

Mutations_explore2

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.

```
library(tidyverse)
```

```
## -- Attaching packages -----
```

```
## 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 -----
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

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

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

```
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

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

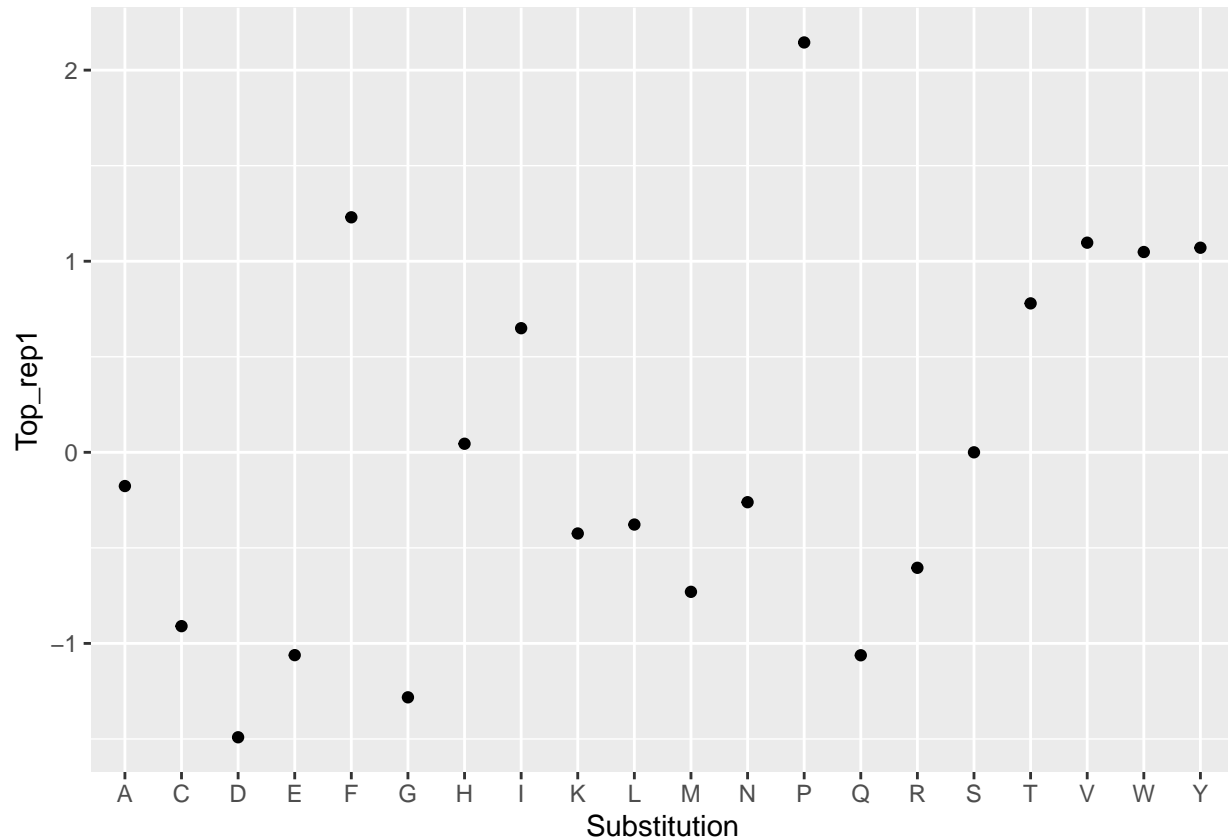
##	Residue..	Substitution	Top_rep1	Top_rep2	Bottom_rep1	Bottom_rep2
## 2335	518	R	0.00000000	0.00000000	0.00000000	0.00000000
## 2336	518	S	0.11209716	0.02784838	-0.3241399	-0.9474402
## 2337	518	T	0.09676146	-0.77840324	-0.6421838	-0.5398540
## 2338	518	V	-0.48007793	-0.35046816	-0.5480825	-0.4105959
## 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.."] # Taking the first residue
fin_start<-file[1,"Residue.."] # Taking the last residue
res_vector<-fin_start:fin_res # Formulating a residue vector
