 0.05 and log2 fold change > 3) to a background of all possible sequences present in the initial peptide library.

*Time Course*

We provide plot\_time\_course() as a convenience function to visualize the intensity of peptides over time by condition using the ggplot2 package.

*Cleavage Position*

We provide the plot\_cleavage\_per\_pos() as a function to easily visualize the number of cleavages per position of the peptide libarary using the ggplot2 package.

**Standardized Report**

We provide the generate\_report() function to generate a .html file containing a standardized set of plots rendered from a .RMD template file. This function takes processed data, the design matrix, the number of cleavages, and the peptide library used in the experiment as arguments. This output is intended to be a standardized method to produce a generic mspms analysis that can quickly and easily be applied to any experiment.

**Data Used for Study**

Data was acquired from MASSIVE from a previously ran, analyzed, and published MSP-MS experiment (PMID: [34153188](https://pubmed.ncbi.nlm.nih.gov/34153188)). In brief, samples were incubated with the peptide library such that each of the 228 peptides were at a concentration of 0.5 μM. After incubation at 37°C and at respective timepoints, the reaction was quenched by addition of GuHCl. These samples were desalted with C18 columns, and ~0.4 μg of each sample was subjected to LC-MS/MS with an Ultimate 3000 HPLC and Q-Exactive mass spectrometer using the previously reported parameters (Thermo Fisher Scientific, PMID: [34153188](https://pubmed.ncbi.nlm.nih.gov/34153188)).

**Peaks Processing**

Data from all mass spectrometry runs was processed using PEAKS software, ultimately producing a protein-peptides-id.csv and protein-peptides-lfq.csv file. For each sample experiment specific parameters were specified as follows: Q-Exactive instrument, HCD fragmentation, no enzyme. Scans were merged with a retention time window of 0.8 min, and precursor m/z error tolerance of 10 ppm. Precursor mass was corrected. Scans were filtered to include retention time between 0 and 95 min with a precursor mass tolerance of 10 ppm. For identification, a precursor mass tolerance of 20 ppm using monoisotopic mass and a fragmentation ion of 0.01 Da was specified. No PTMs were included in the search. The database used was the peptide library used in the MSP-MS experiment (add this file to supplementary info). FDR was estimated using decoy- fusion strategy. Label free quantification was performed with a mass error tolerance of 9 ppm, retention time shift tolerance of 3 minutes. Replicate samples were added to new groups.

The protein-peptides-lfq.csv file was prepared by navigating to the quantification options setting the normalization factor to “No normalization”, changing peptide filters to include all peptides (quality >= 0, Avg.Area >= 0, Peptide ID Count >=0, charge +1 - +10, and at least 1 confident sample). Protein filters were changed so no filtering occurs (Significance >=0). Supporting peptides were then exported and file name was changed to protein-peptides-lfq.csv.

The protein-peptides-id.csv file was prepared by adjusting the FDR setting to 1%. Supporting peptides were then exported and file name was changed to protein-peptides-id.csv.

Results:

The mspms R package was designed to be flexible and robust while providing functionality that prioritizes speed and reproducibility of analysis. This is implemented as a package of modular functions that are unified by one of four goals, preprocessing the data, normalizing, statistics/data preparation, and visualization (Figure 1A).

To initially explore our cathepsin MSP-MS dataset, we applied the mspms package to investigate global patterns in the data through principal coordinate and unsupervised hierarchical clustering analyses. The principal component analysis showed that all replicates within each group (as designated by combination of condition and timepoint) clustered tightly together, as demonstrated by the 95% confidence intervals of the eclipses surrounding the points (Figure 1B). Cathepsin L samples were observed to be distinct and separated from the others along the PC1 axis. Notably, this included cathepsin L samples at T0. We also observed that samples treated for 15 minutes with bovine cathepsin B occupied a distinct space along the PC2 axis. These high-level observations were reflected in our unsupervised hierarchical clustering analysis, but the heatmap provided additional information allowing us to more clearly visualize differences in peptide abundances between groups. Replicates belonging to the same condition and time clustered together almost perfectly, and distinct groups of peptides were observed to be more abundant by condition and time (Figure 1C). Notably, T0 samples also tended to cluster together by the cathepsin present.

