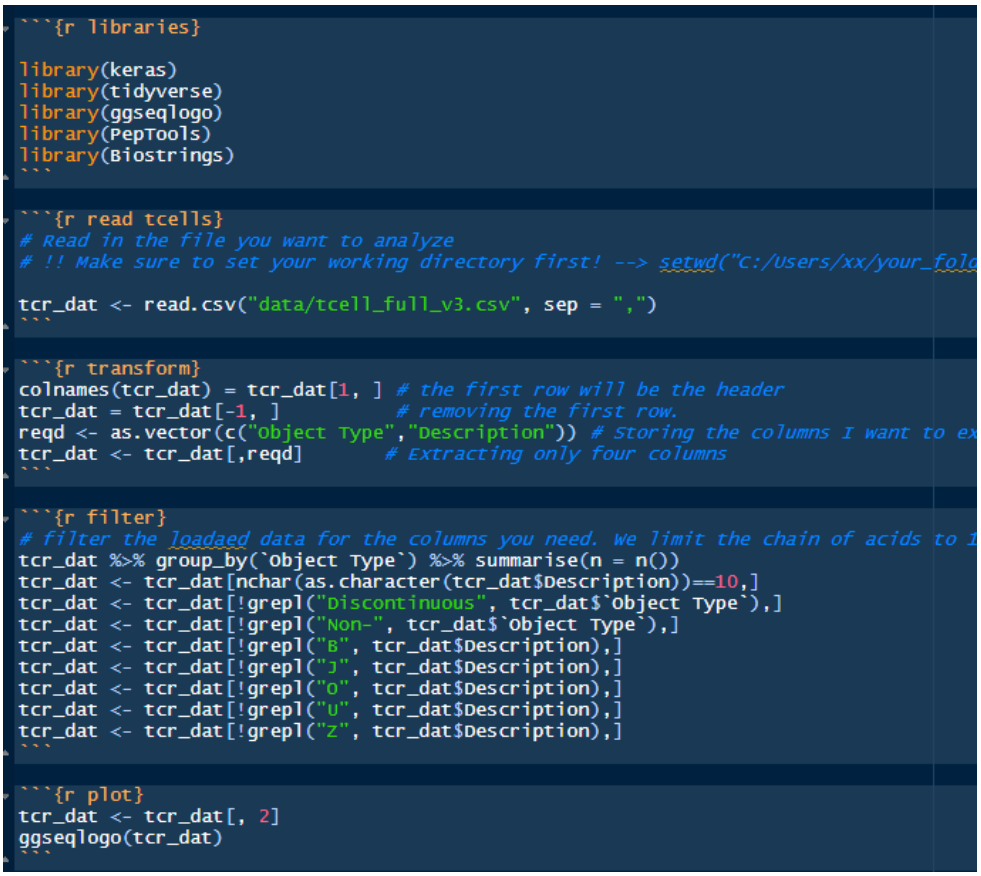
Immunoinformatics

Exercise

Investigate the peptide dataset and find binding patterns in epitopes of BCR versus TCR.



```r
```{r libraries}

library(keras)
library(tidyverse)
library(ggseqlogo)
library(PepTools)
library(Biostrings)
```

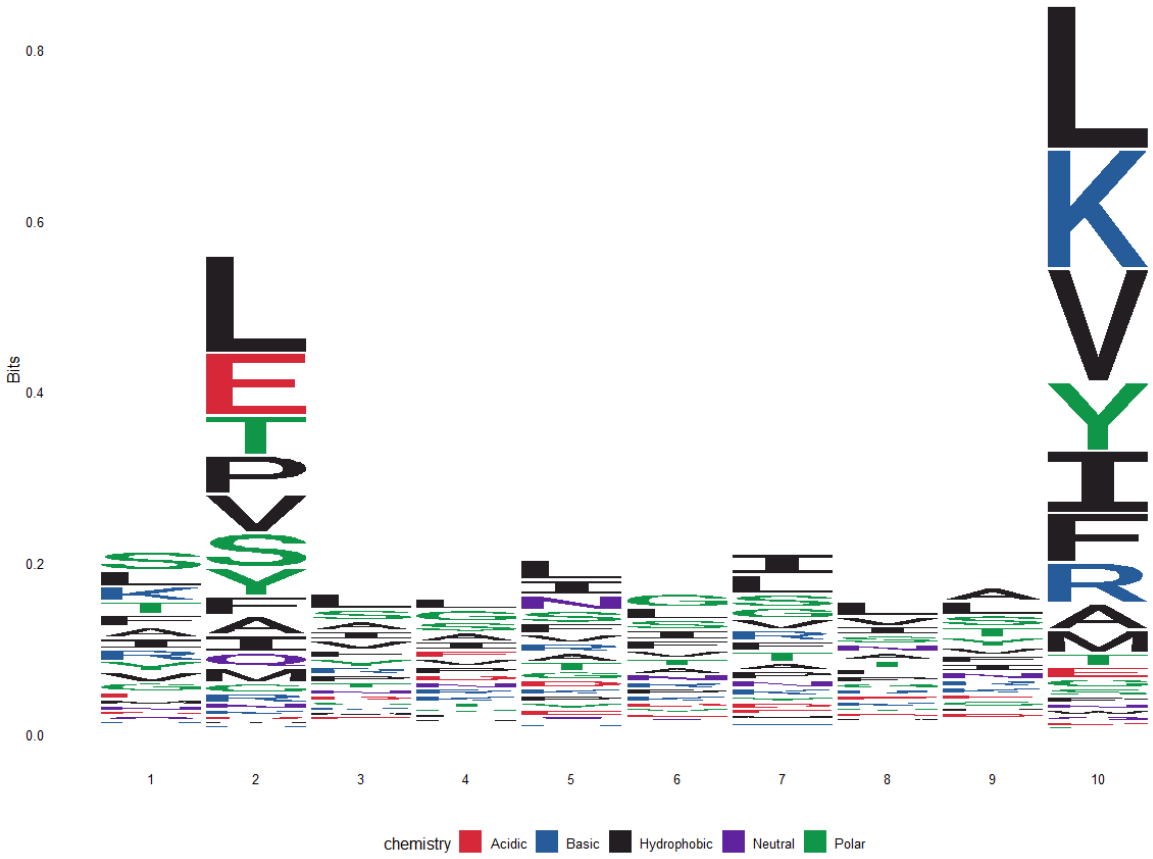
```{r read tcells}
# Read in the file you want to analyze
# !! Make sure to set your working directory first! --> setwd("C:/Users/xx/your_folo

