Data analysis performed in the paper:

The PepSAVI-MS pipeline for natural product bioactive peptide discovery

October 22, 2024

Contents

0 Introduction

The PepSAVIms R package provides a collection of software tools used to facilitate the prioritization of putative bioactive compounds from a complex biological matrix. The package was constructed to provide an implementation of the statistical portion of the laboratory and statistical procedure proposed in *The PepSAVI-MS pipeline for natural product bioactive peptide discovery* (herafter abbreviated to *The PepSAVI-MS pipeline*) [**?**].

This document describes in detail all of the data processing and analysis steps performed in *The PepSAVI-MS pipeline*. By providing this analysis, we hope to facilitate a deeper understanding of the methodology and results described in the paper, and provide a working template for researchers who may wish to apply this methodology to their own research.

1 Data collection

This section describes in brief the data collection procedures that are performed upstream of the data analysis. Please refer to *The PepSAVI-MS pipeline* for a detailed explanation.

1.1 *Viola odorata* sample preparation and LC-MS/MS analysis

V. odorata aerial tissue was aqueously extracted using size exclusion steps to selectively target AMP-like molecules. The resulting extract was crudely fractionated using strong cation exchange (SCX) chromatography for creation of the *V. odorata* peptide library. Bioactivity assays against a panel of microbial pathogens were performed using the generated peptide library. Each SCX fraction was then subject to nano-LC-ESI-MS/MS (Waters nanoAcquity UPLC coupled to an AB Sciex TripleTOF 5600) analysis. After the processing of this dataset, accurate intact mass and relative intensity information for peptide constituents contained within each fraction was obtained, resulting in 30,799 MS features for *Viola odorata* across the SCX fractions.

1.2 Bioactivity screening

Peptide libraries were assayed for growth inhibition against the following pathogens: *E. coli* (ec), *E. faecium* (ef) *S. aureus* (sa), *K. pneumoniae* (kp), *A. baumannii* (ab), *P. aeruginosa* (pa), *E. cloacae* (ecl), *F. graminearum* (fg), and cancer cell lines: breast cancer (bc), prostate cancer (pc) and ovarian cancer (oc). Library fractions were incubated with a microbial or cancer cell culture in a manner such that the presence of bioactive peptides in a given fraction will result in inhibition of culture growth during the incubation period. For bacterial assays, the remaining viable cells were quantified indirectly by spectrophotometric measurement of the irreversible intracellular bioreduction of resazurin. For anticancer bioactivity, cytotoxicity assays were performed using MTT-based assays to measure mitochondrial succinate dehydrogenase activity with absorbance measurements at 570 nm. Values for each fraction are compared to positive and negative controls containing a known therapeutic or water, respectively, to determine a percent activity of each library fraction, where a small value of remaining viable cells indicates high activity. Percent activity of each well was calculated using the formula:

$$\text{percent activity} = \left(1 - \frac{\text{Response of fraction} - \text{Response of positive control}}{\text{Response of negative control} - \text{Response of positive control}}\right) \times 100$$

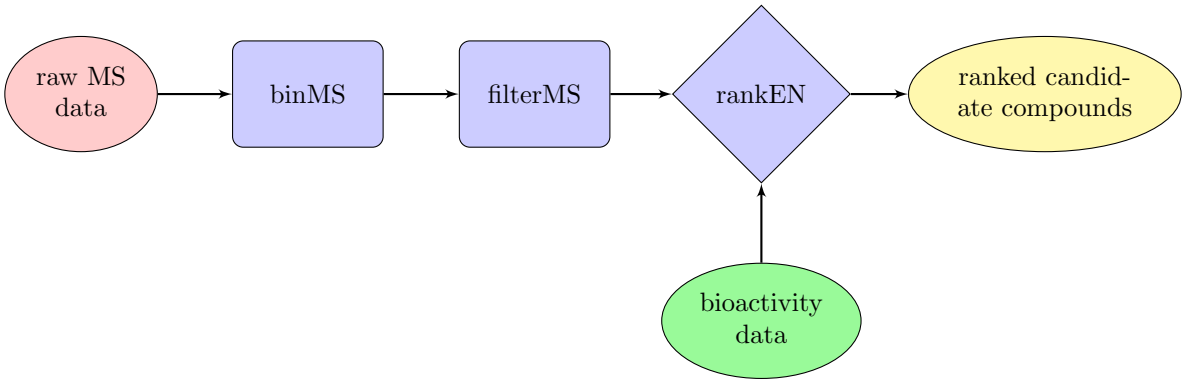
where response refers to relative fluorescence units for antibacterial assays and absorbance for anticancer assays.

