protViz: Visualizing and Analyzing Mass Spectrometry Related Data in Proteomics using R

Christian Panse and Jonas Grossmann

Package Description

protViz is an R package to do quality checks, visualizations, and analysis of mass spectrometry data, coming from proteomics experiments.

The package is developed, tested and used at the Functional Genomics Center Zurich. We use this package mainly for prototyping, teaching, and having fun with proteomics data. But it can also be used to do solid data analysis for small-scale datasets.

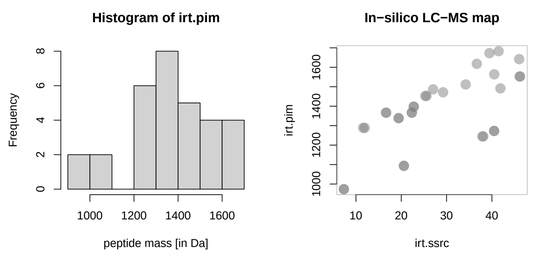
1 Peptide Identification

The currency in proteomics are the peptides. In proteomics, proteins are digested to so-called peptides since peptides are much easier to handle biochemically than proteins. Proteins are very different some are very sticky while others are soluble in aqueous solutions while again are only sitting in membranes. Therefore, proteins are chopped up into peptides because it is fair to assume, that for each protein, there will be some peptides behaving well so that they can be measured with the mass spectrometer. This step introduces another problem, the so-called protein inference problem. In this package here, we do not look at all touch upon the protein inference.

1.1 Computing the Parent Ion Mass

The function parentIonMass determines the mass of a amino acid sequence while the function ssrc returns a hydrophobicity value for a given sequence of amino acids which can be used to predict the retention times [4].

```
R> library(protViz)
R> irt.peptide <- as.character(
+  protViz::iRTpeptides$peptide)
R> irt.pim <- parentIonMass(irt.peptide)
R> op <- par(mfrow = c(1,2),
+  pch = 16,
+  col = rgb(0.5, 0.5, 0.5, alpha = 0.5))
R> hist(irt.pim, xlab="peptide mass [in Da]")
R> irt.ssrc <- sapply(irt.peptide, ssrc)
R> plot(irt.pim ~ irt.ssrc,
+  cex = 2,
+  main = 'In-silico LC-MS map')
R> par(op)
```



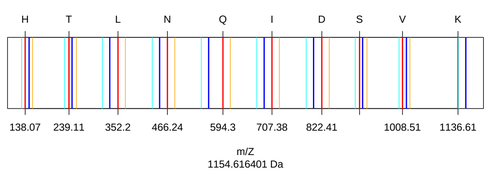
1.2 In-silico Peptide Fragmentation

The fragment ions of a peptide can be computed following the rules proposed in [6]. Beside the b and y ions, the FUN argument of fragmentIon defines which ions are computed. the default ions beeing computed are defined in the function defaultIon. The are no limits for defining other forms of fragment ions for ETD (c and z ions) CID (b and y ions).

```
R> defaultIon

function (b, y)
{
    Hydrogen <- 1.007825
    Oxygen <- 15.994915
    Nitrogen <- 14.003074
    c <- b + (Nitrogen + (3 * Hydrogen))
    z <- y - (Nitrogen + (3 * Hydrogen))
    return(cbind(b, y, c, z))
}
<bytecode: 0x5652514fdfa0>
<environment: namespace:protViz>

R> peptides <- c('HTLNQIDSVK')
R> fi <- fragmentIon(peptides)
```



