specL - Prepare Peptide Spectrum Matches for Use in Targeted Proteomics

Christian Panse*

Christian Trachsel
Jonas Grossmann †
Witold Wolski ‡

*cp@fgcz.ethz.ch †jg@fgcz.ethz.ch ‡wew@fgcz.ethz.ch

November 1, 2022

Contents

1 Introduction

Targeted proteomics is a fast evolving field in proteomics science and was even elected as method of the year in 2012 [1]. Especially targeted methods like SWATH [1] open promising perspectives for the identification and quantification of peptides and proteins. All targeted methods have in common the need of precise MS coordinates composed of precursor mass, fragment masses, and retention time. The combination of this information is kept in so called assays or spectra libraries. Here we present an R package able to produce such libraries out of peptide identification results (Mascot (dat), TPP (pep.xml and mzXMLs), ProteinPilot (group), Omssa (omx)). *specL* (see also [2]) is an easy to use, versatile and flexible function, which can be integrated into already existing commercial or non commercial analysis pipelines for targeted proteomics data analysis. Some examples of today's pipelines are ProteinPilot combined with Peakview (AB Sciex), Spectronaut (Biognosys) or OpenSwath [3].

[1] http://www.nature.com/nmeth/journal/v10/n1/pdf/nmeth.2329.pdf, 2014-09-22

In the following vignette it is described how the *specL* package can be used for the included data sets `peptideStd` and `peptideStd.redundant`.

2 Workflow

2.1 Prologue - How to get the input for the specL package?

Since peptide identification (using, e.g., Mascot, Sequest, xTandem!, Omssa, ProteinPilot) usually creates result files which are heavily redundant and therefore unsuited for spectral library building, the search results must first be filtered. To create non-redundant input files, we use the BiblioSpec [4] algorithm implemented in Skyline [5]. A given search result (e.g. Mascot result file) is loaded into the software Skyline and is redundancy filtered. The 'Skyline workflow step' provides two sqlite readable files[2] as output named `'*.blib'` and `'*.redundant.blib'`. These files are used as ideal input for this packages. Note here, that Skyline is very flexible when it comes to peptide identification results. It means with Skyline you can build the spectrum library files for almost all search engines (even from other spectrum library files such as spectraST [6]).

[2] sqlite uses standart SQL as query language.

The first step which has to be performed on the R shell is loading *specL* library.

```
R> library(specL)
R> packageVersion('specL')

[1] '1.32.0'
```

2.2 Read from redundant plus non-redundant blib files

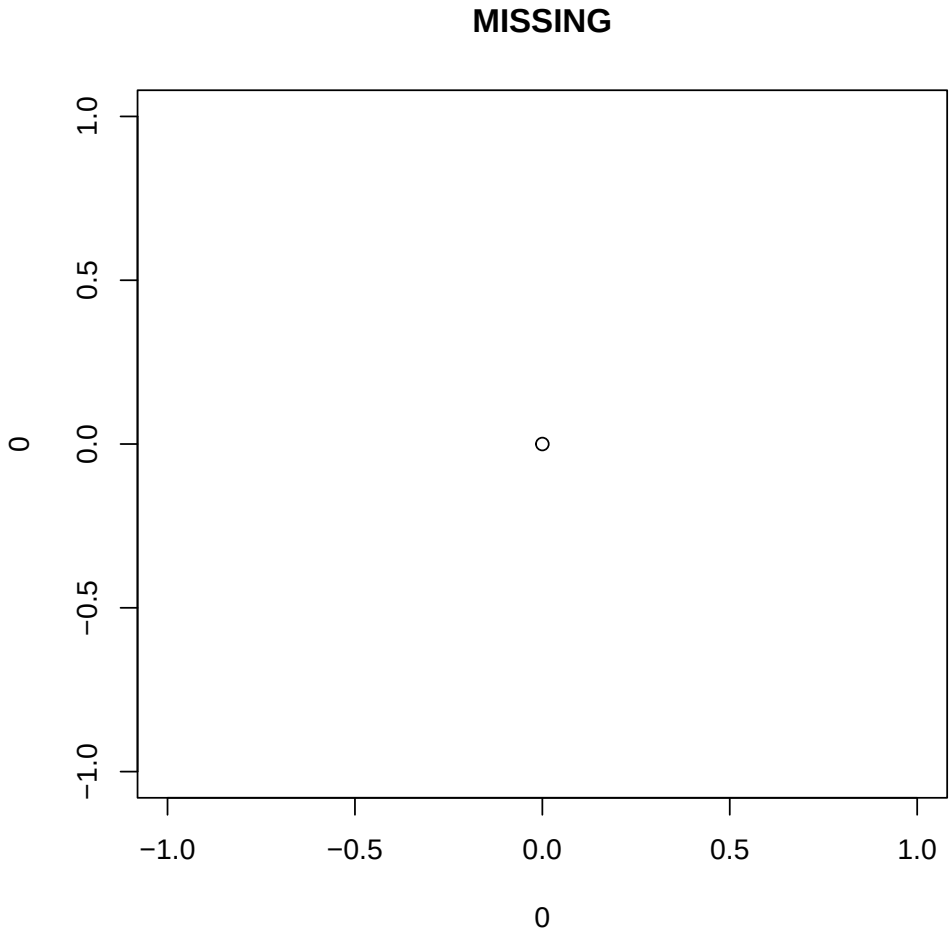
for demonstration *specL* contains the two data sets namely `peptideStd` and `peptideStd.redundant`. This is a data set which comes from two standard run experiments which are routinely used to check if the liquid chromatographic system is still working appropriate. The sample consists of a digest of the Fetuin protein (Bos taurus, uniprot id: P12763). 40 femtomole are loaded on column. Mascot was used to search and identify the respective peptides.

```
R> summary(peptideStd)

Summary of a "psmSet" object.
Number of precursor:
        137
Number of precursors in Filename(s)
        0140910_01_fetuin_400amol_1.raw        21
        0140910_07_fetuin_400amol_2.raw        116
Number of annotated precursor:
        0
```

