# In-silico cleavage of polypeptides using the cleaver package

Sebastian Gibb*

May 2, 2014

## Contents

1 Introduction 1

2 Simple Usage 1

3 Insulin & Somatostatin Example 2

4 Isotopic Distribution Of Tryptic Digested Insulin 3

5 Session Information 5

## 1 Introduction

Most proteomics experiments need protein (peptide) separation and cleavage procedures before these molecules could be analyzed or identified by mass spectrometry or other analytical tools.
*cleaver* allows in-silico cleavage of polypeptide sequences to e.g. create theoretical mass spectrometry data.
The cleavage rules are taken from the ExPASy PeptideCutter tool[4].

## 2 Simple Usage

Loading the *cleaver* package:

```
> library("cleaver")
```

Getting help and list all available cleavage rules:

*mail@sebastiangibb.de

```
> help("cleave")
```

Cleaving of *Gastric juice peptide 1 (P01358)* using *Trypsin*:

```
> cleave("LAAGKVEDSD", enzym = "trypsin")

$LAAGKVEDSD
[1] "LAAGK" "VEDSD"
```

Sometimes cleavage is not perfect and the enzym miss some cleavage positions:

```
> ## miss one cleavage position
> cleave("LAAGKVEDSD", enzym = "trypsin", missedCleavages = 1)

$LAAGKVEDSD
[1] "LAAGKVEDSD"

>
> ## miss zero or one cleavage positions
> cleave("LAAGKVEDSD", enzym = "trypsin", missedCleavages = 0:1)

$LAAGKVEDSD
[1] "LAAGK"      "VEDSD"      "LAAGKVEDSD"
```

Combine *cleaver* and the *Biostrings* R package[5]:

```
> ## create AAStringSet object
> gaju <- AAStringSet("LAAGKVEDSD")
>
> ## cleave it
> cleave(gaju, enzym = "trypsin")

AAStringSetList of length 1
[["LAAGKVEDSD"]] LAAGK VEDSD
```

# 3 Insulin & Somatostatin Example

Downloading *Insulin (P01308)* and *Somatostatin (P61278)* sequences from the UniProt[6] database using the *UniProt.ws* R package[1].

```
> ## load UniProt.ws library
> library("UniProt.ws")
>
> ## download sequences of Insulin/Somatostatin
> s <- select(UniProt.ws, keys = c("P01308", "P61278"),
+     columns = c("SEQUENCE"))

Getting extra data for P01308 P61278 NA etc

>
> ## fetch only sequences
> sequences <- setNames(s$SEQUENCE, s$UNIPROTKB)
```

```
>
> ## remove whitespaces
> sequences <- gsub(pattern = "[[:space:]]", replacement = "",
+     x = sequences)
```

Cleaving using *Pepsin*:

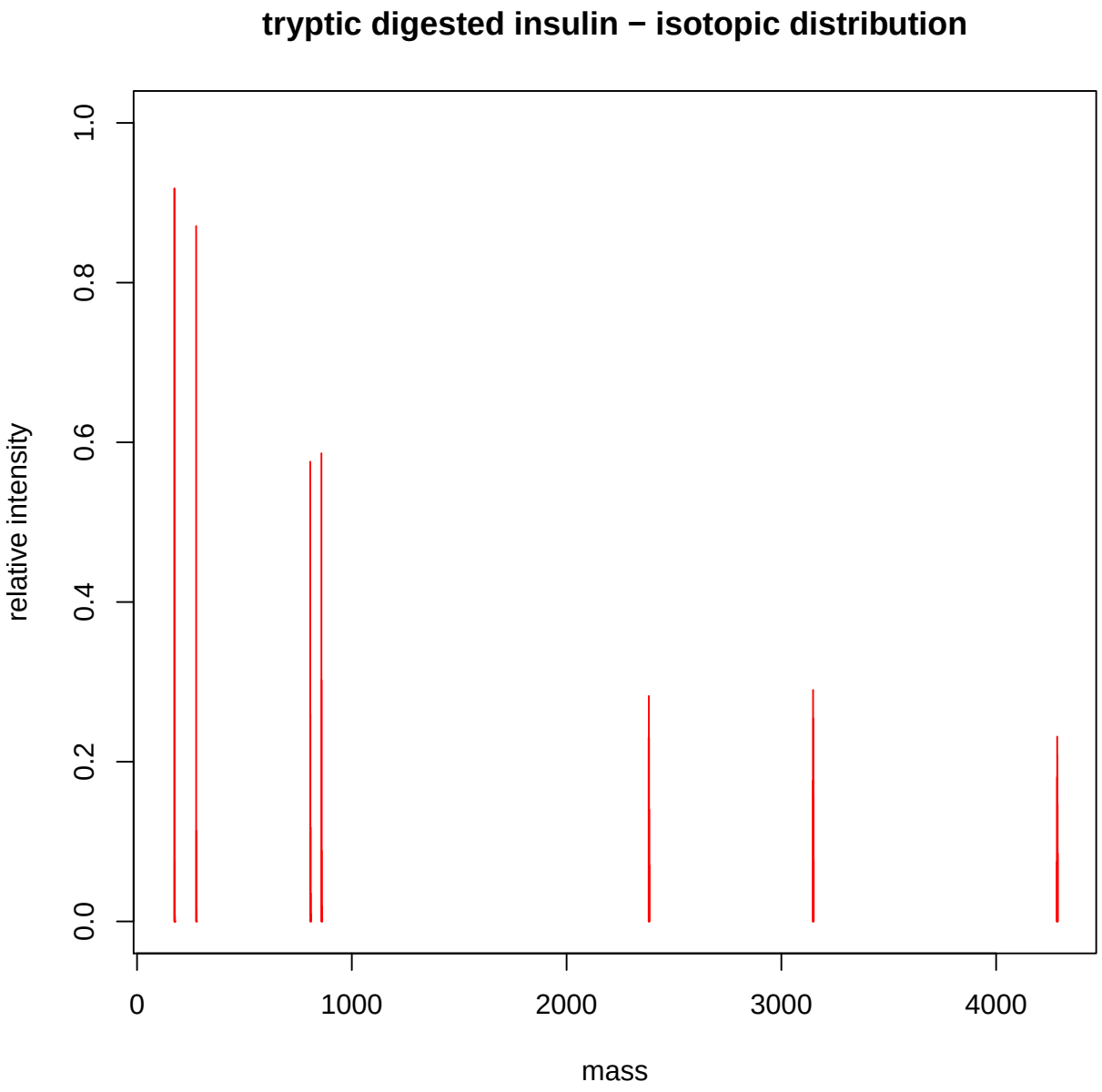
```
> cleave(sequences, enzym = "pepsin")

$P01308
 [1] "MA"              "L"               "WMRLLP"          "LL"
 [5] "A"               "L"               "L"               "A"
 [9] "L"               "WGPDPAAA"        "F"               "VNQH"
[13] "L"               "CGSH"            "L"               "VEA"
[17] "L"               "Y"               "L"               "VCGERG"
[21] "FF"              "YTPKTRREAED"     "L"               "QVGQVE"
[25] "L"               "GGGPGAGS"        "LQP"             "LA"
[29] "L"               "EGS"             "L"               "QKRGIVEQCCTSICS"
[33] "L"               "YQ"              "L"               "ENYCN"

$P61278
 [1] "ML"                    "SCRL"                  "QCA"
 [4] "L"                     "AA"                    "L"
 [7] "SIV"                   "L"                     "A"
[10] "L"                     "GCVTGAPSDPRL"          "RQ"
[13] "FL"                    "QKS"                   "LAAAAGKQEL"
[16] "AKY"                   "FL"                    "AE"
[19] "L"                     "L"                     "SEPNQTENDA"
[22] "LEPED"                 "L"                     "SQAAEQDEMRL"
[25] "EL"                    "QRSANSNPAMAPRERKAGCKN" "FF"
[28] "WKT"                   "FTSC"
```