2 Data analysis presented in *The PepSAVI-MS pipeline*

This section describes all of the steps and shows the results of the data analyses performed in *The PepSAVI-MS pipeline*, as depicted in Figure 1. In this scenario, mass spectrometry abundance values for each compound across the SCX fractions are obtained using LC-MS/MS analysis as described in section 1.1, and is represented by the pink oval. The mass spectrometry data is consolidated via the `binMS` function, and the resulting consolidated data is filtered by the `filterMS` function.

Bioactivity data is collected as described in section 1.2, and is represented by the green oval. The bioactivity data is then then provided along with the consolidated and filtered data as inputs to the `rankEN` function, which calculates and returns the data analysis results.

Figure 1: Data analysis flow chart for the data analysis performed in *The PepSAVI-MS pipeline*



2.1 Loading the mass spectrometry and bioactivity data

The first step in the data analysis is to load the mass spectrometry data collected for *The PepSAVI-MS pipeline*. The raw data produced by this method is stored as a comma-seperated values (csv) file. However, in order to minimize the size of the PepSAVIms package, we have removed the columns in the data corresponding to variables not used in the data analysis, and have converted the data to a compressed native R format. The values of the data relevent to this data analysis remain unchanged from their original form.

Readers who wish to obtain the data in its raw form may obtain the data as well as the R script used to convert the raw data into the form provided with the package from the PepSAVIms package developement repository located at `https://github.com/dpritchLibre/PepSAVIms` in the `data-raw` directory.

2.1.1 Mass spectrometry data

The mass spectrometry data yielded intact mass and relative intensity information for peptide constituents contained within each library fraction; the analysis obtained 30,799 MS features for *Viola odorata* across fractions 1-43.

```
library(PepSAVIms)

# Load mass spectrometry data into memory
data(mass_spec)
```

2.1.2 Bioactivity data

Peptide libraries were assayed for growth inhibition against the pathogens listed in Section 1.2. Three replicate observations were collected for each of the assays with the exception of *F. graminearum*, which had two replicates. The resulting percent activity of each fraction was quantified using cell viability assays, and the average response across all replicates was used for this analysis. Assays showing promising activity in the region of cyO2 elution include *E. coli* (ec), *A. baumannii* (ab), *P. aeruginosa* (pa), *F. graminearum* (fg), breast cancer (bc), prostate cancer (pc) and ovarian cancer (oc).

```
# Load bioactivity data into memory
data(bioact)
```

2.2 Consolidating mass spectrometry data

Now that the mass spectrometry data is loaded into memory, we begin data processing using PepSAVIms pipeline by consolidating MS features in the data believed to belong to the same compound. In this way, all of the abundance belonging to the same compound will be grouped together and thus resemble a more accurate depiction of the compound abundance accross SCX fractions. This is performed via the `binMS` function.

2.2.1 Specification of criteria

The `binMS` function contains a number of criteria to exclude unwanted compounds and thereby restrict compounds to those of potential interest. Retention time limitations are put in place to eliminate background ions that are detected either before (equilibration) or after (wash) the gradient is applied; for the gradient performed in our laboratory, 14 and 45 minutes are appropriate retention time lower and upper bounds. Mass and charge limitations of 2 to 15 kDa and +2 to +10, respectively, were chosen to restrict compounds to the mass and charge ranges of known bioactive peptides.