We then further investigated the high-level differences between bovine and human cathepsin B that were apparent in the PCA and unsupervised hierarchical clustering analysis by looking at the overlap of differentially abundant peptides relative to time 0 between the bovine and human cathepsin B. We observed that 162 (60%) of the significant cleavages were exclusive to bovine cathepsin B, 61 (23%) were shared between bovine cathepsin B and human cathepsin B at all time points, while virtually no significant cleavages were exclusive to human cathepsin B, either after 15 minutes (0, 0%), or after 60 minutes (3, 1%) (Supplementary Figure 1). The number of significant peptides shared uniquely between human and bovine cathepsin B increased over time, with 1 (0%) after 15 minutes, and 41 (15%) after 60 minutes.

We then analyzed the significant differences for each cathepsin relative to time 0 by performing t-tests, calculating the log2 fold change, and plotting the resulting data as volcano plots (Figure 2A). The number of significantly different peptides (as defined as peptides having a log2 fold change >3 and p.adj <= 0.05 relative to time 0) increased for each cathepsin as incubation time increased (Supplementary Figure 2).

Next, we examined the amino acid positions of the original peptide library these significant cleavages occurred at in order to determine the positional substrate specificity of each cathepsin. We observed a clear enrichment of cleavage events at the C terminus of the peptide library with cathepsin A, enriched cleavage events at position 12 and 10 with cathepsin B, peaks of enriched cleavage events at positions 12, 10, 8, 6, 4, and 2 with bovine cathepsin B, peaks of enriched cleavage events at position 2 and 4 with cathepsin C, enriched cleavage events throughout the interior positions, peaking at position 8, with bovine cathepsin D, and enriched cleavage events throughout the interior positions (2 – 11) with cathepsin L. (Figure 2B)

In order to visualize amino acid substrate preferences at each position, we used mspms to conduct an Icelogo analysis, generating substrate specificity plots for positions (P4’-P1’ and P1 – P4). Cleavage sequences from peptides determined to have a log2 fold change >= 3 and FDR adjusted p value of <= 0.05 relative to time 0 were compared to the background universe of all possible cleavages present in the peptide library for each cathepsin where residues with a p val <= 0.05 are reported.

Cathepsin A displayed a strong preference for X at P2’- P4’, and a strong aversion for X at P- P4-P3 (Figure 2C). Hydrophobic residues phenylalanine and norleucine were observed to be enriched immediately proximal to the cleavage site (P1 and P1’), while slight aversion to aspartic acid was observed at P2, P1, and P1’.

Human and bovine cathepsin B showed a strong preference for X at P3’- P4’ as well as enrichment of arginine, lysine, and norleucine at P1, phenylalanine and norleucine at P1’, Human cathepsin B showed a strong aversion for X at P4-P3 (Figure 2C).

Cathepsin C showed a strong preference of X at P4 and P3, and an aversion for X at P3’- P4’. Histidine was preferred at P2 and P1, while phenylalanine was preferred at P1’ and P2’ (Figure 2C).

Bovine Cathepsin D showed an aversion for X at P4-P2 and P2’-P4’, enrichments for phenylalanine and norleucine at P1 and P1’, as well as arginine at P2, P2’, and P3’ (Figure 2C).

Lastly, Cathepsin L showed an aversion to X at P4-P2 and P3’-P4’, as well as a preference for phenylalanine and leucine at P2 (Figure 2C).

Discussion:

Prior to the development of the mspms package, analysis for the multiplex substrate profiling by mass spectrometry for proteases method was performed using a series of in house developed R scripts. This approach was functional but was limited in three primary ways. (1) These scripts consisted of a monolithic code base that was decentralized and poorly documented. (2) These scripts were not entirely self-contained and relied on preforming key components of analysis by manually manipulating intermediate spreadsheets and then inputting into additional tools as needed- making fully traceable and reproducible analyses extremely hard, if not impossible to perform. (3) Adjusting the script to extend to any experiment deviating from a standard protocol was a daunting task that required a non-trivial amount of R programming experience.

