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 Espressione genica

I passi fondamentali per capire come un gene esprime una particolare proteina prevedono:

1. Trascizione: sostituzione della Timina (T) con Uracile (U), ottenendo pre-mRNA
2. Splicing: eliminazione delle sequenze introniche e concatenazione degli estroni per ottenere mRNA (oppure trascritto)
3. Traduzione: traduzione in proteine dei **codoni** (triplette di basi) della CDS (sottostringa dellmRNA)

Risulta triviale come un'alterazione della sequenza di DNA iniziale possa protrarsi fino alla fase di traduzione, andando ad alterare la produzione delle proteine di un particolare gene.

Attraverso le informazioni raccolte in questa fase saremo in grado di identificare in quali geni si concentrano le variazioni rilevate, dove queste avvengano e in che modo alterino i codoni e le rispettive proteine codificate.

3 Formato di output

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`

gene_id	gene_start	gene_end	cds_start	cds_end	relative_start	relative_end	alteration	original_codone	altered_codone	original_aminoacid	encoded_aminoacid
M	26523	27191	26523	27191	197	197	C	GUG	GUC	V	V
N	28274	29533	28274	29533	1100	1100	A	GAG	GAA	E	E
	28274	29533	28274	29533	414	414	C	UUG	CUG	L	L
	28274	29533	28274	29533	561	561	C	UCA	CCA	S	P
	28274	29533	28274	29533	556	556	U	UCC	UUC	S	F
	28274	29533	28274	29533	607	609	AAC	AGGGGA	AAACGA	RG	KR
	28274	29533	28274	29533	604	604	A	AGU	AAU	S	N
ORF10	29558	29674	29558	29674	4	5	U	GGC	GGU	G	G
ORF1ab	266	21555	266	13483	1131	1131	A	GUA	AUA	V	I
	266	21555	266	21555	1131	1131	A	GUA	AUA	V	I
	266	21555	266	13483	10817	10817	U	UUG	UUU	L	F
	266	21555	266	21555	10817	10817	U	UUG	UUU	L	F
	266	21555	266	21555	18111	18111	U	ACA	UCA	T	S
	266	21555	266	13483	793	793	U	ACC	AUC	T	I
	266	21555	266	21555	793	793	U	ACC	AUC	T	I
	266	21555	266	13483	2771	2771	U	UUC	UUU	F	F
	266	21555	266	21555	2771	2771	U	UUC	UUU	F	F
	266	21555	266	21555	14142	14142	U	CCU	UCU	P	S
	266	21555	266	13483	3637	3637	U	CCA	CUA	P	L
	266	21555	266	21555	3637	3637	U	CCA	CUA	P	L
	266	21555	266	13483	9248	9248	G	UUA	UUG	L	L
	266	21555	266	21555	9248	9248	G	UUA	UUG	L	L
	266	21555	266	13483	13110	13110	G	ACC	GCC	T	A
	266	21555	266	21555	13110	13110	G	ACC	GCC	T	A
	266	21555	266	13483	13210	13210	U	GCG	GUG	A	V
	266	21555	266	21555	13210	13210	U	UGC	UUC	C	F
	266	21555	266	21555	19218	19218	U	GCU	UCU	A	S
	266	21555	266	13483	8441	8441	C	GGU	GGC	G	G
	266	21555	266	21555	8441	8441	C	GGU	GGC	G	G
	266	21555	266	21555	20621	20621	A	AAA	AAA	K	K
	266	21555	266	13483	47	47	U	CUC	CUU	L	L
	266	21555	266	21555	47	47	U	CUC	CUU	L	L
	266	21555	266	13483	8516	8516	U	AGC	AGU	S	S
	266	21555	266	21555	8516	8516	U	AGC	AGU	S	S
	266	21555	266	13483	618	618	U	CGU	UGU	R	C
	266	21555	266	21555	618	618	U	CGU	UGU	R	C
	266	21555	266	13483	1082	1082	U	CCC	CCU	P	P
	266	21555	266	21555	1082	1082	U	CCC	CCU	P	P
	266	21555	266	13483	8893	8893	U	CCU	CUU	P	L
	266	21555	266	21555	8893	8893	U	CCU	CUU	P	L
	266	21555	266	13483	9455	9469	----------	CAUUUCUAUUGGUUCUUU	CA----------U	HFYWFF	
	266	21555	266	21555	9455	9469	----------	CAUUUCUAUUGGUUCUUU	CA----------U	HFYWFF	
	266	21555	266	21555	19250	19270	----------	AGAUUGUAUCUCGAUGCUUAUAAC	AG----------C	RLYLDAYN	
ORF8	27894	28259	27894	28259	250	250	C	UUA	UCA	L	S
S	21563	25384	21563	25384	64	64	U	ACU	AUU	T	I
	21563	25384	21563	25384	1172	1172	U	UGC	UGU	C	C
	21563	25384	21563	25384	1840	1840	G	GAU	GGU	D	G
	21563	25384	21563	25384	905	905	U	ACG	ACU	T	T
	21563	25384	21563	25384	3751	3751	U	GGA	GUA	G	V
	21563	25384	21563	25384	2313	2313	A	GUU	AUU	V	I

Figura 1: Tabella di output delle alterazioni

4 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 fasta della sequenza di Wuhan `NC_045512.2`.

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 contenente le informazioni sui geni e le CDS della sequenza di riferimento. In particolare, una delle CDS analizzate derivava dall'unione (join) di due sequenze. In tal caso 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. Identifica le CDS nelle quali avviene l'alterazione rispetto alla sequenza reference.

2. Effettua una sostituzione della `Timina` (`T`) con l'`Uracile` (`U`) nelle sezioni coinvolte delle sequenze.

3. Recupera le informazioni del gene associato alle CDS rilevate calcolando le posizioni globali e relative alla CDS dell'alterazione.

4. 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.

5. Memorizza tutte le informazioni ricavate in una struttura dati apposita tramite cui derivare la tabella per l'output finale associando i valori a chiavi prestabilite.

La struttura dati così ottenuta sarà esportata in formato CSV per permettere una visualizzazione agevolata attraverso programmi terzi (come riportato in fig. 1).

5 Analisi dei risultati e conclusioni

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.

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

5.1 Distribuzione delle variazioni all'interno delle CDS

Rispetto alle variazioni identificate ed analizzate, più del 65% risultano appartenenti ad almeno una CDS e riportate in tabella.

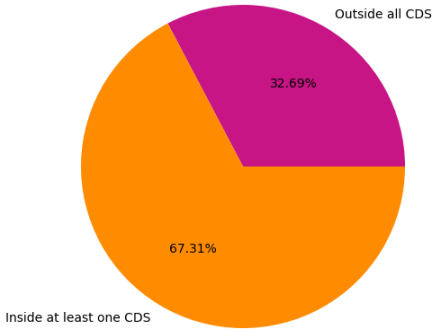

Figura 2: Percentuale di variazioni interne ed esterne alle CDS

Analizzando la distribuzione delle alterazioni che coinvolgono geni riportata in fig. 3, quasi i tre quarti del numero totale di variazioni si concentrano nel gene `ORF1ab`. Questo era un risultato atteso, essendo un gene composto da più di 21.000 basi.

Infine, dal grafico delle alterazioni appartenenti ad una CDS divise per sequenza (fig. 4) notiamo una distribuzione piuttosto omogenea, eccetto per le sequenze `MT262993.1` e `MT276597.1` dove abbiamo un basso numero di alterazioni. Al contrario, la sequenza nella quale si manifestano più alterazioni nelle CDS risulta essere `EPI_ISL_437334`, che anche nelle analisi precedenti

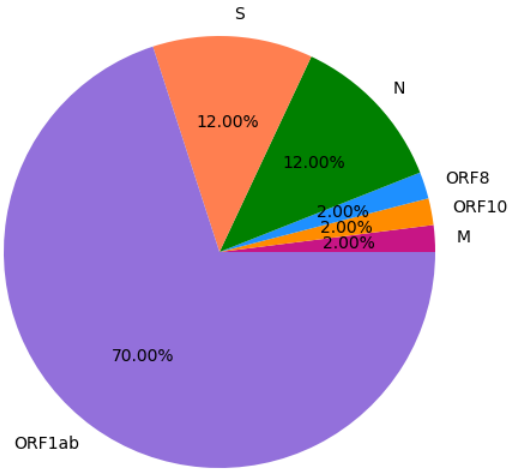

Figura 3: Percentuale di variazioni per gene

risultava avere il maggior numero di sostituzioni rispetto alle altre sequenze.

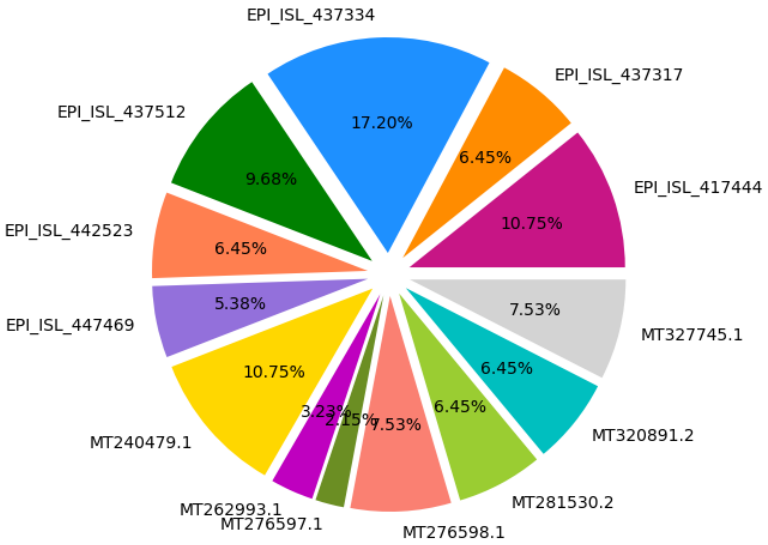

Figura 4: Tipologia di variazioni

5.2 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.

6 Listati di codice

Code Listing 1: Tabella per la traduzione in amminoacidi

```
aminoacids_lookup_table = {
  'F': ['UUU', 'UUC'],
  'L': ['UUA', 'UUG', 'CUU', 'CUA', 'CUC', 'CUG'],
  'I': ['AUU', 'AUC', 'AUA'],
  'M': ['AUG'],
  'V': ['GUU', 'GUA', 'GUC', 'GUG'],
  'S': ['UCU', 'UCA', 'UCC', 'UCG', 'AGU', 'AGC'],
  'P': ['CCU', 'CCA', 'CCC', 'CCG'],
  'T': ['ACU', 'ACA', 'ACC', 'ACG'],
  'A': ['GCU', 'GCA', 'GCC', 'GCG'],
  'Y': ['UAU', 'UAC'],
  'H': ['CAU', 'CAC'],
  'Q': ['CAA', 'CAG'],
  'N': ['AAU', 'AAC'],
  'K': ['AAA', 'AAG'],
  'D': ['GAU', 'GAC'],
  'E': ['GAA', 'GAG'],
  'C': ['UGU', 'UGC'],
  'W': ['UGG'],
  'R': ['CGU', 'CGA', 'CGC', 'CGG', 'AGA', 'AGG'],
  'G': ['GGU', 'GGA', 'GGC', 'GGG'],
  'STOP': ['UAA', 'UAG', 'TGA']
}
```