For both `peptideStd`, `peptideStd.redandant` data sets the Skyline software was used to generate the bibliospec files which contain the peptide sequences with the respective peptide spectrum match (PSM). The `specL::read.bibliospec` function was used to read the blib files into R.

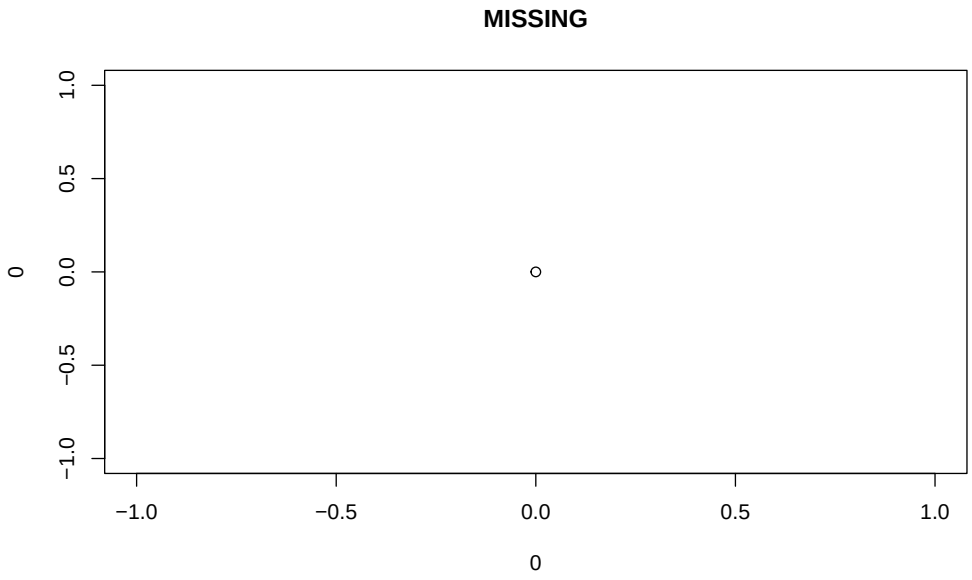
The from `read.bibliospec` generated object has its own plot functions. The LC-MS map graphs peptide mass versus retention time.

```
R> # plot(peptideStd)
R> plot(0,0, main='MISSING')
```



The individual peptide spectrum match (psm) is displayed by using the *protViz* `peakplot` function.

```
R> demoIdx <- 40
R> # str(peptideStd[[demoIdx]])
R> #res <- plot(peptideStd[[demoIdx]], ion.axes=TRUE)
R> plot(0,0, main='MISSING')
```



2.3 Read from Mascot result files

Alternatively Mascot search result files (dat) can be used by applying *protViz* perl script `protViz_mascotDat2RData.pl`.

The perl script can be found in the exec directory of the *protViz* package. The mascot mod_file can be found in the configurations of the mascot server. An example on our Linux shell looks as follow:

```
$ /usr/local/lib/R/site-library/protViz/exec/protViz_mascotDat2RData.pl \
-d=/usr/local/mascot/data/20130116/F178287.dat \
-m=mod_file
```

`mascotDat2RData.pl` requires the Mascot server `mod_file` keeping all the configured modification.

