



UNIVERSITÀ DEGLI STUDI DI MILANO - BICOCCA
Scuola di Scienze
Dipartimento di Informatica, Sistemistica e Comunicazione
Corso di Laurea Magistrale in Informatica

Multiple Sequence Alignment (MSA) di sequenze SARS-CoV2

Alberici Federico - 808058
Bettini Ivo Junior - 806878
Traversa Silvia - 816435

Anno Accademico 2019 - 2020

Indice

SARS-CoV-2	2
Obiettivo del progetto	3
Sequenze	3
Tool	3
Analisi verticale	4
New York (USA)	4
Russia	4
Italia	4
Spagna	6

SARS-CoV-2

I coronavirus sono una famiglia di virus RNA a filamento positivo con aspetto simile a una corona se esaminati al microscopio elettronico (da cui il nome) e noti per causare malattie che vanno dal comune raffreddore a patologie più gravi come la Sindrome respiratoria mediorientale (MERS) o la Sindrome respiratoria acuta grave (SARS).

Nel Dicembre 2019 a Wuhan, in Cina, è stato isolato un nuovo ceppo di coronavirus, denominato SARS-CoV-2 (*Severe Acute Respiratory Syndrome - Coronaviru*s - 2), la cui sequenza virale ha un'omologia del 76% con il virus che causò la pandemia della Sars nel 2002. La malattia respiratoria causata da questo nuovo coronavirus è stata chiamata COVID-19.

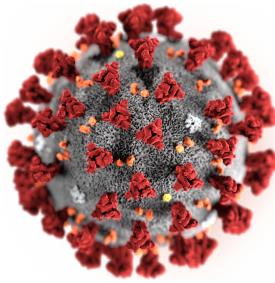


Figure 1: L'illustrazione, creata dai Centers for Disease Control and Prevention (CDC) statunitense, rivela la morfologia ultrastrutturale del SARS-CoV-2

Ogni virione (singola particella virale) SARS-CoV-2 ha un diametro di circa 50-200 nanometri. Come altri coronavirus, SARS-CoV-2 presenta quattro proteine strutturali: S (spike), E (inviluppo), M (membrana) e N (nucleocapside). La proteina N contiene il genoma dell'RNA mentre le proteine S, E, M creano insieme l'inviluppo virale. In particolare la proteina spike è quella che permette al virus di attaccarsi alla membrana di una cellula ospite.

Fra le sequenze genomiche SARS-CoV-2 note è stata evidenziata una bassa variabilità, si ritiene dunque che il ceppo sia stato rilevato dalle autorità sanitarie entro poche settimane dalla sua comparsa tra la popolazione umana alla fine del 2019. Dopo il primo caso di infezione noto, avvenuto in Cina verso la fine del 2019, il virus si è diffuso in tutte le province della Cina e in oltre 150 altri paesi in Asia, Europa, Nord America, Sud America, Africa e Oceania. La trasmissione da uomo a uomo di SARS-CoV-2 è stata confermata nel gennaio 2020 e avviene principalmente attraverso goccioline respiratorie da tosse e starnuti entro un raggio di circa 1,8 metri oppure un'altra possibile causa di infezione è il contatto indiretto tramite superfici contaminate.

Il numero di riproduzione di base del virus è stato stimato tra 1.4 e 3.9, il che significa che ogni infezione dal virus dovrebbe causare da 1.4 a 3.9 nuove infezioni nel caso in cui nessun membro della comunità sia immune e non vengano prese misure preventive.

Al 25 maggio 2020, ci sono stati 5.428.605 casi confermati totali di infezione da SARS-CoV-2 nella pandemia in corso, con un totale di decessi attribuiti al virus di 345.375.

Obiettivo del progetto

L'obiettivo del nostro progetto è riuscire ad allineare le sequenze scaricate (compresa quella che utilizziamo come riferimento per i confronti) con i tool scelti, e produrre in output le variazioni delle sequenze scaricate rispetto alla sequenza di riferimento e creare un'apposita documentazione.

Per poter raggiungere questo obiettivo, abbiamo deciso di dividere la nostra analisi in due parti:

- nella prima parte eseguiamo un'analisi verticale fra la sequenza di riferimento (ossia la prima sequenza che presente nella banca dati raccolta a Wuhan) e delle sequenze prese rispettivamente da Italia, New York, