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 codifcato
- il nuovo codone generato dalla variazione (o i nuovi codoni generati) specifcando la sequenza del codone e il nuovo amminoacido codifcato

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 stato 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 costuzione 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 cacellazione 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	267	21555	267	13483	1131	1131	A	GTA	V	ATA	I
43740578	267	21555	267	21555	1131	1131	A	GTA	V	ATA	I
43740578	267	21555	267	13483	10817	10817	T	TTG	L	TTT	F
43740578	267	21555	267	21555	10817	10817	T	TTG	L	TTT	F
43740578	267	21555	267	21555	18111	18111	T	ACA	T	TCA	S
43740575	28275	29533	28275	29533	1100	1100	A	GAG	E	GAA	E
43740578	267	21555	267	13483	793	793	T	ACC	T	ATC	I
43740578	267	21555	267	21555	793	793	T	ACC	T	ATC	I
43740578	267	21555	267	13483	2771	2771	T	TTC	F	TTT	F
43740578	267	21555	267	21555	2771	2771	T	TTC	F	TTT	F
43740578	267	21555	267	21555	14142	14142	T	CCT	P	TCT	S
43740578	267	21555	267	13483	3637	3637	T	CCA	P	CTA	L
43740578	267	21555	267	21555	3637	3637	T	CCA	P	CTA	L
43740578	267	21555	267	13483	9248	9248	G	TTA	L	TTG	L
43740578	267	21555	267	21555	9248	9248	G	TTA	L	TTG	L
43740578	267	21555	267	13483	13110	13110	G	ACC	T	GCC	A
43740578	267	21555	267	21555	13110	13110	G	ACC	T	GCC	A
43740578	267	21555	267	13483	13210	13210	T	GCG	A	GTG	V
43740578	267	21555	267	21555	13210	13210	T	TGC	C	TTC	F
43740578	267	21555	267	21555	19218	19218	T	GCT	A	TCT	S
43740571	26524	27191	26524	27191	197	197	C	GTG	V	GTC	V
43740575	28275	29533	28275	29533	414	414	C	TTG	L	CTG	L
43740575	28275	29533	28275	29533	561	561	C	TCA	S	CCA	P
43740578	267	21555	267	13483	8441	8441	C	GGT	G	GGC	G
43740578	267	21555	267	21555	8441	8441	C	GGT	G	GGC	G
43740578	267	21555	267	21555	20621	20621	A	AAA	K	AAA	K
43740568	21564	25384	21564	25384	64	64	T	ACT	T	ATT	I
43740568	21564	25384	21564	25384	1172	1172	T	TGC	C	TGT	C
43740575	28275	29533	28275	29533	556	556	T	TCC	S	TTC	F
43740578	267	21555	267	13483	47	47	T	CTC	L	CTT	L
43740578	267	21555	267	21555	47	47	T	CTC	L	CTT	L
43740568	21564	25384	21564	25384	1840	1840	G	GAT	D	GGT	G
43740575	28275	29533	28275	29533	607	609	AAC	AGGGGA	RG	AAACGA	KR
43740578	267	21555	267	13483	8516	8516	T	AGC	S	AGT	S
43740578	267	21555	267	21555	8516	8516	T	AGC	S	AGT	S
43740568	21564	25384	21564	25384	905	905	T	ACG	T	ACT	T
43740568	21564	25384	21564	25384	3751	3751	T	GGA	G	GTA	V
43740577	27895	28259	27895	28259	250	250	C	TTA	L	TCA	S
43740575	28275	29533	28275	29533	604	604	A	AGT	S	AAT	N
43740578	267	21555	267	13483	618	618	T	CGT	R	TGT	C
43740578	267	21555	267	21555	618	618	T	CGT	R	TGT	C
43740578	267	21555	267	13483	1082	1082	T	CCC	P	CCT	P
43740578	267	21555	267	21555	1082	1082	T	CCC	P	CCT	P
43740578	267	21555	267	13483	8893	8893	T	CCT	P	CTT	L
43740578	267	21555	267	21555	8893	8893	T	CCT	P	CTT	L
43740568	21564	25384	21564	25384	2313	2313	A	GTT	V	ATT	I
43740576	29559	29674	29559	29674	5	5	T	GGC	G	GGT	G
43740578	267	21555	267	13483	9455	9469	----------	CATTTCTATTGGTTCTTT	HFYWFF	CA----------T	
43740578	267	21555	267	21555	9455	9469	----------	CATTTCTATTGGTTCTTT	HFYWFF	CA----------T	
43740578	267	21555	267	21555	19250	19270	----------	AGATTGTATCTCGATGCTTATAAC	RLYLDAYN	AG----------C	