In creating the mspms R package, we have addressed these limitations. (1) This tool is accessed through a centralized location, and functionality is broken down into discrete, modular functions. This logical approach makes package maintenance/testing and the development of new features easy to perform. (2) Analysis performed by the mspms package is entirely self-contained, facilitating reproducible research while allowing us to develop mspms-shiny, a GUI application that allows for the user to access the core features of the package without requiring any R programming knowledge. To enable this, key features of Icelogo, a highly cited Java based application that is no longer actively supported, were implemented in R. Our implementation of Icelogo is even more adaptable to mspms experiments, as it can handle nonstandard amino acid letters, X which represent a post terminal site, and n representing Norleucine. (3) Mspms was developed with applicability in mind and supports a broad range of experimental designs. For example, Mspms allows for the use of a user defined peptide library, opening up the possibility to easily extend MSP-MS analysis beyond the previously described 228-peptide library in the future if desired. Additionally, mspms allows the user to analyze the cleavage sequence motifs any arbitrary number of amino acids from the cleavage site, while previous analysis was limited to four amino acids on each side of the cleavage site. Lastly, mspms supports analysis of data exported by multiple upstream proteomics software. PEAKS and Proteome Discoverer are currently supported, and there is the option to easily add more based on user demand in the future. As there is substantial variability in the proteomic software of choice lab to lab, this is a crucial feature enabling broad applicability of the tool.

In this paper, we demonstrate the core features of mspms as applied to publicly available MSP-MS data profiling several well-studied cathepsin proteases spanning all three families of cathepsins: serine (A), aspartic (D), and cysteine (B,C,L). Here we use the mspms package to profile the substrate specificity of cathepsins A, B, D, C, and L, and subsequently compare our results with those reported in the public literature

In our experience, one of the first analyses that should be performed on high-dimensional MSP-MS data is a simple quality control check to make sure that the experimental groups are behaving as expected. In the mspms package, two different avenues of this are supported, principle ordination analysis23 or a heatmap displaying unsupervised hierarchical clustering results by using the plot\_pca() or plot\_heatmap() functions respectively. This allows the user to inspect the broad level features of the data and ensure that the experiment performed as expected. A proper MSP-MS experiment will have been designed with appropriate positive and negative controls, allowing the user to definitively know what was expected and how the observed results agree. In these experiments, time 0 can be thought of as a negative control that should be the same between conditions. All of the cathepsins used in this experiment could be thought of as positive controls to some extent, as their substrate specificities are well-documented.

Initial high-level analysis of our experimental data showed that the cathepsin L exposed samples were very different from the others on a global scale. Importantly, this was apparent at time 0, suggesting that these differences were due either to batch-to-batch variability in the peptide library used from experiment to experiment or batch effects of the mass spectrometry runs. This variability in the initial state of the peptide library is further observed by the results of the hierarchical clustering analysis. If there was no difference at time 0 as would be expected, time 0 samples would cluster randomly. Instead, we observed that replicates of each cathepsin tended to cluster together. Together, this highlights the importance of using differences relative to time 0 with MSP-MS data or using experimental methods to limit this variability (such as pooling aliquots of peptide library for all experimental trials to be compared, or performing mass spectrometry runs in random order.)

We then applied the mspms R package to investigate whether we could recapitulate the expected substrate specificity for each of the cathepsins profiled through the use of cleavage position and Icelogo plots. First, we investigated cathepsin A, a serine protease that processes carboxypeptidase activity6. Our mspms analysis showed that Cathepsin A has carboxypeptidase activity, as indicated by the count of significant peptide cleavages per position of the peptide library. This is also apparent through the enrichment of X at P2’, P3’, and P4’ of the ice logo as well as the negative enrichment of X at the P4, P3, and P2 positions.

