

# Multiple Sequence Alignment (MSA) di sequenze SARS-CoV-2

EDOARDO SILVA 816560  
DAVIDE MARCHETTI 815990

A.A.: 2019/2020

## 1 Abstract

La seconda parte del progetto prevede di elaborare i file prodotti in precedenza ricavando informazioni relative alle alterazioni rilevate e producendo in output una tabella riassuntiva contenente:

- il gene id del gene in cui cade la variazione con lo start e l'end della sua CDS rispetto alla reference
- il codone (o i codoni) alterato della reference, con posizione di inizio rispetto alla CDS, sequenza del codone e amminoacido codificato
- il nuovo codone generato dalla variazione (o i nuovi codoni generati) specificando la sequenza del codone e il nuovo amminoacido codificato

## 2 Algoritmo

L'algoritmo inizia caricando tutti i file necessari per l'elaborazione, in particolare quelli prodotti in output nella parte precedente del progetto:

1. Caricamento della sequenza reference dal file corrispondente memorizzato in `/project-1/input/reference.fasta`.
2. Caricamento di uno dei file di output prodotti nella prima parte di progetto. Nel nostro caso è stata utilizzata l'analisi dell'allineamento di `ClustalW`.
3. Lettura del file `Genes-CDS.xlsx` contenente le informazioni sui geni e le CDS della sequenza di reference. In particolare, per le CDS che derivano dalla join di due sequenze è possibile specificare il punto di unione della sequenza.

Dopo la lettura del materiale rilevante a questa fase di elaborazione, l'algoritmo itera le variazioni rilevate nell'allineamento e per ciascuna di esse esegue i seguenti step:

1. Trova le CDS nelle quali avviene l'alterazione rispetto alla reference.
2. Recupera le informazioni del gene associato alle CDS rilevate calcolando le posizioni globali e relative alla CDS dell'alterazione.
3. Identifica i codoni alterati e ne effettua la ritraduzione in amminoacidi grazie ad una look-up table (listato 1). Vengono ignorate le alterazioni che presentano sequenze di soli -, derivate probabilmente da un sequenziamento errato o un'alterazione posta ai capi dell'allineamento.
4. Memorizza tutte le informazioni ricavate in una struttura dati tramite cui derivare la tabella per l'output finale associando i valori a chiavi prestabilite (listato 2).

Al termine dell'elaborazione di tutte le alterazioni, viene costruito un oggetto di tipo `DataFrame` fornito dalla libreria `pandas`.

Le chiavi utilizzate nella costruzione della struttura dati a lista diventeranno le colonne del `DataFrame`. Questo sarà esportato in CSV nella cartella `/project-2/output/alteration-table.csv` per permettere una visualizzazione più semplice tramite programmi terzi (come riportato in fig. 1)

### 3 Informazioni memorizzate

Ad ogni variazione analizzata corrisponde un'entrata nella struttura dati a lista contenente le seguenti informazioni:

- **gene\_id**: id del gene in cui cade la variazione
- **gene\_start**: inizio del gene in cui cade la variazione (1-based)
- **gene\_end**: fine del gene in cui cade la variazione (1-based)
- **cds\_start**: inizio della **Coding DNA Sequence** della porzione del gene in cui cade la variazione (1-based)
- **cds\_end**: fine della **Coding DNA Sequence** della porzione del gene in cui cade la variazione (1-based)
- **relative\_start**: inizio della variazione in rispetto all'inizio della cds (1-based)
- **relative\_end**: fine della variazione in rispetto all'inizio della cds (1-based)
- **alteration**: sequenza della variazione
- **original\_codone**: codone della reference prima della modifica
- **original\_aminoacid**: amminoacido codificato da **original\_codone**
- **altered\_codone**: codone della reference modificati dalla variazione
- **encoded\_aminoacid**: amminoacido codificato da **altered\_codone**

## 4 Output

Come riportato in fig. 1 la maggior parte delle alterazioni coinvolgono un singolo codone e quelli ottenuti rimangono traducibili.

