Mutation Hotspots of SARS Covid19 Variants

Approach: Firstly the gene data were collected from the SARS-CoV-2 interactive dashboard with *RefSeq: NC_045512.2*. This produced *sequences.fasta* with the data. Then, each of the 12 genes in the combined data was split into *FASTA* files of their own (namely: *envelope*, *membrane*, *nucleocapsid*, *ORF1a*, *ORF1ab*, *ORF3a*, *ORF6*, *ORF7a*, *ORF7b*, *ORF8*, *ORF10*, and *surface*). Following the creation of the gene data files, they were saved in a folder called *Genes*. In the same directory as the *Genes* folder, another folder was created named *Viruses*. Virus data was collected from the same dashboard list of *RefSeq: NC_045512.2* (namely: *OZ194575.1*, *OZ194709.1*, *OZ195329.1*, *OZ195368.1*, and *OZ196954.1*). Please note that the virus data files were selected from different pages of the *RefSeq: NC_045512.2* list to ensure randomness.

Next, a *Python* script (namely: *saha_dibbyo_assignment2_code.py*) was developed that compares each of the aforementioned genes in the reference genome against that in the corresponding virus variant. The *Python* script allowed a number of mismatches to be calculated as 1% of the gene's length ranging from (a minimum of 5 and a maximum of 20 mismatches). Mismatches were counted using the gene's index, like a mismatch at position 26 in the reference gene having '*C*' whereas the corresponding position in the gene of the variant had '*T*', was noted as *26C/T*. Finally, the data gets organized in a *CSV* file (namely: *saha_dibbyo_assignment2_spreadsheet.csv*). Each row produced entails a virus instance and the columns report the list of mutations for the respective virus instance as mentioned earlier.

Observations: The *CSV* file had the following data which was produced from running the *Python* code on the genes and viruses data as mentioned in the above section:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Virus	envelope	ORF7a	ORF8	ORF3a	nucleocapsid	ORF7b	ORF1a	ORF10	surface	ORF1ab	ORF6	membrane
2	OZ195329.1	['C26T']	[]	['G229A']	['C192T', 'G328T', 'C668T']	[]	['C52T', 'T55C']	[]	[]	[]	[]	['A58C', 'G181C', 'A182T', 'T183C']	[]
3	OZ196954.1	['C26T']	[]	['A28T']	['C192T', 'C232T', 'C668T']	[]	['C52T', 'T55C']	[]	[]	[]	[]	['A58C', 'G181C', 'A182T', 'T183C']	[]
4	OZ194709.1	['C26T', 'G139N']	[]	[]	[]	[]	['C52T', 'T55C']	[]	[]	[]	[]	['A58C', 'G181C', 'A182T', 'T183C']	[]
5	OZ194575.1	['C26T']	[]	[]	['C192T', 'C668T']	[]	['C52T', 'T55C']	[]	[]	[]	[]	['A58C', 'G181C', 'A182T', 'T183C']	[]
6	OZ195368.1	['C26T']	[]	[]	['C192T', 'C668T']	[]	['C52T', 'T55C']	[]	[]	[]	[]	['A58C', 'C105T', 'G181C', 'A182T', 'T183C']	[]
7													

Figure 1*: Image of the data in the CSV file produced after analysis of mutation hotspots of SARS Covid19 variants.*

As can be noted from *Figure 1*, the *envelope* gene had a mutation, *C26T*, for all the variants as well as mutation *G139N*, for *OZ194709.1*. But, the *ORF8* gene had mutation *G229A* in *OZ195329.1* and mutation *A185T* in *OZ196954.1*. On the other hand, the *ORF3a* had mutations *C192T* and *C668T* for *OZ195329.1*, *OZ196954.1*, *OZ194575.1*, and *OZ195368.1,* and mutations *G328T* and *C232T* for *OZ195329.1* and *OZ196954.1 respectively.* For the gene *ORF7b*, there were mutations *C52T* and *T55C* across all variants. And finally, *ORF6* had mutations *A58C*, *G181C*, *A182T*, and *T183C* across all variants with an additional mutation *C105T* for variant *OZ195368.1*.