The next lines compute the singly and doubly charged fragment ions of the HTLNQIDSVK peptide. The in-silico ions are usually the ones that can be used to perform the identification.

```
R> fi.HTLNQIDSVK.1 <-
+  fragmentIon('HTLNQIDSVK')[[1]]
R> Hydrogen <- 1.007825
R> fi.HTLNQIDSVK.2 <-
+  (fi.HTLNQIDSVK.1 + Hydrogen) / 2
```

b1	y1	c1	z1	b2	y2	c2	z2
138.07	147.11	155.09	130.09	69.54	74.06	78.05	65.55
239.11	246.18	256.14	229.15	120.06	123.59	128.57	115.08
352.20	333.21	369.22	316.19	176.60	167.11	185.12	158.60
466.24	448.24	483.27	431.21	233.62	224.62	242.14	216.11
594.30	561.32	611.33	544.30	297.65	281.17	306.17	272.65
707.38	689.38	724.41	672.36	354.20	345.20	362.71	336.68
822.41	803.43	839.44	786.40	411.71	402.22	420.22	393.70
909.44	916.51	926.47	899.48	455.23	458.76	463.74	450.25
1008.51	1017.56	1025.54	1000.53	504.76	509.28	513.27	500.77
1136.61	1154.62	1153.63	1137.59	568.81	577.81	577.32	569.30

Table 1: Singly and doubly charged fragment ions of the HTLNQIDSVK tryptic peptide of the SwissProt P12763 FETUA BOVIN Alpha-2-HS-glycoprotein protein are listed.

1.3 Fragment Ion Matching

Given a peptide sequence and a tandem mass spectrum. For the assignment of a candidate peptide an in-silico fragment ion spectra fi is computed. The function findNN determines for each fragment ion the closed peak in the MS2. If the difference between the in-silico mass and the measured mass is inside the 'accuracy' mass window of the mass spec device the in-silico fragment ion is considered as a potential hit.

```
R> peptideSequence <- 'HTLNQIDSVK'
R> str(spec, nchar.max = 25, vec.len = 2)

List of 6
 $ scans      : num 1138
 $ title      : chr "178: (rt"| __truncated__
 $ rtinseconds: num 1343
 $ charge     : num 2
 $ mZ         : num [1:31] 195 221 ...
 $ intensity  : num [1:31] 932 322 ...

R> fi <- fragmentIon(peptideSequence)
R> n <- nchar(peptideSequence)
R> by.mZ <- c(fi[[1]]$b, fi[[1]]$y)
R> idx <- findNN(by.mZ, spec$mZ)
R> mZ.error <- abs(spec$mZ[idx]-by.mZ)
R> which(mZ.error < 0.3)

 [1]  2  3  4  5  6  7  8  9 13 14 16 17 18 19
```

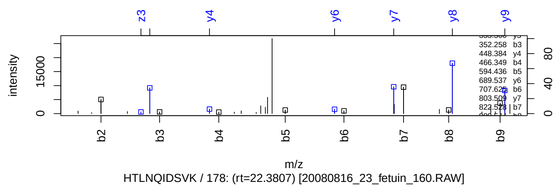
The function fragmentIon is also used for generating ion libraries in the specL Bioconductor package [5].

1.4 MS2 Labeling

The above-described peptide assignment is handled by the peakplot function.

```
R> p <- peakplot('HTLNQIDSVK', spec)
```

The plot below graphs a peptide-spectrum match of the 'HTLNQIDSVK' peptide.

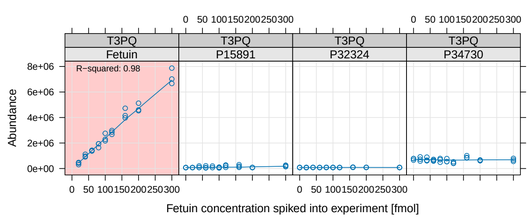


2 Quantification

For an overview of Quantitative Proteomics read [1, 2]. The authors are aware that meaningful statistics usually require a much higher number of biological replicates. In almost all cases there are not more than three to six repetitions. For the moment there are limited options due to the availability of machine time and the limits of the technologies.

2.1 Absolute Label-Free

The data set fetuinLFQ contains a subset of our results descriped in [3]. The example below shows a visualization using trellis plots. It graphs the abundance of four protein dependency from the fetuin concentration spiked into the sample.