Next, we profiled cathepsin B, which is well established to have dipeptidyl carboxypeptidase activity24,7. In our study, we profiled both bovine cathepsin B and human cathepsin B. These proteases share 83.58% identity with each other but any differences in substrate specificities are poorly explored. Our analysis showed clear evidence of dipeptidyl carboxypeptidase activity for both human and bovine cathepsin B as indicated by peaks every 2 AAs with the greatest being at position 12 ( 2 from the C terminal end). This cleavage pattern suggests cleavages occur 2 peptides from the C-terminal end, and then the resulting cleaved peptides are subsequently cleaved 2 peptides from their new C-terminal end. Notably, we observed evidence of more sequential cleavages with bovine cathepsin B compared to human cathepsin B. However, more total cleavages were identified from bovine cathepsin B and we therefore hypothesize the effective concentration or bovine cathepsin B was higher or the bovine protease is more active than the human, which would also explain the additional sequential cleavages observed. Evidence for dipeptidyl carboxypeptidase activity was also apparent by the strong preference for X at positions P4’ and P3’ in the Icelogo plot. Additionally, we noted a trypsin like specificity for both forms of cathepsin B, with arginine and lysine being enriched at P1 and proline being negatively enriched at P1’. Despite these very similar substrate preferences, our principal component and unsupervised hierarchical clustering analyses showed that bovine cathepsin B produced a significantly different profile from human cathespin B on a global scale. Further investigation showed that virtually all of the peptides found to be significant after incubation with human cathepsin B were shared with those found to be significant using the bovine preparation, but a large number of peptides were observed in bovine but not human cathepsin B. Altogether, these results suggest that bovine cathepsin B and human cathepsin B have similar substrate specificities, and the high-level differences observed in the MSP-MS experiment are likely related to the higher activity of the bovine cathepsin B preparation used in the experiment relative to its human counterpart, either due to a higher concentration, or greater activity of the preparation in the experimental buffer.

Next, we assessed cathepsin C, also known as dipeptidyl peptidase I (DPP-I), which is a lysosomal dipeptidyl-peptidase belonging to the cystine cathepsin protease family. It is known to have relatively broad substrate specificity and acts by the stepwise removal of dipeptides from the N-terminus of the cleavage site. It has been reported to stop peptidase activity in three cases: when the N-terminal amino group is blocked, when a proline residue is adjacent to the cleavage site (P1 or P1’ position), or when the N-terminal residues are lysine or arginine8. Our mspms based analysis showed clear indications of dipeptidyl aminopeptidase activity, as the most significant cleavages were found to occur at position 2 of the library. We also observed a peak in significant cleavages at position 4 of the peptide library that increased over time. This suggests sequential cleavage of products 2 positions from the N terminal end. We see these aspects reflected in the icelogo by the strong preference for X at P4 and P3, as well as aversion for X at P3’ and P4’.

We were not able to detect evidence supporting the two reported remaining rules via a standard Icelogo based analysis, as we saw no aversion to proline at P1 or P1’, or lysine and arginine at N-terminal residues indicated in the Icelogo plot. However, the flexible nature of the mspms R package permitted further exploration of the data which allowed us to filter the data using a regular expression capturing these described rules of cathepsin C ("XX[^L|K|X][^P][^P]...") to see if we observe their cleavage over time. We saw that peptides with cleavage sequences fitting this rule were significantly increased over time (Supplementary Figure 3). For comparison, we reversed the rule ("...[^P][^P][^L|K|X]XX") and saw no significance over time. Together, this suggests that the previously described rules for cathepsin C specificity are accurate, but additional features defining optimal substrate specificity may also exist.

We then moved on to profile cathepsin D, a lysosomal aspartyl protease that has endoprotease activity. It is known to prefer hydrophobic residues on both sides of the cleavage sensitive bond (P1 and P1’)10,9. Mspms based analysis showed clear indications that bovine cathepsin D has endopeptidase activity, both by the position plot, and Icelogo. The Icelogo further displays a clear preference for Phenylalanine (F) and neuroleucine (n), which are both hydrophobic residues, recapitulating the expected results near perfectly.

Lastly, we profiled cathepsin L, a lysosomal cysteine protease that primarily has endopeptidase activity and preferentially cleaves peptide bonds with aromatic residues in P2 and hydrophobic residues in P3 position 11. Our mspms based analysis recapitulated these expected results, showing clear indications that cathepsin L possesses endopeptidase activity, and that the hydrophobic residues phenylalanine and tryptophan were enriched at P2 as well as the hydrophobic residue leucine was observed to be enriched at P3.

Together, these results show that the mspms R package is able to analyze MSP-MS data recapitulating expected substrate specificities. While we analyzed 6 well-studied cathepsins in the current work, this type of analysis can be extended to any set of samples profiled by the MSP-MS method, whether that be complex biomixtures or purified proteases of interest. Due to the involvement of proteases in a wide variety of biological processes and the importance of determining their substrate specificities we expect mspms to be a widely used tool facilitating the analysis of MSP-MS data across many research groups.

