

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

Attraverso le informazioni raccolte in questa fase saremo in grado di identificare in quali geni si concentrano le variazioni rilevate, all'interno di essi, dove queste avvengono e come alterino i codoni causando la produzione di una proteina diversa.

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 fasta 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, una delle CDS analizzate derivava dall'unione (join) di due sequenze. In tal caso è possibile specificare il punto di unione della sequenza.

3 Listati di codice

Code Listing 1: Tabella per la traduzione in amminoacidi

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

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 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 apposita tramite cui derivare la tabella per l'output finale associando i valori a chiavi prestabilite.

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)

4 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**

gene_id	gene_start	gene_end	cds_start	cds_end	relative_start	relative_end	alteration	original_codone	altered_codone	original_aminoacid	encoded_aminoacid
M	26524	27191	26524	27191	197	197	C	GUG	GUC	V	V
	28275	29533	28275	29533	1100	1100	A	GAG	GAA	E	E
	28275	29533	28275	29533	414	414	C	UUG	CUG	L	L
	28275	29533	28275	29533	561	561	C	UCA	CCA	S	P
N	28275	29533	28275	29533	556	556	U	UCC	UUC	S	F
	28275	29533	28275	29533	607	609	AAC	AGGGGA	AAACGA	RG	KR
	28275	29533	28275	29533	604	604	A	AGU	AAU	S	N
	29559	29674	29559	29674	5	5	U	GGC	GGU	G	G
ORF10	267	21555	267	13483	1131	1131	A	GUA	AUA	V	I
	267	21555	267	13483	1131	1131	A	GUA	AUA	V	I
	267	21555	267	13483	10817	10817	U	UUG	UUU	L	F
	267	21555	267	13483	10817	10817	U	UUG	UUU	L	F
	267	21555	267	13483	18111	18111	U	ACA	UCA	T	S
	267	21555	267	13483	793	793	U	ACC	AUC	T	I
	267	21555	267	13483	793	793	U	ACC	AUC	T	I
	267	21555	267	13483	2771	2771	U	UUC	UUU	F	F
	267	21555	267	13483	2771	2771	U	UUC	UUU	F	F
	267	21555	267	13483	14142	14142	U	CCU	UCU	P	S
	267	21555	267	13483	3637	3637	U	CCA	CUA	P	L
	267	21555	267	13483	3637	3637	U	CCA	CUA	P	L
	267	21555	267	13483	9248	9248	G	UUA	UUG	L	L
	267	21555	267	13483	9248	9248	G	UUA	UUG	L	L
	267	21555	267	13483	13110	13110	G	ACC	GCC	T	A
	267	21555	267	13483	13110	13110	G	ACC	GCC	T	A
ORF1ab	267	21555	267	13483	13210	13210	U	GGG	GUG	A	V
	267	21555	267	13483	13210	13210	U	UGC	UUC	C	F
	267	21555	267	13483	19218	19218	U	GGU	UCU	A	S
	267	21555	267	13483	8441	8441	C	GGU	GGC	G	G
	267	21555	267	13483	8441	8441	C	GGU	GGC	G	G
	267	21555	267	13483	20621	20621	A	AAA	AAA	K	K
	267	21555	267	13483	47	47	U	CUC	CUU	L	L
	267	21555	267	13483	47	47	U	CUC	CUU	L	L
	267	21555	267	13483	8516	8516	U	AGC	AGU	S	S
	267	21555	267	13483	8516	8516	U	AGC	AGU	S	S
	267	21555	267	13483	618	618	U	CGU	UGU	R	C
	267	21555	267	13483	618	618	U	CGU	UGU	R	C
	267	21555	267	13483	1082	1082	U	CCC	CCU	P	P
	267	21555	267	13483	1082	1082	U	CCC	CCU	P	P
	267	21555	267	13483	8893	8893	U	CCU	CUU	P	L
	267	21555	267	13483	8893	8893	U	CCU	CUU	P	L
	267	21555	267	13483	9455	9469	-----	CAUUUCUAUUGGUUUUU	CA-----U	HFYWFF	
	267	21555	267	13483	9455	9469	-----	CAUUUCUAUUGGUUUUU	CA-----U	HFYWFF	
	267	21555	267	13483	19250	19270	-----	AGAUUGUAUCGUAUGUAUAC	AG-----C	RLYDAYN	
	267	21555	267	13483	250	250	C	UUA	UCA	L	S
ORF8	27895	28259	27895	28259	64	64	U	ACU	AUU	T	I
	21564	25384	21564	25384	1172	1172	U	UGC	UGU	C	C
	21564	25384	21564	25384	1840	1840	G	GAU	GGU	D	G
	21564	25384	21564	25384	905	905	U	ACG	ACU	T	T
S	21564	25384	21564	25384	3751	3751	U	GGA	GUA	G	V
	21564	25384	21564	25384	2313	2313	A	GUU	AUU	V	I

Figura 1: Tabella di output delle alterazioni

5 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.

6 Analisi dei risultati e conclusioni

Le alterazioni appartenenti a geni si concentrano per quasi i tre quarti del totale sul gene **ORF1ab**. Variazioni minori seguiti 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.

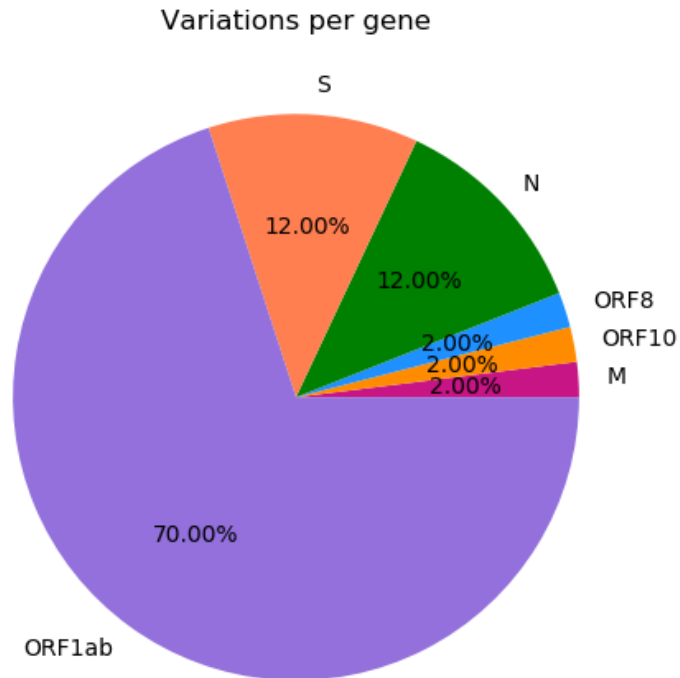


Figura 2: Tipologia di variazioni

6.1 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.