MS features that satisfy this initial exclusion process are then consolidated when features are believed to belong to the same underlying compound. This consolidation step requires choices to be made through the specification of criteria deeming two MS features to be considered the same compound. An m/z difference of 0.05 Da was chosen for our data analysis to be the maximum difference for which two compounds could have in m/z to be considered to belong to the same compound. The value of 0.05 Da was chosen by performing multiple LC-MS runs studying

the variation in the observations; this value was chosen so that fluctuations in mass accuracy and retention time for the test data allowed the same peak to be picked multiple times despite representing the same compound.

The other criterion that must be met for two MS feature observations to be considered to belong to the same compound is that the retention times for the two features cannot be more than a specified amount of time apart. For this analysis, because we used prior retention time alignment, we chose to allow all MS features that were below the threshold for m/z difference and had the same charge state to be considered to belong to the same compound. The option to effectively ignore this criterion is implemented by specifying the appropriate parameter to be equal to the run time, which in this case was 60 minutes (likewise, any value greater than 60 would have the same effect).

2.2.2 Consolidating data with `binMS`

Here we perform the mass spectrometry consolidation prodedure, using the criteria as described in Section 2.2.1. We can see from the summary output below that in totality, the number of of candidate compounds has been reduced from an initial amount of 30,799 observations to 6,258 compounds.

The reduction of candidate compounds can be considered more granularly as occuring in two separate steps. The first is an exclusion process, and we can see that the number of elution levels that satisfied the criterion for time of peak retention, mass, and charge level were 26,387, 11,482, and 19,250 levels, respectively. The number of elution levels that satisfied all of the inclusion criteria was 10,902. Then the consolidation step consolidated those 10,902 elution levels into 6,258 candidate compounds.

```
# Perform mass spectrometry levels consolidation
bin_out <- binMS(mass_spec = mass_spec,
                 mtoz = "m/z",
                 charge = "Charge",
                 mass = "Mass",
                 time_peak_reten = "Reten",
                 ms_inten = NULL,
                 time_range = c(14, 45),
                 mass_range = c(2000, 15000),
                 charge_range = c(2, 10),
                 mtoz_diff = 0.05,
                 time_diff = 60)

# Show some summary information describing the consolidation process
summary(bin_out)

##
## The inclusion criteria was specified as follows:
## -------------------------------------------------
##     time of peak retention:  between    14 and     45
##     mass:                    between 2,000 and 15,000
##     charge:                  between     2 and     10
##
## m/z levels were consolidated when each of the following criteria were met:
## ------------------------------------------------------------------------
##     m/z levels were no more than 0.05 units apart
##     the time peak retention occured no farther apart than 60 units
##     the charge states were the same
##
## The mass spectrometry data prior to binning had:
## -------------------------------------------------
##     30,799 m/z levels
##
## The number of remaining m/z levels after filtering by the inclusion criteria was:
## ------------------------------------------------------------------------------
##     time of peak retention:  26,387
##     mass:                    11,482
```

```
##     charge:                  19,250
##     satisfied all:           10,902
##
## After consolidating the m/z levels, there were:
## ------------------------------------------------
##     6,258 levels
```

2.3 Filtering mass spectrometry data

The next step in the PepSAVIms pipeline is to remove any potential candidate compounds with observed abundances for which it is unlikely that they might be a compound with an effect on the observed bioactivity. This step is performed by the `filterMS` function.

2.3.1 Selecting fraction region of interest

Bioactivity regions of interest are selected based on each individual bioactivity data set. Because of the crude nature of SCX, a given peptide should elute over a minimum of three fractions in a Gaussian manner. Because mass spectrometry is more sensitive than the bioactivity assays, the defined bioactivity region can extend 1-2 fractions beyond the visible activity range on either end. The bioactivity regions for each species that were deemed to be active are depicted by blue bars in Figure 2. The regions chosen for each bioactivity data ranged from a size of as small as 3 fractions to as large as 6 fractions, and each region was contained within the window of fractions 17-25.