## 4 Isotopic Distribution Of Tryptic Digested Insulin

A common use case of in-silico cleavage is the calculation of the isotopic distribution of peptides (which were enzymatic digested in the in-vitro experimental workflow). Here the *BRAIN* R package[2;3] is used to calculate the isotopic distribution of *cleaver*'s output. (please note: it is only a toy example, e.g. the relation of intensity values between peptides isn't correct).

```
> ## load BRAIN library
> library("BRAIN")
>
> ## cleave insulin
> cleavedInsulin <- cleave(sequences[1], enzym = "trypsin")[[1]]
>
> ## create empty plot area
> plot(NA, xlim = c(150, 4300), ylim = c(0, 1),
+     xlab = "mass", ylab = "relative intensity",
```

```
+     main = "tryptic digested insulin - isotopic distribution")
>
> ## loop through peptides
> for (i in seq(along = cleavedInsulin)) {
+     ## count C, H, N, O, S atoms in current peptide
+     atoms <- BRAIN::getAtomsFromSeq(cleavedInsulin[[i]])
+     ## calculate isotopic distribution
+     d <- useBRAIN(atoms)
+     ## draw peaks
+     lines(d$masses, d$isoDistr, type = "h", col = 2)
+ }
```



# 5 Session Information

- R version 3.1.0 (2014-04-10), `x86_64-apple-darwin13.1.0`
- Locale: `C/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8`
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, utils
- Other packages: BRAIN 1.10.0, BiocGenerics 0.10.0, Biostrings 2.32.0, DBI 0.2-7, IRanges 1.22.5, PolynomF 0.94, RCurl 1.95-4.1, RSQLite 0.11.4, UniProt.ws 2.4.1, XVector 0.4.0, bitops 1.0-6, cleaver 1.2.0, knitr 1.5, lattice 0.20-29
- Loaded via a namespace (and not attached): AnnotationDbi 1.26.0, Biobase 2.24.0, BiocStyle 1.2.0, GenomeInfoDb 1.0.2, evaluate 0.5.5, formatR 0.10, grid 3.1.0, highr 0.3, stats4 3.1.0, stringr 0.6.2, tools 3.1.0, zlibbioc 1.10.0

# References

[1] Marc Carlson. *UniProt.ws: R Interface to UniProt Web Services*. R package version 2.0.0.

[2] Jürgen Claesen, Piotr Dittwald, Tomasz Burzykowski, and Dirk Valkenborg. An efficient method to calculate the aggregated isotopic distribution and exact center-masses. *Journal of The American Society for Mass Spectrometry*, 23(4):753–763, 2012.

[3] Piotr Dittwald, Jürgen Claesen, Tomasz Burzykowski, Dirk Valkenborg, and Anna Gambin. Brain: A universal tool for high-throughput calculations of the isotopic distribution for mass spectrometry. *Analytical chemistry*, 85(4):1991–1994, 2013.

[4] Elisabeth Gasteiger, Christine Hoogland, Alexandre Gattiker, S'everine Duvaud, Marc R. Wilkins, Ron D. Appel, and Amos Bairoch. Protein identification and analysis tools on the expasy server. In John M. Walker, editor, *The Proteomics Protocols Handbook*, pages 571–607. Humana Press, 2005. ISBN 978-1-58829-343-5. doi: 10.1385/1-59259-890-0:571. URL http://dx.doi.org/10.1385/1-59259-890-0%3A571.

[5] H. Pages, P. Aboyoun, R. Gentleman, and S. DebRoy. *Biostrings: String objects representing biological sequences, and matching algorithms*. R package version 2.28.0.

[6] The UniProt Consortium. Reorganizing the protein space at the universal protein resource (uniprot). *Nucleic Acids Research*, 40(D1):D71–D75, 2012. doi: 10.1093/nar/gkr981. URL http://nar.oxfordjournals.org/content/40/D1/D71.abstract.