**Progetto3 (tre studenti)**

Si richiede di prendere in input il file `covid-sequences.fasta` contenente genomi di SARS-CoV-2 sequenziati nel novembre 2021 e scaricati dal sito di NCBI: (https://www.ncbi.nlm.nih.gov/sars-cov-2/). Il primo, con identificatore NC_045512.2, è il genoma di riferimento sequenziato nell'autunno 2019. Usare MAFFT (https://www.ebi.ac.uk/Tools/msa/mafft/) per allinearli e utilizzare la matrice di allineamento multiplo ottenuta per trovare in seguito tutte le variazioni puntuali dei genomi rispetto al riferimento. Non considerare gli eventuali gaps iniziali/finali.
Esempio per tre genomi G_REF (*reference*) G1 e G2:

```
G_REF     AAGCTGATTGCACGC-TCG
G1        --GCAGAGTG-ACGCCT--
G2        --GCCGAGTGCACGCCT--
```

Variazioni di G1:

- Posizione 5: sostituzione T→A
- Posizione 8: sostituzione T→G
- Posizione 11: inserimento di C
- Posizione 16: cancellazione di C

Variazioni di G2:

- Posizione 5: sostituzione T→C
- Posizione 8: sostituzione T→G
- Posizione 16: cancellazione di C

Implementare uno script Python (non un notebook) che produca un report di tutte le variazioni puntuali rilevate rispetto al *reference,* ciascuna riferita alla posizione nel *reference* in cui occorre. Per ognuna di esse specificare il tipo (sostituzione, inserimento nel *reference*, cancellazione nel *reference*), le basi coinvolte (o la base inserita/cancellata) e il numero di genomi che presentano la variazione rispetto al *reference*. Ignorare le sostituzioni che coinvolgono la base *unknown* N (gli inserimenti/cancellazioni che coinvolgono N sono invece da considerare). Il report deve anche riportare:

- il genoma con più variazioni e quello con meno variazioni rispetto al *reference*.
- le posizioni del *reference* rispetto a cui tutti gli altri genomi variano
- le posizioni del *reference* rispetto a cui tutti gli altri genomi variano allo stesso modo