Mutations_explorev2

Adarsh

6/20/2020

Mutagenic data of spike protein

This is a notebook exploring the SARS-CoV-2 mutagenic data. The first step is to import the tidyverse library which will be used to conduct exploratory data analysis, and import the data.

```r
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.1     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ---------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
file<-read.csv("data.csv",header= TRUE, sep=",")
dim(file)
```

```
## [1] 2340    6
```

```r
head(file,N=6L)
```

```
##   Residue.. Substitution   Top_rep1   Top_rep2 Bottom_rep1 Bottom_rep2
## 1        19            A -0.1764999  0.2978451  -0.2039027 -0.24789819
## 2        19            C -0.9096144 -1.4284898   0.5559624  0.03971889
## 3        19            D -1.4911309 -1.2155133   0.9672447  1.27736869
## 4        19            E -1.0613240 -1.8116679   1.4735906  1.77240860
## 5        19            F  1.2298997  1.0344321  -1.0662619 -1.18149682
## 6        19            G -1.2820811 -1.4374131   0.9239352  1.21642268
```

```r
tail(file, N=-6L)
```

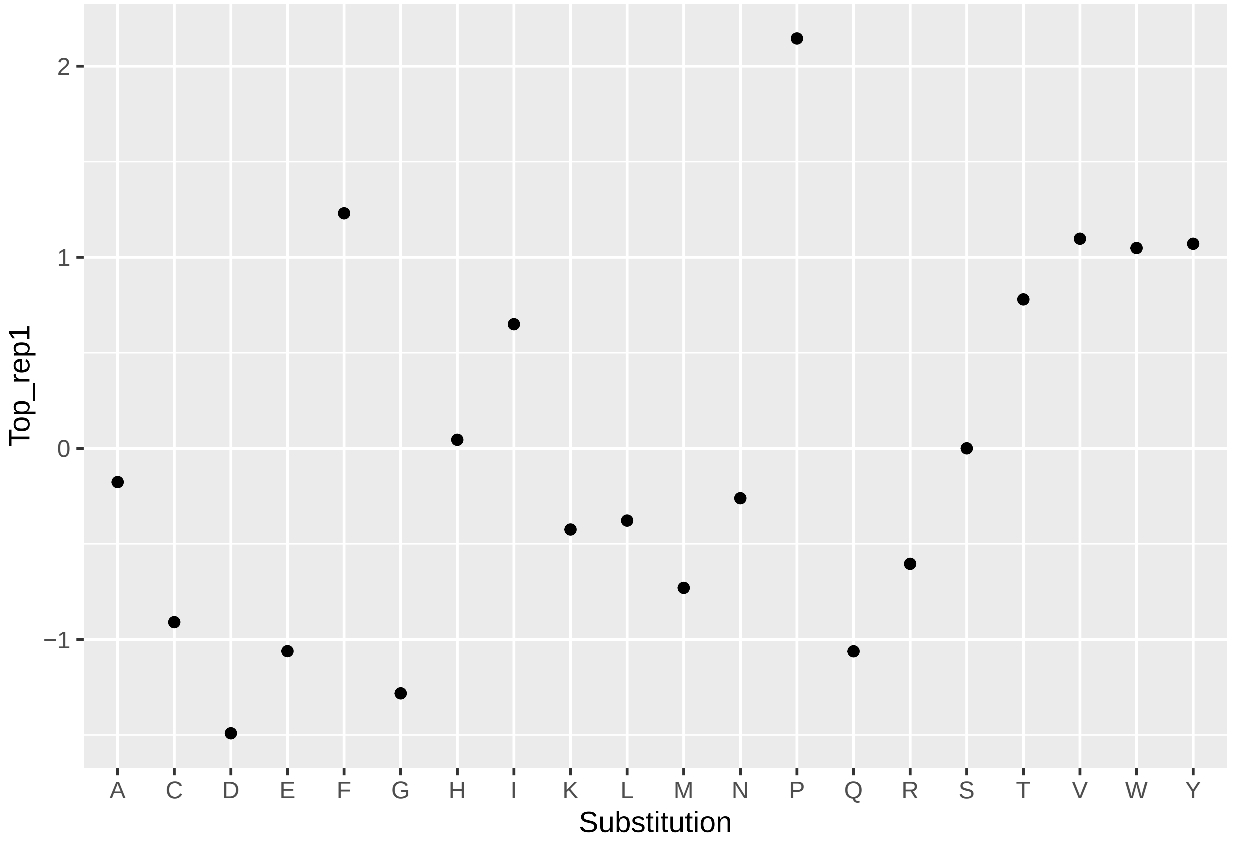
```
##      Residue.. Substitution    Top_rep1    Top_rep2 Bottom_rep1 Bottom_rep2
## 2335       518            R  0.00000000  0.00000000   0.0000000   0.0000000
## 2336       518            S  0.11209716  0.02784838  -0.3241399  -0.9474402
## 2337       518            T  0.09676146 -0.77840324  -0.6421838  -0.5398540
## 2338       518            V -0.48007793 -0.35046816  -0.5480825  -0.4105959
## 2339       518            W -0.14315194 -0.68674607  -0.4058743  -0.3200862
## 2340       518            Y -1.91622539 -0.05349618  -0.3532171  -0.9761237
```

```
fin_res=file[2340,"Residue.."]
```