Once the perl script is finished, the resulting RData file can be read into the R session using `load`.

Next, the variable modifications, and the S3 psmSet object has to be generated. This can be done by using `specL:::.mascot2psmSet`

```
R> specL:::.mascot2psmSet

function (dat, mod, mascotScoreCutOff = 40)
{
    res <- lapply(dat, function(x) {
        x$MonoisotopicAAmass <- protViz::aa2mass(x$peptideSequence)[[1]]
```

```
            modString <- as.numeric(strsplit(x$modification, "")[[1]])
            modIdx <- which(modString > 0) - 1
            modString.length <- length(modString)
            x$varModification <- mod[modString[c(-1, -modString.length)] +
                1]
            if (length(modIdx) > 0) {
                warning("modified varModification caused.")
                x$varModification[modIdx] <- x$varModification[modIdx] -
                    x$MonoisotopicAAmass[modIdx]
            }
            rt <- x$rtinseconds
            x <- c(x, rt = rt, fileName = "mascot")
            class(x) <- "psm"
            return(x)
        })
        res <- res[which(unlist(lapply(dat, function(x) {
            x$mascotScore > mascotScoreCutOff && length(x$mZ) > 10
        })))]
        class(res) <- "psmSet"
        return(res)
    }
    <bytecode: 0x55d7038e7458>
    <environment: namespace:specL>
```

If you are processing Mascot result files you can continue reading in section 2.5.

However, please note, due do the potential high redundancy of peptide spectrum matches in a database search approach, it might not result in useful ion library for targeted data extraction, unless redundancy filtering is handled. However in a future release a redundancy filter algorithm might be proposed to resolve this problem.

2.4 Annotate protein ids using FASTA

The information to which protein a peptide-spectrum-match belongs (PSM) is not stored by BiblioSpec. Therefore *specL* provides the `annotate.protein_id` function which uses R's internal `grep` to 'reassigning' the protein information. Therefore a `fasta` object has to be loaded into the R system using `read.fasta` of the *seqinr* package. For this, not necessarily, the same `fasta` file needs to be provided as in the original database search.

The following lines demonstrate a simple sanity check with a single FASTA style formated protein entry. Also it demonstrates the use case how to identify entries in the R-object which are from one or a few proteins of interest.

```
R> irtFASTAseq <- paste(">zz|ZZ_FGCZCont0260|",
+  "iRT_Protein_with_AAAAK_spacers concatenated Biognosys\n",
+  "LGGNEQVTRAAAAKGAGSSEPVTGLDAKAAAAKVEATFGVDESNAKAAAAKYILAGVENS",
+  "KAAAAKTPVISGGPYEYRAAAAKTPVITGAPYEYRAAAAKDGLDAASYYAPVRAAAAKAD",
+  "VTPADFSEWSKAAAAKGTFIIDPGGVIRAAAAKGTFIIDPAAVIRAAAAKLFLQFGAQGS",
+  "PFLK\n")
R> Tfile <- file();  cat(irtFASTAseq, file = Tfile);
R> fasta.irtFASTAseq <- read.fasta(Tfile, as.string=TRUE, seqtype="AA")
R> close(Tfile)
```

As expected the `peptideStd` data, e.g., our demo Object, does not contain any protein information yet.

```
R> peptideStd[[demoIdx]]$proteinInformation

[1] ""
```

The protein information can be added as follow:

```
R> peptideStd <- annotate.protein_id(peptideStd,
+      fasta=fasta.irtFASTAseq)
```

The following lines show now the object indices of those entries which do hav a protein information now.

```
R> (idx<-which(unlist(lapply(peptideStd,
+      function(x){nchar(x$proteinInformation)>0}))))

[1] 1 2 3 4 5 6
```

As expected, there are now a number of peptide sequences annotated with the protein ID.

```
R> peptideStd[[demoIdx]]$proteinInformation

[1] "zz|ZZ_FGCZCont0260|"
```

Please note, that the default digest pattern is defined as

```
R> digestPattern = "(([RK])|(^)|(^M))"
```

for tryptic peptides. For other enzymes, the pattern has to be adapted. For example, for semi-tryptic identifications use `digestPattern = ""`.

2.5 Generate the spectral library (assay)

`genSwathIonLib` is the main contribution of the *specL* package. It generates the spectral library used in a targeted data extraction workflow from a mass spectrometric measurement. Generating the ion library using iRT peptides is highly recommended as described. However if you have no iRT peptide continue reading in section 2.7.