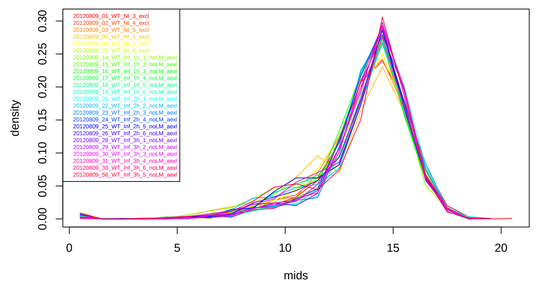


The plot shows the estimated concentration of the four proteins using the top three most intense peptides. The Fetuin peptides are spiked in with increasing concentration while the three other yeast proteins are kept stable in the background.

2.2 Relative Label-Free

LCMS based label-free quantification is a very popular method to extract relative quantitative information from mass spectrometry experiments. At the FGCZ we use the software ProgenesisLCMS for this workflow http://www.nonlinear.com/products/progenesis/lc-ms/overview/. Progenesis is a graphical software which does the aligning and extracts signal intensities from LCMS maps.

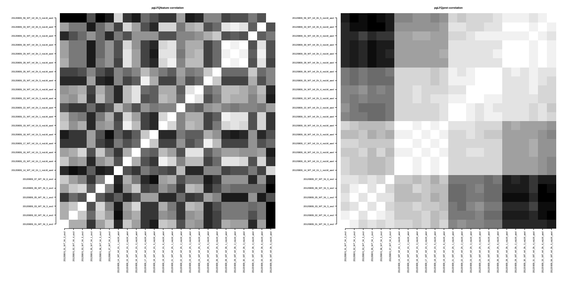
```
R> par(mfrow = c(1,1));
R> data(pgLFQfeature)
R> data(pgLFQprot)
R> protViz:::.featureDensityPlot(
+  asinh(
+  pgLFQfeature$"Normalized abundance"),
+  nbins=25)
```



The plots graph the normalized signal intensity distribution (logarithmic scaled) over the 24 LCMS runs aligned in this experiment.

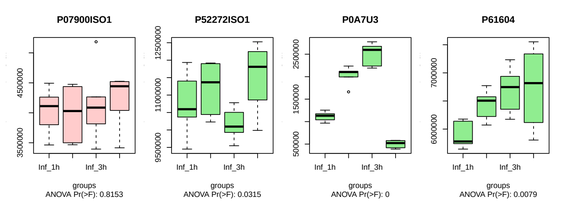
```
R> op <- par(mfrow=c(1,1),
+  mar = c(18,18,4,1),
+  cex=0.5)
R> samples <-
+  names(pgLFQfeature$"Normalized abundance")
R> image(cor(
+    asinh(
+      pgLFQfeature$"Normalized abundance")),
+  col = gray(seq(0,1,length=20)),
+  asp = 1,
+  main = 'pgLFQfeature correlation',
+  axes=FALSE)
R> axis(1,
+  at=seq(from = 0, to = 1,
+  length.out=length(samples)),
+  labels=samples, las=2)
R> axis(2,
+  at=seq(from = 0, to = 1,
+  length.out=length(samples)),
+  labels=samples, las=2)
R> par(op)
```

This left figure below shows the correlation between runs on feature level (values are asinh transformed). White color indicates the perfect correlation while black indicates a poor correlation.



This right figure above shows the correlation between runs on protein level (values are asinh transformed). White is the perfect correlation while black indicates a poor correlation. Striking is the fact that the six biological replicates for each condition cluster very well.

```
R> par(mfrow=c(1,4),mar=c(6,3,4,1))
R> ANOVA <- pgLFQaov(
+     pgLFQprot$"Normalized abundance",
+       groups=as.factor(pgLFQprot$grouping),
+       names=pgLFQprot$output$Accession,
+       idx=c(15,16,196,107),
+       plot=TRUE)
```



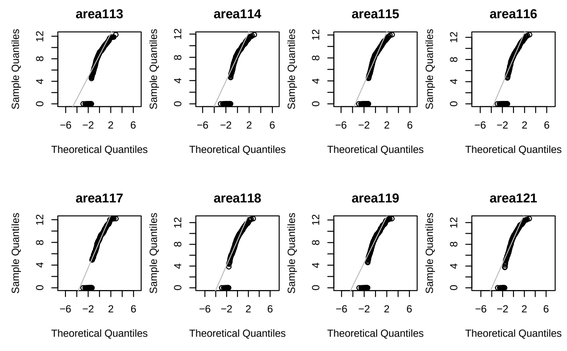
This figure shows the result of four proteins which either differ significantly in expression across conditions (green boxplots) using an analysis of variance test, or nondiffering protein expression (red boxplot).