tcr_dat <- read.csv("data/tcell_full_v3.csv", sep = ",")
```

```{r transform}
colnames(tcr_dat) = tcr_dat[1, ] # the first row will be the header
tcr_dat = tcr_dat[-1, ]          # removing the first row.
reqd <- as.vector(c("Object Type","Description")) # storing the columns I want to ex
tcr_dat <- tcr_dat[,reqd]       # Extracting only four columns
```

```{r filter}
# filter the loadaed data for the columns you need. We limit the chain of acids to 1
tcr_dat %>% group_by(`Object Type`) %>% summarise(n = n())
tcr_dat <- tcr_dat[nchar(as.character(tcr_dat$Description))==10,]
tcr_dat <- tcr_dat[!grepl("Discontinuous", tcr_dat$`Object Type`),]
tcr_dat <- tcr_dat[!grepl("Non-", tcr_dat$`Object Type`),]
tcr_dat <- tcr_dat[!grepl("B", tcr_dat$Description),]
tcr_dat <- tcr_dat[!grepl("J", tcr_dat$Description),]
tcr_dat <- tcr_dat[!grepl("O", tcr_dat$Description),]
tcr_dat <- tcr_dat[!grepl("U", tcr_dat$Description),]
tcr_dat <- tcr_dat[!grepl("Z", tcr_dat$Description),]
```

```{r plot}
tcr_dat <- tcr_dat[, 2]
ggseqlogo(tcr_dat)
```
```

- Install and load the needed libraries
- Read in the data-file (we start with dataset of T-cells)
- Transform and filter your data
    - In this scenario we focus on linear peptides
    - And set maximum length of sequence to “10”
- Create a plot for being able to recognize binding patterns