Figura 1: Tabella di output delle alterazioni

5 Listati di codice

Code Listing 1: Tabella per la traduzione in amminoacidi

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

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

```
for key, value in variations:
  for index, cds in affected_cdses.iterrows():
    ...
    variations_to_genes.append({
      'gene_id': gene_id,
      'gene_start': gene_start + 1, # 1-based position
      'gene_end': gene_end,
      'cds_start': cds_start + 1, # 1-based position
      'cds_end': cds_end,
      'original_codone': original_codone,
      'altered_codone': altered_codone,
      'relative_start': relative_start + 1, # 1-based position
      'relative_end': relative_end,
      'alteration': sequence,
      'original_aminoacid': original_aminoacid,
      'encoded_aminoacid': encoded_aminoacid
    })
```

6 Analisi dei risultati e conclusioni

Sulla base degli allineamenti prodotti ed analizzati in questa prima parte, è stato prodotto uni grafico riassuntivo per fornire un idea generale delle sequenze prese in esame.

6.1 Geni coinvolti

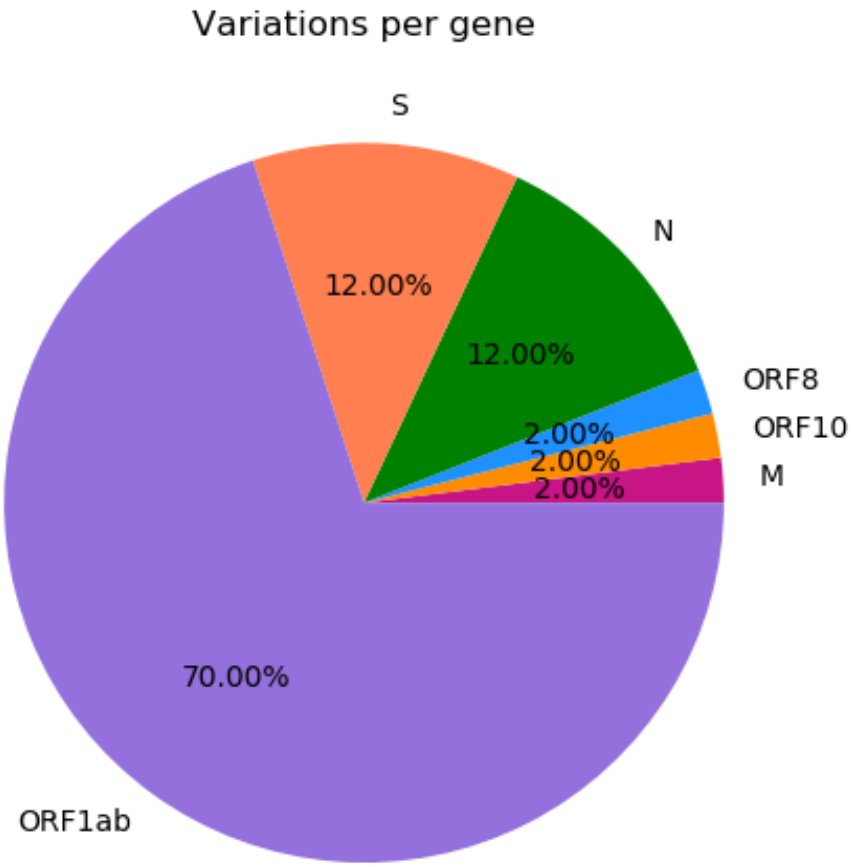

Figura 2: Tipologia di variazioni

Si può notare che quasi tutte le modifiche (70%) si concentrano sul primo gene della CDR: `gene\_name= ORF1ab`; seguito dai geni `gene\_name= S` e `gene` ↪ `\_name= N` (12% ciascuno) e che gli altri siano quasi invariati. Un'analisi finale si trova nella terza e ultima parte del progetto.

7 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.