```

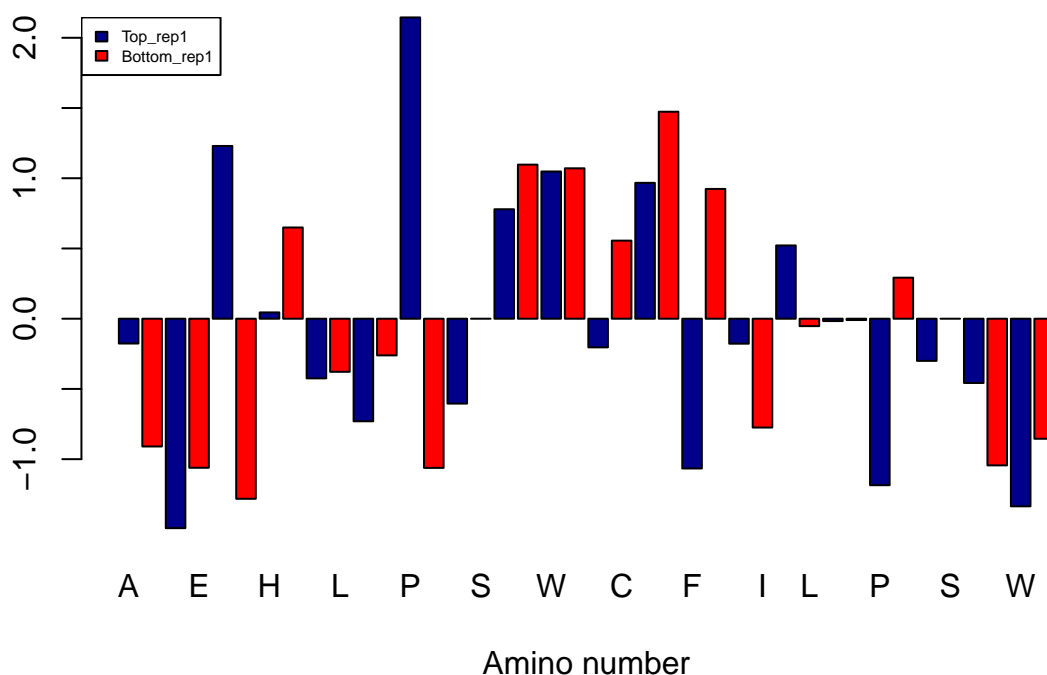
Then, the data can be compartmentalized into each residue. Residue 19 has been used as an example

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

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



Spike plot for '19' residue rep1



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 (amino acid p). Therefore, the difference between the top_rep value and the bottom_rep value will be highest for the amino acid that is most likely to act as the receptor (ie bind to the spike protein) instead of the wild type protein. This is a crude analysis but has been done as a first pass.

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

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

The above analysis can be repeated for all valid residues, and a vector of likely mutagens has been compiled for both experiment 1 (mut1) and experiment 2 (mut2). These can be compared with the original amino acid chain that binds to the spike protein.