For typical analyses we recommend filtering the MS dataset using the region chosen as dictated by the bioactivity data (as described above). However, for the analysis performed for *The PepSAVI-MS pipeline*, for simplicity of presentation and since a single region rather tightly encapsulates the chosen region for all of the bioactivity datasets, we simply filtered the MS dataset with the region chosen to be the smallest one that included the chosen region for every individual datset, namely fractions 17-25.

2.3.2 Filtering criteria selection

In addition to selecting the region of interest for which to filter by, we must also specify a bordering region; this is the region that we consider when filtering for candidate compounds by criterion 3: a candidate compound's elution in the bordering region cannot be greater than a specified proportion of its maximum abundance, and criterion 4: a candidate compound must have a nonzero level of elution in the right adjacent fraction to the fraction with the maximum elution. For the data analysis performed for *The PepSAVI-MS pipeline*, we specified the most conservative choice for bordering region, which is that every fraction not in the region of interest is to be considered the bordering region.

Another criterion that we filter by is that a candidate compound may not by more abundant in the bordering region than some specified proportion of its maximum abundance in the region of interest. Allowing for a small nonzero elution outside of the region of interest accounts for small fluctations from absolute 0 that are most likely due to noise in the instrumentation readings. The allowable proportion of abundance relative to the maximum is data dependent; we selected a value of 0.01 by inspecting the distribution of MS features in the data that fit the profile of what we might expect for a compound with an effect on bioactivity levels in the region.

Another filtering criterion that we consider for a candidate compound is a minimum value for the maximum intensity value over the LC-MS profile. This level should be chosen to reflect the appropriate amount of noise apparent in the instrumentation; for our analysis we selected 1000 Da as an appropriate value.

Finally, an option to filter the data by a maximum allowed charge state is included in the pipeline. This has already been considered in the consolidation step, but is included in the filtering step in the event that researcher does not wish to use the `binMS` consolidation function with the data analysis in hand. We selected a maximum charge state of +10, consistent with our choice in Section 2.2.1.

2.3.3 Filtering data with `filterMS`

Here we perform the filtering operation to remove any potential candidate compounds with observed abundances for which it is scientifically unlikely that they might be a compound with an effect on bioactivity via the `filterMS`