Conclusions

Here we present the mspms R package and show that it is a robust, flexible, and simple-to-use tool allowing researchers to analyze data produced by the MSP-MS method to probe protease specificity in their sample(s) of choice.

Availability of Source Code and Requirements:

* Project name: mspms
* Project home page: <https://github.com/baynec2/mspms>.
* Operating system(s): Platform independent
* Programming language: R
* Other requirements: R 4.1.0, dplyr, magrittr, MASS, NormalyzerDE, outliers, purrr, stats, tidyr, truncnorm, utils, ggplot2, ggseqlogo, heatmaply, readr, readxl, rstatix, stringr, tibble, SummarizedExperiment, rlang, forcats, ggpubr, rmarkdown, DT.
* License: MIT
* Bioconductor repository: Submitted, TBD.
* Shiny app instance: <https://gonzalezlab.shinyapps.io/mspms_shiny/>
* Shiny app repository: <https://github.com/baynec2/mspms-shiny>
* Reproducible vignette: <https://github.com/baynec2/mspms>/vignettes/mspms\_analysis.RMD
* Manuscript repo: <https://github.com/baynec2/mspms-manuscript>.

Availability of Supporting Data and Materials:

All data used to build this manuscript can be found in the github repository for the manuscript. All supporting data and materials are additionally available in the GigaScience and GigaDB database.

Supplementary Data:

**Supplemental Fig. S1.** Venn diagram showing the number of significant peptides shared between human and bovine cathepsin B relative to time 0.

**Supplemental Fig. S2.** Scatterplot showing the number of peptides significantly different relative to time 0 over time for each cathepsin.

**Supplemental Fig.S3.** Violin plot investigating changes over time of peptides with cleavages containing sequences consistent with previously described cathepsin C specificity rules (A) compared to reversed specificity rules (B).

**Supplemental Fig.S4.** Screenshots of mspms-Shiny graphic user interface.

Abbreviations:

MSP-MS: Multiplex Substrate Profiling by Mass Spectrometry. T0: Time zero. PCA: Principal Component Analysis. PC1: Principal component 1. PC2: Principal component 2.

Competing Interests:

The authors declare that they have no competing interests.

Funding:

Charlie Bayne and Brianna Hurysz were supported in part by the UCSD Graduate Training Program in Cellular and Molecular Pharmacology through an institutional training grant from the National Institute of General Medical Sciences, T32 GM007752.

Author contributions

C.B wrote the R package, shiny app, documentation, and manuscript; B.H contributed to software development. D.J.G. and A.O oversaw the project.

Acknowledgements

We would like to acknowledge Dr. Zhenze Jiang for his initial conception of the previous R scripts, as well as Dr. Lawrence Liu and Dr. Michael Yoon for maintaining and updating the scripts. We also thank Dr. Jiang and Dr. Yoon for running the cathepsins MSP-MS experiment analyzed in this manuscript. Finally, we would like to acknowledge Diego F. Trujillo for suggestions to improve mspms and for educating new users how to perform the data analysis.

Thank the people who did the physical experiments. Whoever wrote the initial R script. Anyone else?

References

1. O’Donoghue, A. J. *et al.* Global identification of peptidase specificity by multiplex substrate profiling. *Nat Methods* **9**, 1095–1100 (2012).

2. Rohweder, P. J., Jiang, Z., Hurysz, B. M., O’Donoghue, A. J. & Craik, C. S. Multiplex substrate profiling by mass spectrometry for proteases. in *Methods in Enzymology* vol. 682 375–411 (Elsevier, 2023).

3. Miller, G. A Scientist’s Nightmare: Software Problem Leads to Five Retractions. *Science* **314**, 1856–1857 (2006).

4. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* (2004).

5. Yadati, T., Houben, T., Bitorina, A. & Shiri-Sverdlov, R. The Ins and Outs of Cathepsins: Physiological Function and Role in Disease Management. *Cells* **9**, 1679 (2020).

6. Hiraiwa, M. Cathepsin A/protective protein: an unusual lysosomal multifunctional protein. *Cellular and Molecular Life Sciences (CMLS)* **56**, 894–907 (1999).

