# TRGN 527: Applied Data Science and Bioinformatics

## UNIT I. Introduction and Basic Data Science

## Week 4 – Lecture 2 – Case Study Part 2

**Enrique I. Velazquez Villarreal, M.D., Ph.D., M.P.H., M.S. | Assistant Professor**

Dept. of Translational Genomics

USC **|** Keck School of Medicine | Norris Comprehensive Cancer Center

Leader of the USC Bioinformatics Core – *USC CaRE2 Health Equity Center*


**David W. Craig, Ph.D. | Professor and Vice Chair**

Dept. of Translational Genomics

USC **|** Keck School of Medicine | Norris Comprehensive Cancer Center

Co-Director, Institute of Translational Genomics

- Peptide Identification by using MS/MS spectra
  - For complex protein samples it is useful the application of tandem MS spectra (MS/MS-based peptide identification).
  - Naming conventions of peptide fragments after fragmenting a protein at a peptide bond.
    - Fragment: a, b, c
    - N-terminus fragment: R1-, R1-CO-, R1-CO-NH-
    - C-terminus fragment: -CO-NH-R2, -NH-R2, -R2
    - Fragment name: x, y, z

Picture, fragmentation of a protein at a peptide bond
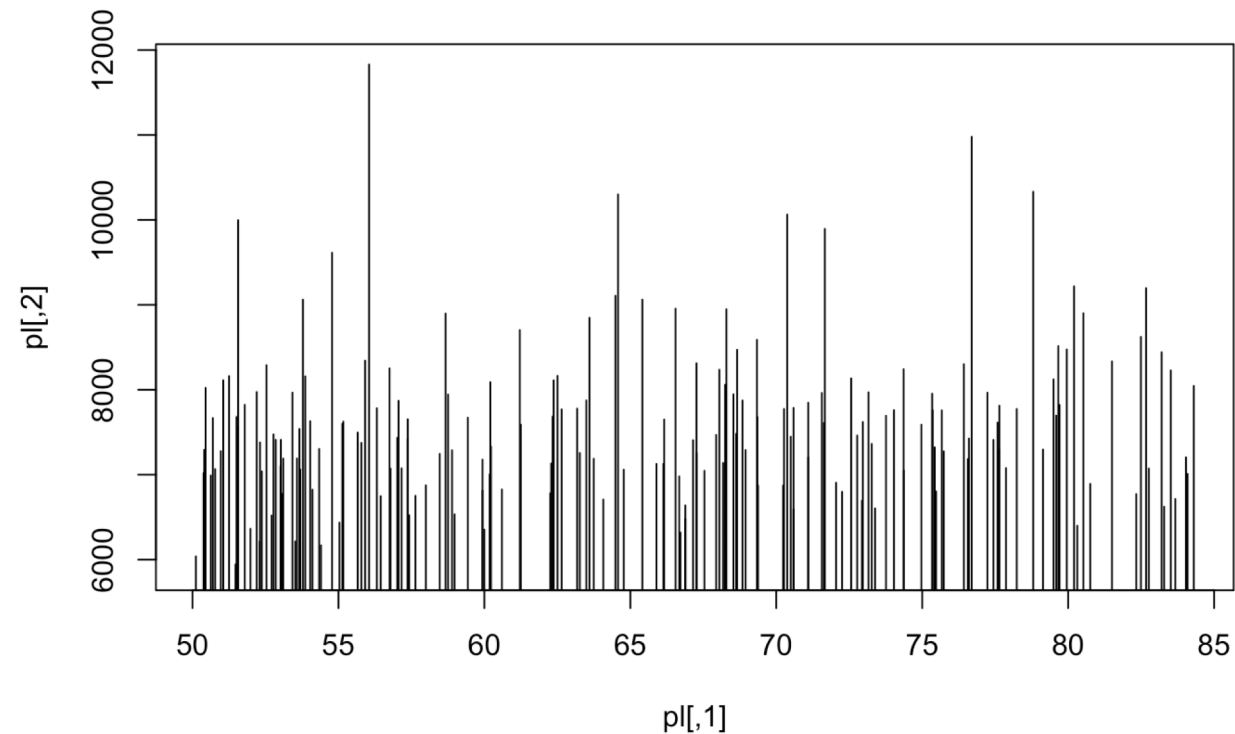
- Load the dataset:

```r
94 # Peptide Isentification by using MS/MS spectra
95 # Load the dataset:
96 ```{r}
97
98 library(mzR)
99 msfile <- system.file("threonine/threonine_i2_e35_pH_tree.mzXML",
100                       package = "msdata")
101 my.data <- openMSfile(msfile)
102 pl <- peaks(my.data,7)
103
104 ```
```

# Case Study – Proteomics Analysis

- Let us have a look at the spectrum:



Note: Noise with some strong peaks (m/z, intensity)

- Selecting the strongest peaks:

```r
113   # Pick up only the strongest peaks:
114
115   ```{r}
116
117   topnum <- 15
118   pl.top <- pl[pl[,2] %in% head(sort(pl[,2],decreasing=T),topnum),1]
119
120   ```
```

- Creating a matrix that contains the m/z differences between the different peaks:

```r
#Creating a matrix that contains the m/z differences between the different peaks:
```{r}

peakdiff <- outer(pl.top, pl.top, '-')
plot(density(abs(peakdiff)))

```
```

**density.default(x = abs(peakdiff))**



N = 225   Bandwidth = 2.464