In alcuni casi, l'amminoacido risultante dalla traduzione dell'alterazione non viene modificato. La maggior parte delle variazioni si concentrano nel gene **ORF1ab** identificato da **gene\_id = 43740578**.

Le ultime righe della tabella riportano delle alterazioni che determinano la cancellazione di alcune basi rispetto alla sequenza reference. Queste sono relative solo alla sequenza **MT262993.1** e si pensa possano derivare da un errore in fase di sequenziamento.

gene_id	gene_start	gene_end	cds_start	cds_end	relative_start	relative_end	alteration	original_codone	original_aminoacid	altered_codone	encoded_aminoacid
43740578	266	21555	266	13483	1132	1132	A	AGT	S	AAT	N
43740578	266	21555	266	21555	1132	1132	A	AGT	S	AAT	N
43740578	266	21555	266	13483	10818	10818	T	GTA	V	TTA	L
43740578	266	21555	266	21555	10818	10818	T	GTA	V	TTA	L
43740578	266	21555	266	21555	18112	18112	T	TAC	Y	TTC	F
43740575	28274	29533	28274	29533	1101	1101	A	GCC	A	ACC	T
43740578	266	21555	266	13483	794	794	T	CAC	H	CAT	H
43740578	266	21555	266	21555	794	794	T	CAC	H	CAT	H
43740578	266	21555	266	13483	2772	2772	T	CTA	L	TTA	L
43740578	266	21555	266	21555	2772	2772	T	CTA	L	TTA	L
43740578	266	21555	266	21555	14143	14143	T	ACC	I	ATC	I
43740578	266	21555	266	13483	3638	3638	T	GCC	A	GCT	A
43740578	266	21555	266	21555	3638	3638	T	GCC	A	GCT	A
43740578	266	21555	266	21555	9249	9249	G	ACT	T	GCT	A
43740578	266	21555	266	21555	9249	9249	G	ACT	T	GCT	A
43740578	266	21555	266	13483	13111	13111	G	TAC	Y	TGC	C
43740578	266	21555	266	21555	13111	13111	G	TAC	Y	TGC	C
43740578	266	21555	266	13483	13211	13211	T	TGC	C	TGT	C
43740578	266	21555	266	21555	13211	13211	T	TTC	L	TTT	F
43740578	266	21555	266	21555	19219	19219	T	TGC	C	TTC	F
43740571	26523	27191	26523	27191	198	198	C	GCT	A	CCT	P
43740575	28274	29533	28274	29533	415	415	C	CTT	L	CCT	P
43740575	28274	29533	28274	29533	562	562	C	ATC	I	ACC	T
43740578	266	21555	266	13483	8442	8442	C	TGG	W	CGG	R
43740578	266	21555	266	21555	8442	8442	C	TGG	W	CGG	R
43740578	266	21555	266	21555	20622	20622	A	AGG	R	AGG	R
43740568	21563	25384	21563	25384	65	65	T	AAC	N	AAT	N
43740568	21563	25384	21563	25384	1173	1173	T	CTT	L	TTT	F
43740575	28274	29533	28274	29533	557	557	T	TTC	F	TTT	F
43740578	266	21555	266	13483	48	48	T	CAG	Q	TAG	STOP
43740578	266	21555	266	21555	48	48	T	CAG	Q	TAG	STOP
43740568	21563	25384	21563	25384	1841	1841	G	GGA	G	GGG	G
43740575	28274	29533	28274	29533	608	610	AAC	TAGGGG	STOPG	TAAACG	STOPT
43740578	266	21555	266	13483	8517	8517	T	CCA	P	TCA	S
43740578	266	21555	266	21555	8517	8517	T	CCA	P	TCA	S
43740568	21563	25384	21563	25384	906	906	T	GTT	V	TTT	F
43740568	21563	25384	21563	25384	3752	3752	T	TGG	W	TGT	C
43740577	27894	28259	27894	28259	251	251	C	TTT	F	TTC	F
43740575	28274	29533	28274	29533	605	605	A	CAG	Q	CAA	Q
43740578	266	21555	266	13483	619	619	T	ACG	T	ATG	START
43740578	266	21555	266	21555	619	619	T	ACG	T	ATG	START
43740578	266	21555	266	13483	1083	1083	T	CCA	P	TCA	S
43740578	266	21555	266	21555	1083	1083	T	CCA	P	TCA	S
43740578	266	21555	266	13483	8894	8894	T	TCC	S	TCT	S
43740578	266	21555	266	21555	8894	8894	T	TCC	S	TCT	S
43740568	21563	25384	21563	25384	2314	2314	T	TGT	C	TAT	Y
43740576	29558	29674	29558	29674	6	6	T	CTA	L	TTA	L
43740578	266	21555	266	13483	9456	9470	-----	TTTCTATTGGTCTT	FLVL	-----	
43740578	266	21555	266	21555	9456	9470	-----	TTTCTATTGGTCTT	FLVL	-----	
43740578	266	21555	266	21555	19251	19271	-----	ATTGTATCTGATGCTTATAA	IVSRCLSTOP	-----	