Then, the data can be compartmentalized into each residue.Residue 19 has been used as an example

```
bool<-file[file$Residue..==19, ]
```

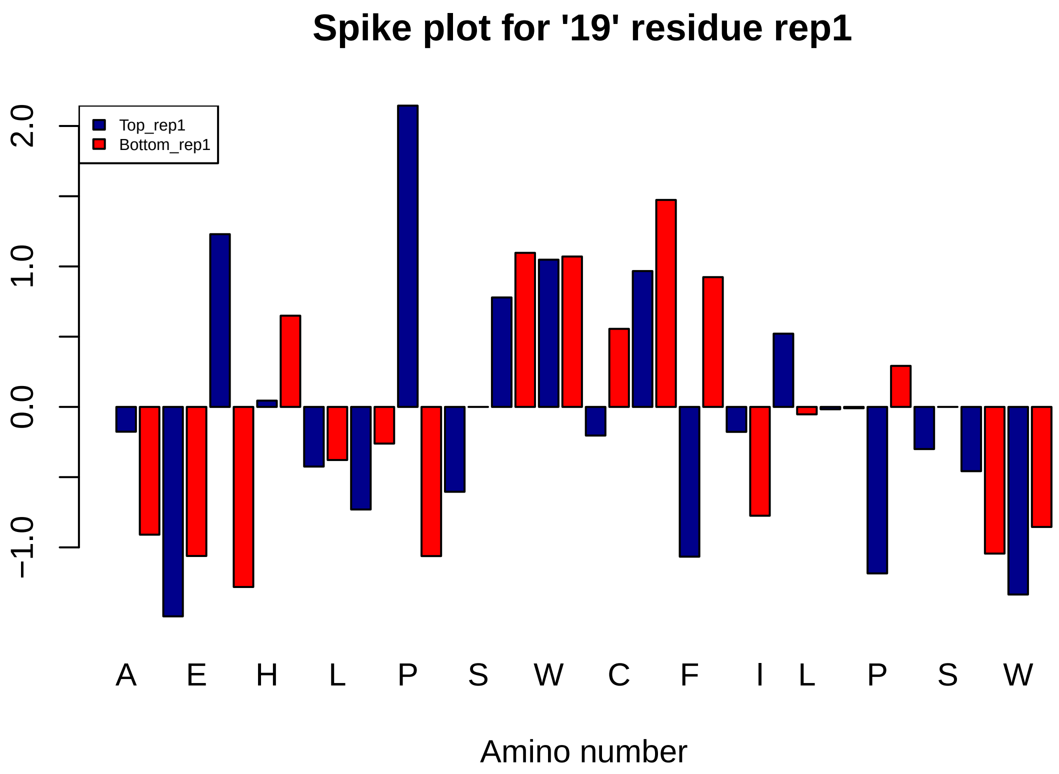
```
ggplot(data=bool)+geom_point(mapping=aes(x=Substitution,y=Top_rep1))
```



```
counts <- c(bool$Top_rep1,bool$Bottom_rep1)
x <- c(bool$Substitution, bool$Substitution)
regions <- c("Top_rep1", "Bottom_rep1")
col <- c("darkblue","red")

barplot(counts, names.arg=x, main="Spike plot for '19' residue rep1",
  xlab="Amino number", col=c("darkblue","red"),
  legend = rownames(counts), beside=TRUE)
```

```
legend("topleft", regions, cex = 0.5, fill = col)
```



Now, it is likely that the amino acid with the highest top_rep value and the least bottom_rep value is the mutagen with the highest affinity for the reactive binding domain.

```
tdiff1=abs(bool$Top_rep1-bool$Bottom_rep1)

mut1=bool$Substitution[which.max(tdiff1)]
```

```
#for
```