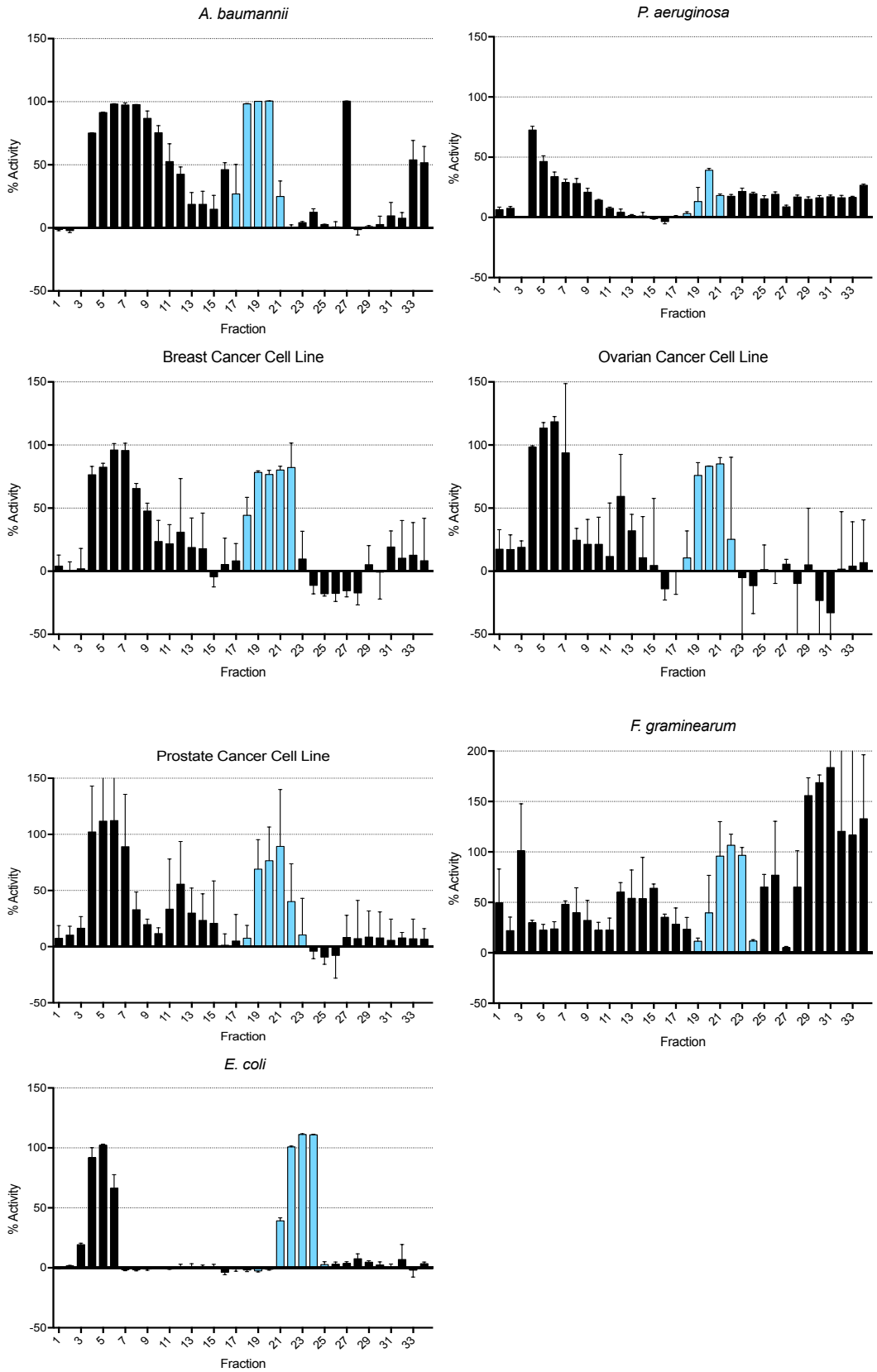


Figure 2: *Viola odorata* peptide library bioactivity against all pathogens in which strong activity was seen in the region of cyO2 elution. Average percentage of activity values for each fraction are plotted with error bars representing +1 standard deviation. The fractions selected as the region of interest are shown using blue bars. *(note that this figure was not created using PepSAVIms or R)*

function. We can see from the summary output below that 3,428 compounds satisfied criterion 1, 3,428 compounds satisfied criterion 2, 3,428 compounds satisfied criterion 3, 3,428 compounds satisfied criterion 4, and 3,428 compounds satisfied criterion 5 (see *The PepSAVI-MS pipeline* or the `binMS` documentation for a description of the criteria). In total, 225 compounds satisfied each each of the 5 criteria.

```
# Perform mass spectrometry levels filtering
filter_out <- filterMS(msObj = bin_out,
                       region = paste0("VO_", 17:25),
                       border = "all",
                       bord_ratio = 0.01,
                       min_inten = 1000,
                       max_chg = 10)

# Show summary information describing the filtering process
summary(filter_out)

##
## The region of interest was specified as (9 fractions):
## ------------------------------------------------------
##     20150207_CLK_BAP_VO_17
##     20150207_CLK_BAP_VO_18
##     20150207_CLK_BAP_VO_19
##     20150207_CLK_BAP_VO_20
##     20150207_CLK_BAP_VO_21
##     20150207_CLK_BAP_VO_22
##     20150207_CLK_BAP_VO_23
##     20150207_CLK_BAP_VO_24
##     20150207_CLK_BAP_VO_25
##
## The bordering regions were specified as "all"
## ---------------------------------------------
##     * fraction names omitted for brevity *
##
## The filtering criteria was specified as:
## ----------------------------------------
##     minimum intensity:      1,000
##     maximum charge:            10
##     bordering region ratio:     0.01
##
## The mass spectrometry data prior to filtering had:
## --------------------------------------------------
##     6,258 compounds
##        34 fractions
##
## Individually, each criterion reduced the 6,258 m/z levels to the following number:
## ----------------------------------------------------------------------------------
##     criterion 1:  3,428    (fraction with max. abundance is in region of interest)
##     criterion 2:  1,080    (fractions in bordering region have < 1% of max. abundance)
##     criterion 3:  2,818    (nonzero abundance in right adjacent fraction to max.)
##     criterion 4:  4,012    (at least 1 intensity > 1,000 in region of interest)
##     criterion 5:  6,258    (must have charge <= 10)
##
## The total number of candidate compounds was reduced to:
## -------------------------------------------------------
##     225
```