Generation of the spec Library with default (see Table 1) settings.

```
R> res.genSwathIonLib <- genSwathIonLib(data = peptideStd,
+     data.fit = peptideStd.redundant)
```

parameter	description	value
`max.mZ.Da.error`	max ms2 tolerance	`0.1`
`topN`	the n most intense fragment ion	10
`fragmentIonMzRange`	mZ range filter of fragment ion	c(300, 1800)
`fragmentIonRange`	min/max number of fragment ions	`c(5,100)`
`fragmentIonFUN`	desired fragment ion types	`b1+,y1+,b2+,y2+,b3+,y3+`

Table 1: `genSwathIonLib` default settings

```
R> summary(res.genSwathIonLib)

Summary of a "specLSet" object.

Parameter:

Number of precursor (q1 and peptideModSeq) = 137
Number of unique precursor
(q1.in-silico and peptideModSeq) = 126
Number of iRT peptide(s) = 8
Which std peptides (iRTs) where found in which raw files:
         0140910_01_fetuin_400amol_1.raw GAGSSEPVTGLDAK
          0140910_01_fetuin_400amol_1.raw TPVITGAPYEYR
          0140910_01_fetuin_400amol_1.raw VEATFGVDESNAK
          0140910_07_fetuin_400amol_2.raw ADVTPADFSEWSK
          0140910_07_fetuin_400amol_2.raw DGLDAASYYAPVR
          0140910_07_fetuin_400amol_2.raw GTFIIDPGGVIR
          0140910_07_fetuin_400amol_2.raw LFLQFGAQGSPFLK
          0140910_07_fetuin_400amol_2.raw TPVISGGPYEYR
```

```
Number of transitions frequency:
        4        1
        5        5
        6        10
        7        7
        8        18
        9        32
        10        64

Number of annotated precursor = 6
Number of file(s)
        2

Number of precursors in Filename(s)
        0140910_01_fetuin_400amol_1.raw        21
        0140910_07_fetuin_400amol_2.raw        116

Misc:

Memory usage        =         676976 bytes
```

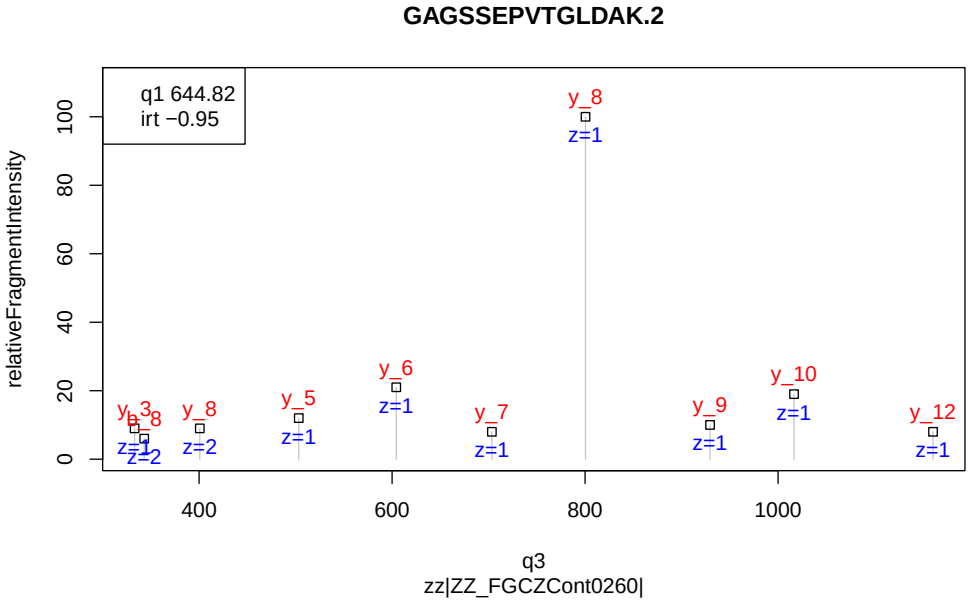
The determined mass spec coordinates of the selected tandem mass spectrum `demoIdx` look like this:

```
R> res.genSwathIonLib@ionlibrary[[demoIdx]]

An "specL" object.


content:
group_id = GAGSSEPVTGLDAK.2
peptide_sequence = GAGSSEPVTGLDAK
proteinInformation = zz|ZZ_FGCZCont0260|
q1 = 644.8219
q1.in_silico = 1288.638
q3 = 800.4497 604.3285 1016.522 503.2805 929.4925 400.7282
333.176 1160.581 703.3948 343.1235
q3.in_silico = 800.4512 604.3301 1016.526 503.2824 929.4938
400.7295 333.1769 1160.579 703.3985 343.1615
prec_z = 2
frg_type = y y y y y y y y b
frg_nr = 8 6 10 5 9 8 3 12 7 8
```

```
frg_z = 1 1 1 1 1 2 1 1 1 2
relativeFragmentIntensity = 100 21 19 12 10 9 9 8 8 6
irt = -0.95
peptideModSeq = GAGSSEPVTGLDAK
mZ.error = 0.001514 0.00156 0.003685 0.001914 0.001318
0.001313 0.000856 0.001846 0.003686 0.0380015
uclei_diff_extraction_methods\20140910_01_fetuin_400amol_1.raw
score = 41.54902

size:
Memory usage: 4776 bytes
```

It can be displayed using the `specL::plot` function.