2.3 iTRAQ – Two Group Analysis

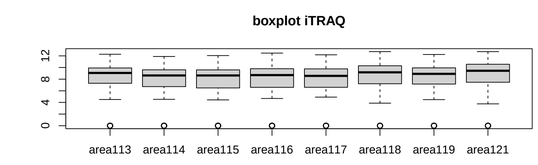
The data for the next section is an iTRAQ-8-plex experiment where two conditions are compared (each condition has four biological replicates)

2.3.1 Sanity Check

```
R> data(iTRAQ)
R> par(mfrow = c(2,4),
+       mar = c(6,4,3,0.5));
R> for (i in 3:10){
+       qqnorm(asinh(iTRAQ[,i]),
+       asp = 1,
+           main=names(iTRAQ)[i])
+       qqline(asinh(iTRAQ[,i]), col='grey')
+   }
```

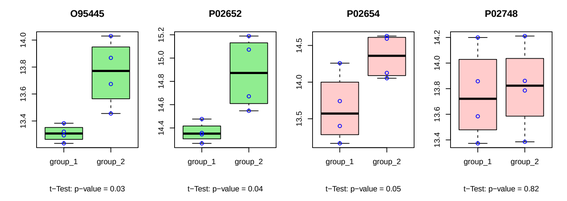



```
R> b <- boxplot(asinh(iTRAQ[, c(3:10)]),
+  main='boxplot iTRAQ')
```



A first check to see if all reporter ion channels are having the same distributions. Shown in the figure are Q-Q plots of the individual reporter channels against a normal distribution. The last is a boxplot for all individual channels.

2.3.2 On Protein Level



This figure shows five proteins which are tested using the t.test function if they differ across conditions using the four biological replicates.

References

[1] M. Bantscheff, S. Lemeer, M. M. Savitski, and B. Kuster. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem*, 404(4):939–965, Sep 2012.

[2] S. Cappadona, P. R. Baker, P. R. Cutillas, A. J. Heck, and B. van Breukelen. Current challenges in software solutions for mass spectrometry-based quantitative proteomics. *Amino Acids*, 43(3):1087–1108, Sep 2012.

[3] J. Grossmann, B. Roschitzki, C. Panse, C. Fortes, S. Barkow-Oesterreicher, D. Rutishauser, and R. Schlapbach. Implementation and evaluation of relative and absolute quantification in shotgun proteomics with label-free methods. *J Proteomics*, 73(9):1740–1746, Aug 2010.

[4] O. V. Krokhin, R. Craig, V. Spicer, W. Ens, K. G. Standing, R. C. Beavis, and J. A. Wilkins. An improved model for prediction of retention times of tryptic peptides in ion pair reversed-phase HPLC: its application to protein peptide mapping by off-line HPLC-MALDI MS. *Mol. Cell Proteomics*, 3(9):908–919, Sep 2004.

[5] C. Panse, C. Trachsel, J. Grossmann, and R. Schlapbach. specL–an R/Bioconductor package to prepare peptide spectrum matches for use in targeted proteomics. *Bioinformatics*, 31(13):2228–2231, Jul 2015.

[6] P. Roepstorff and J. Fohlman. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.*, 11(11):601, Nov 1984.

This poster has been produced by using the R CMD Sweave poster.Rnw commandline, R version 4.3.2 (2023-10-31), and protViz package version 0.7.9.

Contact: Functional Genomics Center Zurich, Y32 H94, Winterthurerstr. 190, CH-8057 Zurich, SWITZERLAND, Phone: +41 44 635 39 12, Fax: +41 44 635 39 22, EMail: {cp,jg}@fgcz.ethz.ch, URL: https://CRAN.R-project.org/package=protViz.