7. Yoon, M. C., Hook, V. & O’Donoghue, A. J. Cathepsin B Dipeptidyl Carboxypeptidase and Endopeptidase Activities Demonstrated across a Broad pH Range. *Biochemistry* **61**, 1904–1914 (2022).

8. Stojkovska Docevska, M. & Novinec, M. Cathepsin C: structure, function, and pharmacological targeting. *Rare Dis Orphan Drugs J* **2**, (2023).

9. Ferguson, J. B., Andrews, J. R., Voynick, I. M. & Fruton, J. S. The Specificity of Cathepsin D. *Journal of Biological Chemistry* **248**, 6701–6708 (1973).

10. Sun, H. *et al.* Proteolytic Characteristics of Cathepsin D Related to the Recognition and Cleavage of Its Target Proteins. *PLoS ONE* **8**, e65733 (2013).

11. Brömme, D., Bonneau, P. R., Lachance, P. & Storer, A. C. Engineering the S2 subsite specificity of human cathepsin S to a cathepsin L- and cathepsin B-like specificity. *Journal of Biological Chemistry* **269**, 30238–30242 (1994).

12. Ma, B. *et al.* PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Comm Mass Spectrometry* **17**, 2337–2342 (2003).

13. Orsburn, B. C. Proteome Discoverer—A Community Enhanced Data Processing Suite for Protein Informatics. *Proteomes* **9**, 15 (2021).

14. Willforss, J., Chawade, A. & Levander, F. NormalyzerDE: Online Tool for Improved Normalization of Omics Expression Data and High-Sensitivity Differential Expression Analysis. *J. Proteome Res.* **18**, 732–740 (2019).

15. Dixon, W. J. Processing Data for Outliers. *Biometrics* **9**, 74 (1953).

16. Ripley, B. *et al.* MASS: Support Functions and Datasets for Venables and Ripley’s MASS. (2024).

17. Mersmann, O., Trautmann, H., Steuer, D. & Bornkamp, B. truncnorm: Truncated Normal Distribution. (2023).

18. Kassambara, A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. (2023).

19. Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J. & Gevaert, K. Improved visualization of protein consensus sequences by iceLogo. *Nat Methods* **6**, 786–787 (2009).

20. Galili, T., O’Callaghan, A., Sidi, J. & Sievert, C. heatmaply: an R package for creating interactive cluster heatmaps for online publishing. *Bioinformatics* **34**, 1600–1602 (2018).

21. Wickham, H. *et al.* ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. (2024).

22. Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645–3647 (2017).

23. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Phil. Trans. R. Soc. A.* **374**, 20150202 (2016).

24. Yoon, M. C. *et al.* Selective Neutral pH Inhibitor of Cathepsin B Designed Based on Cleavage Preferences at Cytosolic and Lysosomal pH Conditions. *ACS Chem. Biol.* **16**, 1628–1643 (2021).

Figure Legends

**Figure 1. Overview of the mspms R package and cathepsin MSP-MS profiling.** (A) Schematic of the functions contained within the mspms R package. (B) Principal component analysis displaying PC1 and PC2. Samples are colored by time, and the shape and line type displays the type of cathepsin. Eclipses represent the 95% confidence interval. (C) Heatmap showing the results of the experiment as clustered using unsupervised hierchical clustering. Rows of the heatmap represent the samples while columns represent the peptides. Color of the heatmap cells represent the normalized, column centered, and scaled values. White cells represent those that were found to be outliers and converted to NA. Colored bars to the right of the heatmap indicate the cathepsin and time of the samples in each row.

**Figure 2. Differentially abundant peptide cleavages over time.** (A) Volcano plots displaying the log2 fold change of the timepoint as indicated by color relative to and -log10 FDR corrected p values for each cathepsin. (B) Plot showing the number of significant cleavage events at each position of the peptide library (as defined as having a log2fc >= 3 and FDR adjusted p values <= 0.05) (C) Icelogo plots as implemented in the mspms package. Amino acid residues (with X representing positions past the terminus) four positions to the left and right of the cleavage site are displayed. Only residues with significantly higher proportions relative to the proportion of all possible cleavage sequences present in the initial peptide library (pval <= 0.05) are shown, with the height representing the percentage differences compared to background.