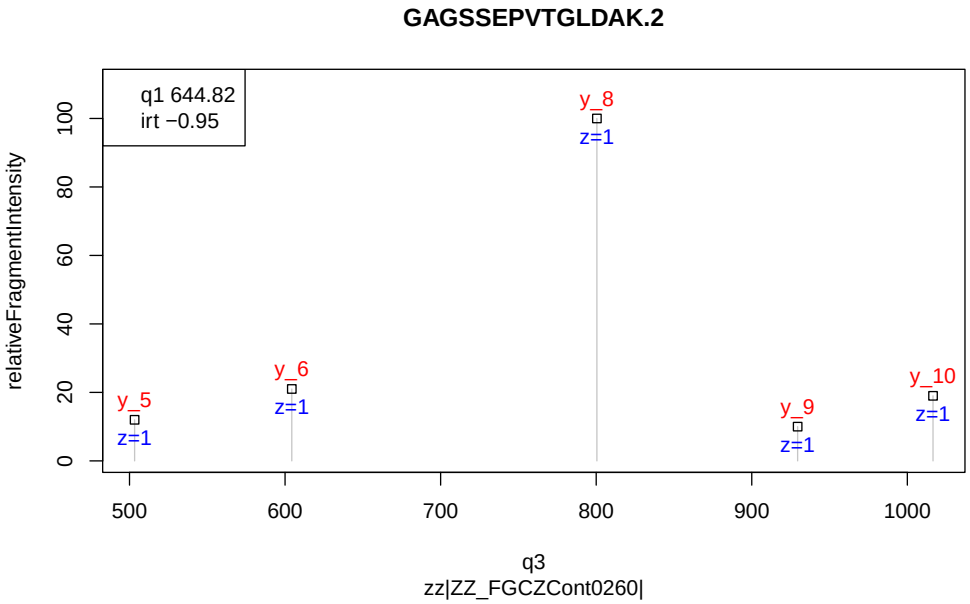
```
R> plot(res.genSwathIonLib@ionlibrary[[demoIdx]])
```



The following code considers only the top five y ions.

```
R> # define customized fragment ions
R> # for demonstration lets consider only the top five singly charged y ions.
R>
R> r.genSwathIonLib.top5 <- genSwathIonLib(peptideStd,
+      peptideStd.redundant, topN=5,
+      fragmentIonFUN=function (b, y) {
+        return( cbind(y1_=y) )
```

```
+         }
+       )
R> plot(r.genSwathIonLib.top5@ionlibrary[[demoIdx]])
R>
```



2.6 Normalizing the retention time using iRT peptides

Retention time is a very important parameter in targeted data extraction. However retention times are not easy to transfer between different reverse phase columns or HPLC systems. To make transfer applicable and account for inter run shift in retention time Biognosys [7] invented the iRT normalization based on iRT / HRM peptides. For this, a set of well behaving peptides (good flying properties, good fragmention characteristics, completely artifical) which cover the whole rt-gradient are spiked into each sample. For this set of peptide an idependent retention time (dimension less) is suggested by Biognosys. With this at hand, the set of peptides can later be used to apply a linear regression model to adapt all measured retention times into an independent retention time scale.

If the identification results contain iRT peptides the package supports the conversion to the iRT scale. For this (if the identification result are based on multiple input files) the redundant BiblioSpec file is required where all iRT peptides from all measurements are stored. For the most representative spectrum in the non-redundant R-object the original filename is identified and the respective

linear model for this one particular MS experiment is applied to normalize the retention time to the iRT scale. The iRT peptides as well as their independent retention times are stored in the `iRTpeptides` object.

specL uses by default the iRT peptide table to normalize into the independent retention time but could also be extended or changed to custom iRT peptides if available.

```
R> iRTpeptides
```

The method genSwathIonLib uses:

```
R> fit <- lm(formula = rt ~ aggregateInputRT * fileName,
+    data=m)
```

to build the linear models for each MS measurement individually. For defining `m` both data sets were aggregated over the attributes `peptide` and `fileName` using the `mean` operator.