2.4 Candidate compound ranking

In this section potential candidate compounds are ranked with the goal of facilitating the investigation of compounds with a deleterious effect on each of the bioactivity peptide libraries via the `rankEN` function. For each library the mass spectrometry data remains the same. Also note that the region of interest changes over the peptitde libraries according to the selections shown in Figure 2.

2.4.1 Selecting the quadratic penalty parameter

The `rankEN` function works by selecting a choice of quadratic penalty parameter, and for fixed value of quadratic penalty parameter tracking the order in which the coefficients corresponding to candidate compounds first become nonzero along the elastic net [2] path as the ℓ_1 penalty parameter changes. Then it is presumed that compounds with corresponding coefficients that become nonzero earlier in the path may be better candidates for having an effect on bioactivity levels then those with corresponding coefficients that become nonzero later in the path. To select a quadratic penalty parameter we recommend a small nonzero value with the thought that being near to the lasso penalty [3] we might achieve good behavior in terms of variable selection. The reason for not choosing a penalty of 0 is that in that case the elastic net reduces to the lasso model, which can have no more than one less than the number of bioactivity data points of nonzero coefficients, and thus severly reduces the size of the list of candidate compounds generated by this approach. We also note that we do not consider this choice to be data-driven; for our analysis we chose a value of 0.001.

```
# Rank the candidate compounds using the ranking procedure for each of the
# bioactivity datasets

rank_oc <- rankEN(msObj      = filter_out,
                  bioact     = bioact$oc,
                  region_ms  = paste0("VO_", 18:22),
                  region_bio = paste0("VO_", 18:22),
                  lambda     = 0.001)

rank_bc <- rankEN(msObj      = filter_out,
                  bioact     = bioact$bc,
                  region_ms  = paste0("VO_", 18:22),
                  region_bio = paste0("VO_", 18:22),
                  lambda     = 0.001)

rank_pc <- rankEN(msObj      = filter_out,
                  bioact     = bioact$pc,
                  region_ms  = paste0("VO_", 18:23),
                  region_bio = paste0("VO_", 18:23),
                  lambda     = 0.001)

rank_ab <- rankEN(msObj      = filter_out,
                  bioact     = bioact$ab,
                  region_ms  = paste0("VO_", 17:21),
                  region_bio = paste0("VO_", 17:21),
                  lambda     = 0.001)

rank_pa <- rankEN(msObj      = filter_out,
                  bioact     = bioact$pa,
                  region_ms  = paste0("VO_", 18:21),
                  region_bio = paste0("VO_", 18:21),
                  lambda     = 0.001)

rank_ec <- rankEN(msObj      = filter_out,
                  bioact     = bioact$ec,
                  region_ms  = paste0("VO_", 18:25),
```

```
                  region_bio = paste0("VO_", 18:25),
                  lambda     = 0.001)

rank_fg <- rankEN(msObj      = filter_out,
                  bioact     = bioact$fg,
                  region_ms  = paste0("VO_", 19:24),
                  region_bio = paste0("VO_", 19:24),
                  lambda     = 0.001)
```