Figura 1: Tabella di output delle alterazioni

## 5 Listati di codice

**Code Listing 1:** Tabella per la traduzione in amminoacidi

```
1 aminoacids_lookup_table = {
2     'START': 'ATG',
3     'STOP': ['TAA', 'TAG', 'TGA'],
4     'F': ['TTT', 'TTC'],
5     'L': ['TTA', 'TTG', 'CTT', 'CTA', 'CTC', 'CTG'],
6     'I': ['ATT', 'ATC', 'ATA'],
7     'M': ['ATG'],
8     'V': ['GTT', 'GTA', 'GTC', 'GTG'],
9     'S': ['TCT', 'TCA', 'TCC', 'TCG', 'AGT', 'AGC'],
10    'P': ['CCT', 'CCA', 'CCC', 'CCG'],
11    'T': ['ACT', 'ACA', 'ACC', 'ACG'],
12    'A': ['GCT', 'GCA', 'GCC', 'GCG'],
13    'Y': ['TAT', 'TAC'],
14    'H': ['CAT', 'CAC'],
15    'Q': ['CAA', 'CAG'],
16    'N': ['AAT', 'AAC'],
17    'K': ['AAA', 'AAG'],
18    'D': ['GAT', 'GAC'],
19    'E': ['GAA', 'GAG'],
20    'C': ['TGT', 'TGC'],
21    'W': ['TGG'],
22    'R': ['CGT', 'CGA', 'CGC', 'CGG', 'AGA', 'AGG'],
23    'G': ['GGT', 'GGA', 'GGC', 'GGG']
24 }
```

**Code Listing 2:** Memorizzazione dei risultati nella struttura dati a lista

```
1  for key, value in variations:
2      for index, cds in affected_cdses.iterrows():
3          ...
4          variations_to_genes.append({
5              'gene_id': gene_id,
6              'gene_start': gene_start + 1, # 1-based position
7              'gene_end': gene_end,
8              'cds_start': cds_start + 1, # 1-based position
9              'cds_end': cds_end,
10             'original_codone': original_codone,
11             'altered_codone': altered_codone,
12             'relative_start': relative_start + 1, # 1-based position
13             'relative_end': relative_end,
14             'alteration': sequence,
15             'original_aminoacid': original_aminoacid,
16             'encoded_aminoacid': encoded_aminoacid
17         })
```

## 6 Divisone del lavoro

Durante la realizzazione del progetto entrambi i componenti del gruppo hanno partecipato attivamente alla sua realizzazione. In particolare:

- **Edoardo Silva** si è occupato principalmente di recuperare e gestire l'output JSON del progetto1 e delle funzioni di supporto.
- **Davide Marchetti** si è occupato principalmente di generare i file di output e correggere le porzioni di codice relative alle letture delle reference.
- Entrambi hanno lavorato alla creazione ed elaborazione dei dati, alla matrice delle mutazioni e le traduzioni di quest'ultime.