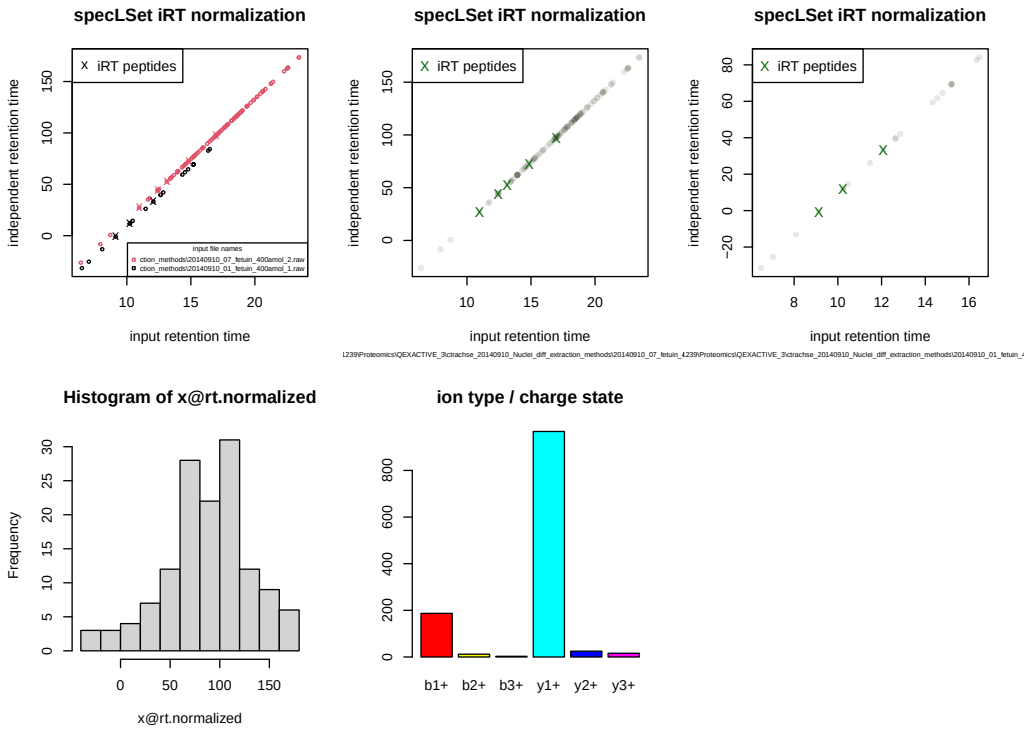
```
R> data <- aggregate(df$rt, by = list(df$peptide, df$fileName),
+    FUN = mean)
R> data.fit <- aggregate(df.fit$rt,
+    by = list(df.fit$peptide, df.fit$fileName),
+    FUN = mean)
```

Afterwards the following join operator was applied.

```
R> m <- merge(iRT, data.fit, by.x='peptide', by.y='peptide')
```

The following graph displays the normalized retention time versus the measured retention time after applying the calculated model to the two data sets.

```
R> # calls the plot method for a specLSet object
R> op <- par(mfrow=c(2,3))
R> plot(res.genSwathIonLib)

[1] 16.83185 13.13262 18.54058 18.36923 15.30478 15.30478
[1]  7.032372  6.490769 14.787681 14.544429 15.207398
[6] 15.207398

R> par(op)
```



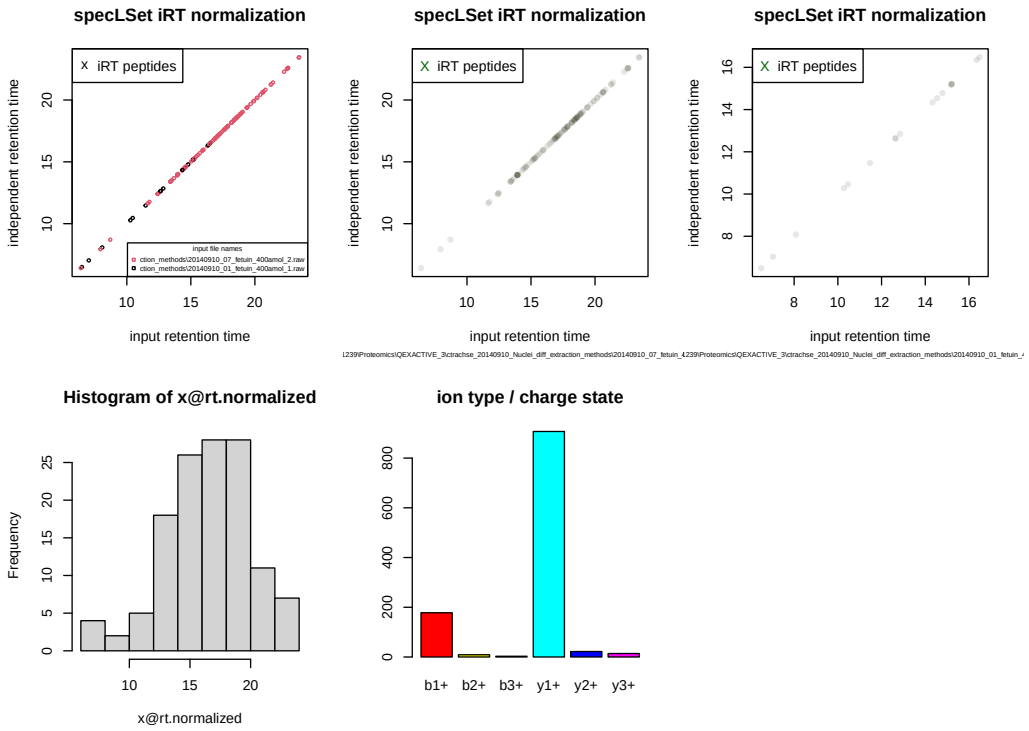
Shown are original retention time (in minutes) and iRT (dimensionless) for two standard run experiments (color black and red). Indicated with black **X** are the iRT peptides which are the base for the regression.