3 PepSAVIms pipeline validation using cyO2

To validate this pipeline, we demonstrate successful detection and identification of a known AMP from the botanical species *Viola odorata*. *Viola odorata*, commonly known as sweet violet, contains many cyclotides - including cycloviolacin O2 (cyO2). CyO2 is a small, cysteine rich cyclotide comprised of 30 amino acids (MWmonoisotopic: 3138.37 Da), which has been shown to have diverse activity against many Gram-negative bacteria (*E. coli*, *K. pneumoniae*, and *P. aeruginosa*), as well as several cancer cell lines. Here, we demonstrate successful detection, prioritization, and identification of cyO2 as a means of validating the use of this statistical analysis approach for bioactive peptide discovery.

3.1 CyO2 compound ranking results

Each of the R objects created in Section 2.4 contains a ranking of the candidate compounds obtained using the procedure. The ranked compounds' m/z values are provided in the `mtoz` element of the `rankEN` object, while the ranked compounds' corresponding charge values are found in the `charge` element (we can also obtain the m/z and charge values through the `extract_ranked` function).

The exact m/z and charge values for cyO2 were obtained by comparing the candidate compounds list after the consolidation and filtering process to the known range of values; the m/z and charge pairs corresponding to cyO2 were determined to be (1047.4898, 3), and (1570.2414, 2) in our data. Thus the R function `find_cyO2_rank` shown below returns a vector of the rankings of each charge state of cyO2.

From the output below we see that for the *E. coli*, and breast cancer data sets, at least one charge state corresponding to cyO2 is identified in the first 20 candidate compounds. For the *A. baumannii*, *P. aeruginosa*, and prostate cancer data cets, CyO2 is identified in the first 50 compounds; and for the *F. graminearum* and ovarian cancer data sets cyO2 is identified in the first 100 compounds.

```
# Function to find the rank of cyO2 compounds
find_cyO2_rank <- function(rankEN_obj) {
    # The m/z values for the two incarnations of cyO2
    mval1 <- 1047.4897758000001886
    mval2 <- 1570.2413587500000176
    # Find the indices (corresponding to the ranks) of the cyO2 incarnations
          which((rankEN_obj$mtoz == mval1 & rankEN_obj$charge == 3) |
                (rankEN_obj$mtoz == mval2 & rankEN_obj$charge == 2))
}

# List the ranks for cyO2
lapply(list(ab=rank_ab, bc=rank_bc, ec=rank_ec, fg=rank_fg,
            oc=rank_oc, pa=rank_pa, pc=rank_pc),
       find_cyO2_rank)

## $ab
## [1] 23
##
```

```
## $bc
## [1]   3 124
##
## $ec
## [1] 20 97
##
## $fg
## [1]  97 112
##
## $oc
## [1]  99 169
##
## $pa
## [1] 41 79
##
## $pc
## [1]  45 172
```

4 Conclusion

This document describes in detail the data processing and analysis steps as implimented in *The PepSAVI-MS pipeline*, including explanations for the choices that were made when performing the analysis. We hope that in conjunction with the PepSAVIms introduction vignette these vignettes provide a valuable template for laboratories wishing to implement this methodology for their own research purposes. As a result of the analysis, we see that out of 30,799 MS features originally measured by the LC-MS process, this pipeline was able to identify cyO2 as a top-20 likely contributor to bioactivity against two of the seven pathogens, and as one of the first 100 candidate compounds for each of the bioactivity data sets.

References

[1] Kirkpatrick, C.L., Broberg, C.A., McCool, E.N., Lee, W.J., Chao, A., McConnell, E.W., Hebert, M., Fleeman, R., Adams, J. and Jamil, A. (2017). The "PepSAVI-MS" Pipeline for Natural Product Bioactive Peptide Discovery. Analytical chemistry, 89(2), 1194-1201.

[2] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320.

[3] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.