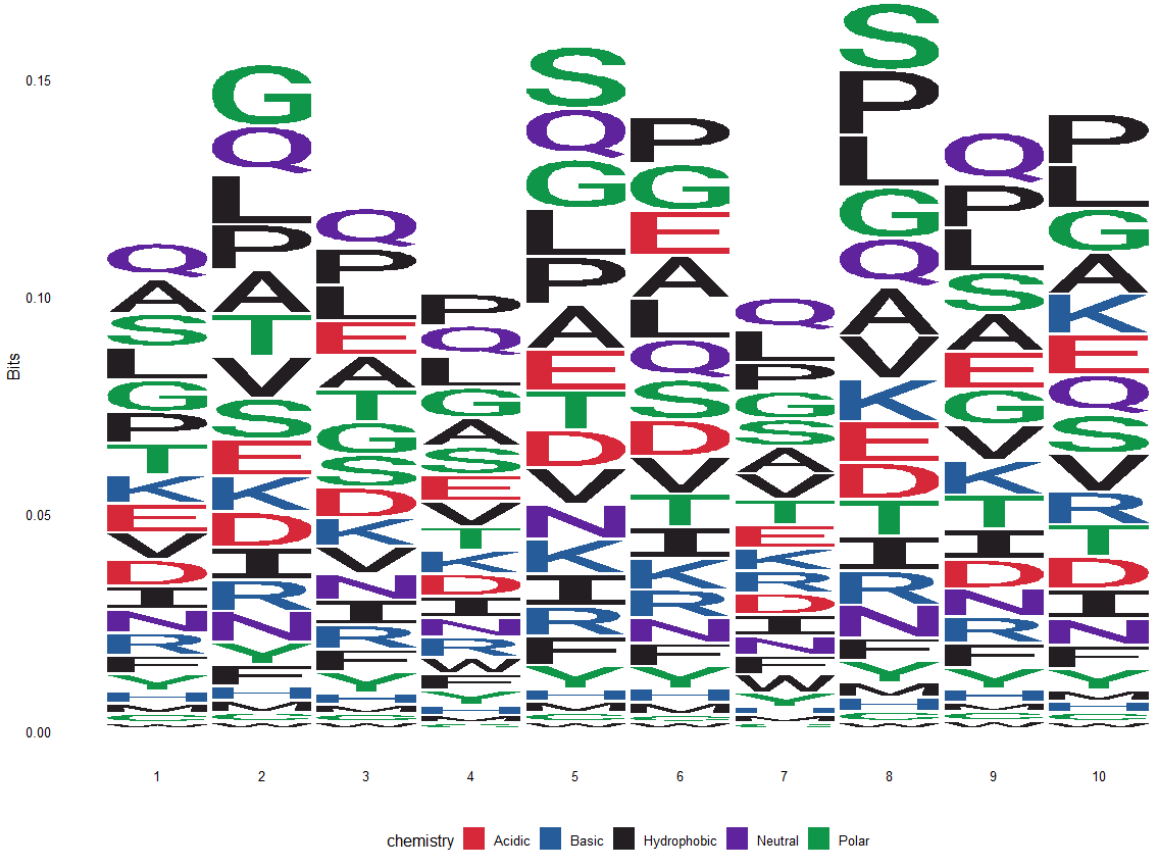


For further investigation print out:

- The count matrix per amino acid
- A frequency matrix per amino acid
- The bits of information matrix

```
> tcr_dat %>% pssm_counts %>% .[1:7,1:10]
     A    R    N    D   C   Q    E    G   H    I
1 1640 1562  787  681 374 634 1091 1287 533 1562
2 1030  504  356  262 181 807 3281  522 206  917
3 1822 1431 1075 1051 545 654  745  920 478 1711
4 1730 1246 1265 1370 416 636 1563 2212 414 1642
5 1360 1366 2194 1190 451 617  939 1248 503 2220
6 1391 1278 1297 1070 527 635  854 2547 465 1808
7 1229 1515 1205  892 436 520  855 1588 513 3002
> # derive the frequency matrix
> tcr_dat %>% pssm_freqs %>% .[1:7,1:10]
          A          R          N          D           C          Q          E          G          H          I
1 0.06551352 0.06239764 0.03143850 0.02720409 0.014940279 0.02532657 0.04358247 0.05141214 0.021291895 0.06239764
2 0.04114569 0.02013342 0.01422123 0.01046618 0.007230456 0.03223745 0.13106699 0.02085247 0.008229138 0.03663165
3 0.07278393 0.05716454 0.04294331 0.04198458 0.021771262 0.02612551 0.02976072 0.03675149 0.019094795 0.06834978
4 0.06910878 0.04977430 0.05053330 0.05472776 0.016618064 0.02540646 0.06243758 0.08836336 0.016538170 0.06559342
5 0.05432829 0.05456797 0.08764431 0.04753725 0.018016219 0.02464747 0.03751049 0.04985419 0.020093477 0.08868294
6 0.05556665 0.05105261 0.05181161 0.04274358 0.021052211 0.02536652 0.03411497 0.10174570 0.018575480 0.07222466
7 0.04909519 0.06052011 0.04813646 0.03563296 0.017417010 0.02077258 0.03415492 0.06343626 0.020492949 0.11992170
> # derive the bits of information matrix
> tcr_dat %>% pssm_freqs %>% pssm_bits %>% .[1:7,1:10]
           A           R           N           D           C           Q           E           G           H           I
1 0.013908380 0.013246884 0.006674326 0.005775370 0.003171789 0.005376776 0.009252465 0.010914686 0.004520224 0.01324688
2 0.022953226 0.011231481 0.007933348 0.005838588 0.004033528 0.017983741 0.073116053 0.011632606 0.004590645 0.02043506
3 0.011884284 0.009333925 0.007011858 0.006855314 0.003554849 0.004265819 0.004859381 0.006000846 0.003117831 0.01116027
4 0.010878460 0.007835007 0.007954481 0.008614734 0.002615861 0.003999249 0.009828343 0.013909337 0.002603285 0.01032510
5 0.011004323 0.011052871 0.017752562 0.009628783 0.003649228 0.004992402 0.007597838 0.010098085 0.004069981 0.01796294
6 0.009032511 0.008298742 0.008422119 0.006948086 0.003422095 0.004123397 0.005545482 0.016539041 0.003019495 0.01174032
7 0.010303245 0.012700908 0.010102042 0.007478026 0.003655179 0.004359387 0.007167839 0.013312899 0.004300703 0.02516708
>
```

Repeat the same scenario by creating a new dataframe for the B-cells-data.


0.15
0.10
0.05
0.00
Bits
1
2
3
4
5
6
7
8
9
10
chemistry
Acidic
Basic
Hydrophobic
Neutral
Polar