```
tdiff1<-list()
tdiff2<-list()
mut1<-vector()
mut2<-vector()
original<-vector()
res<-vector()

for (num in res_vector) {
  bool2<-file[file$Residue==num, ]
  if (is.null(bool2)) {
    next
  }
}
```

```

tdif1=abs(bool2$Top_rep1-bool2$Bottom_rep1)
tdif2=abs(bool2$Top_rep2-bool2$Bottom_rep2)

original<-c(original, bool2[tdif1==0, "Substitution"])

tdiff1<-c(tdiff1,bool2$Residue..[1],abs(bool2$Top_rep1-bool2$Bottom_rep1))
tdiff2<-c(tdiff2,bool2$Residue..[1],tdif2)

mutt1<-bool2$Substitution[which.max(tdif1)]
mutt2<-bool2$Substitution[which.max(tdif2)]
if(length(mutt1) != 0)
{
  res<-c(res,num)
}

mut1<-c(mut1,mutt1)
mut2<-c(mut2,mutt2)

#print(str(num) + str('=') + str(mut1))
}

# print(res)
#
# print(mut1)
#
# print(mut2)
#
# print(original)

smoke <- matrix(c(res,original,mut1,mut2),ncol=4)
colnames(smoke) <- c("Residue","Original","Likely substitute (rep1)","Likely substitute (rep2)")
#rownames <- c("current","former","never")
smoke <- as.table(smoke)
print(smoke)

```

```

##      Residue Original Likely substitute (rep1) Likely substitute (rep2)
## A    19         S          P                               P
## B    20         T          W                               W
## C    21         I          W                               R
## D    22         E          N                               R
## E    23         E          F                               F
## F    24         Q          Y                               D
## G    25         A          V                               V
## H    26         K          D                               D
## I    27         T          L                               L
## J    28         F          Q                               K
## K    29         L          D                               K

```

## L 30	D	E	I
## M 31	K	W	W
## N 33	N	P	P
## O 34	H	A	A
## P 35	E	C	C
## Q 37	E	F	I
## R 38	D	K	P
## S 39	L	R	R
## T 40	F	R	D
## U 41	Y	I	I
## V 42	Q	C	C
## W 45	L	E	C
## X 46	A	P	P
## Y 48	W	Q	H
## Z 49	N	A	A
## A1 50	Y	K	P
## B1 51	N	P	P
## C1 54	I	H	D
## D1 56	E	K	I
## E1 57	E	N	I
## F1 59	V	P	P
## G1 60	Q	T	P
## H1 61	N	Q	E
## I1 62	M	E	P
## J1 63	N	P	P
## K1 64	N	E	P
## L1 65	A	W	W
## M1 67	D	P	P
## N1 68	K	P	P
## O1 69	W	V	V
## P1 71	A	P	P
## Q1 72	F	Y	Y
## R1 73	L	P	P
## S1 74	K	P	H
## T1 75	E	R	R
## U1 76	Q	V	M
## V1 79	L	W	T
## W1 82	M	C	C
## X1 83	Y	Q	E
## Y1 84	P	H	I
## Z1 89	Q	P	D
## A2 90	N	H	Q
## B2 91	L	P	P
## C2 92	T	H	H
## D2 93	V	H	P
## E2 95	L	K	K
## F2 96	Q	F	F
## G2 98	Q	D	P
## H2 102	Q	D	W
## I2 273	R	D	Y
## J2 274	F	D	Q
## K2 276	T	D	G
## L2 277	N	H	K
## M2 290	N	K	Y

##	N2	324	T	E	P
##	O2	325	Q	P	P
##	P2	326	G	P	H
##	Q2	329	E	Y	N
##	R2	330	N	Y	F
##	S2	343	V	P	G
##	T2	345	H	P	F
##	U2	346	P	M	W
##	V2	347	T	Q	Q
##	W2	349	W	N	D
##	X2	350	D	W	Q
##	Y2	351	L	T	F
##	Z2	352	G	K	M
##	A3	353	K	W	A
##	B3	354	G	Q	K
##	C3	355	D	K	K
##	D3	356	F	P	P
##	E3	357	R	A	Y
##	F3	366	M	Q	H
##	G3	367	D	T	F
##	H3	370	L	Y	D
##	I3	371	T	P	P
##	J3	374	H	P	C
##	K3	378	H	K	T
##	L3	382	D	I	I
##	M3	383	M	W	N
##	N3	386	A	L	L
##	O3	387	A	I	C
##	P3	388	Q	D	P
##	Q3	389	P	D	D
##	R3	390	F	E	E
##	S3	393	R	W	W
##	T3	394	N	P	P
##	U3	398	E	Q	M
##	V3	401	H	S	S
##	W3	402	E	F	S
##	X3	406	E	S	A
##	Y3	409	S	C	N
##	Z3	438	F	N	N
##	A4	441	K	S	Q
##	B4	442	Q	I	T
##	C4	445	T	Q	Q
##	D4	446	I	P	D
##	E4	504	F	N	H
##	F4	505	H	I	M
##	G4	509	D	I	K
##	H4	510	Y	L	K
##	I4	511	S	D	Q
##	J4	512	F	D	T
##	K4	514	R	D	P
##	L4	515	Y	T	I
##	M4	518	R	G	G