2.7 Generate the spectral library having no iRTs

If no iRT peptides are contained in the data not iRT normalization is applied. The scatter plot below shows on the y axis that there were not iRT transformation.

```
R> idx.iRT <- which(unlist(lapply(peptideStd,
+    function(x){
+      if(x$peptideSequence %in% iRTpeptides$peptide){0}
+      else{1}
+    })) == 0)
R> # remove all iRTs and compute ion library
R> res.genSwathIonLib.no_iRT <- genSwathIonLib(peptideStd[-idx.iRT])
R> summary(res.genSwathIonLib.no_iRT)

Summary of a "specLSet" object.

Parameter:
```

```
Number of precursor (q1 and peptideModSeq) = 129
Number of unique precursor
(q1.in-silico and peptideModSeq) = 118
Number of iRT peptide(s) = 0
Number of transitions frequency:
        4        1
        5        5
        6        10
        7        7
        8        17
        9        31
        10        58

Number of annotated precursor = 0
Number of file(s)
        2

Number of precursors in Filename(s)
        0140910_01_fetuin_400amol_1.raw        18
        0140910_07_fetuin_400amol_2.raw        111

Misc:

Memory usage        =         630368 bytes

R> op <- par(mfrow = c(2, 3))
R> plot(res.genSwathIonLib.no_iRT)

[1] 16.83185 18.54058 18.36923 15.30478 15.30478 19.36682
[1]  7.032372  6.490769 14.787681 14.544429 15.207398
[6] 15.207398

R> par(op)
```



2.8 Write output to file

The output can be written as an ASCII text file.

```
R> write.spectronaut(res.genSwathIonLib,
+    file="specL-Spectronaut.txt")
```

2.9 Epilogue - What can I do with that library now?

The specL output text file can directly be used as input (assay) for the Spectronaut software from Biognosys or with minimal reshaping for Peakview. Alternatively it can be used as a basis for script based construction of SRM/MRM assays.

2.10 Benchmark

The benchmarks were processed on a 12 core XEON Server (X5650 @ 2.67GHz) running Linux Debian wheezy having R version 3.1.1 (2014-07-10) , specL 1.1.2, and BiocParallel 1.0.0 installed. The default setting of BiocParallel has used eight cores. As FASTA we used a TAIR10 retrived from http://www.arabidopsis.org/ and Human Swissprot.

fasta=TAIR10			blib [unpublished]		runtime	
#proteins	#tryptic peptides	file size	#specs	file size	annotate	generate
71032	3423196	39M	39648/118268	51M	79min	19sec
71032	3423196	39M	65018/136963	120M	130min	30sec
fasta=HUMAN			blib [8, Rosenberger]			
88969	3997085	43M	256908/3060421	4.4G	≈7h	≈5min

The following parameter settings were given to the `genSwathIonLib` function:

```
R> res <- genSwathIonLib(data, data.fit,
+    topN=10,
+    fragmentIonMzRange=c(200,2000),
+    fragmentIonRange=c(2,100))
```

3 Acknowledgement

The authors thank all colleagues of the Functional Genomics Center Zuerich (FGCZ), and especial thank goes to our test users Sira Echevarría Zomeño (ETHZ), Tobias Kockmann (ETHZ), Lukas von Ziegler (Brain Research Institute, UZH|ETH Zurich), and Stephan Michalik (Ernst-Moritz-Arndt-Universität Greifswald, Germany).

4 TODO for next releases

- importer for peakview csv format; enable `compare.specLSet(object0, object1)`
- new option for `specL::genSwathIonLib`; Exclude fragment ions from precursor `window = TRUE, FALSE`
- new option for `specL::genSwathIonLib`; Predict transitions for heavy labeled peptides using information from light peptides `predictHeavy = TRUE,FALSE, LabelFile = "fileWithHeavyAA"`
- new export function into TraML format for compatibility with OpenSWATH [3]

- replace *seqinr* `read.fasta` by using *Biostrings* `readAAStringSet` to handle fasta files
- add varMods to specL class
- replace Mascot score by a generic score
- in-silico rt ion map plot (`plot.specLSet`) split window into SWATH windows (one plot per, e.g., 25Da window)
- assay refinement - replace contaminated fragment ion in library

5 Session information

An overview of the package versions used to produce this document are shown below.

- R version 4.2.1 (2022-06-23), `x86_64-pc-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_GB`, `LC_COLLATE=C`, `LC_MONETARY=en_US.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=en_US.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Running under: `Ubuntu 20.04.5 LTS`
- Matrix products: default
- BLAS: `/home/biocbuild/bbs-3.16-bioc/R/lib/libRblas.so`
- LAPACK: `/home/biocbuild/bbs-3.16-bioc/R/lib/libRlapack.so`
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: BiocStyle 2.26.0, DBI 1.1.3, RSQLite 2.2.18, knitr 1.40, protViz 0.7.3, seqinr 4.2-16, specL 1.32.0
- Loaded via a namespace (and not attached): BiocManager 1.30.19, MASS 7.3-58.1, R6 2.5.1, Rcpp 1.0.9, ade4 1.7-19, bit 4.0.4, bit64 4.0.5, blob 1.2.3, bookdown 0.29, bslib 0.4.0, cachem 1.0.6, cli 3.4.1, codetools 0.2-18, compiler 4.2.1, digest 0.6.30, evaluate 0.17, fastmap 1.1.0, highr 0.9, htmltools 0.5.3, jquerylib 0.1.4, jsonlite 1.8.3, magick 2.7.3, magrittr 2.0.3, memoise 2.0.1, parallel 4.2.1, rlang 1.0.6, rmarkdown 2.17, sass 0.4.2, stringi 1.7.8, stringr 1.4.1, tools 4.2.1, vctrs 0.5.0, xfun 0.34, yaml 2.3.6

References

[1] L. C. Gillet, P. Navarro, S. Tate, H. Rost, N. Selevsek, L. Reiter, R. Bonner, and R. Aebersold. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol. Cell Proteomics*, 11(6):O111.016717, Jun 2012.

[2] Christian Panse, Christian Trachsel, Jonas Grossmann, and Ralph Schlapbach. specl—an r/bioconductor package to prepare peptide spectrum matches for use in targeted proteomics. *Bioinformatics*, 31(13):2228, 2015. URL: +http://dx.doi.org/10.1093/bioinformatics/btv105, arXiv:/oup/backfile/Content_public/Journal/bioinformatics/31/13/10.1093/bioinformatics/btv105/2/btv105.pdf, doi:10.1093/bioinformatics/btv105.

[3] H. L. Rost, G. Rosenberger, P. Navarro, L. Gillet, S. M. Miladinovi?, O. T. Schubert, W. Wolski, B. C. Collins, J. Malmstrom, L. Malmstrom, and R. Aebersold. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.*, 32(3):219–223, Mar 2014.

[4] B. Frewen and M. J. MacCoss. Using BiblioSpec for creating and searching tandem MS peptide libraries. *Curr Protoc Bioinformatics*, Chapter 13:Unit 13.7, Dec 2007.

[5] B. MacLean, D. M. Tomazela, N. Shulman, M. Chambers, G. L. Finney, B. Frewen, R. Kern, D. L. Tabb, D. C. Liebler, and M. J. MacCoss. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*, 26(7):966–968, Apr 2010.

[6] H. Lam, E. W. Deutsch, J. S. Eddes, J. K. Eng, S. E. Stein, and R. Aebersold. Building consensus spectral libraries for peptide identification in proteomics. *Nat. Methods*, 5(10):873–875, Oct 2008.

[7] C. Escher, L. Reiter, B. MacLean, R. Ossola, F. Herzog, J. Chilton, M. J. MacCoss, and O. Rinner. Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics*, 12(8):1111–1121, Apr 2012.

[8] G. Rosenberger.et al. A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Sci. Data*, 2015. 1:140031 doi: 10.1038/sdata.2014.31.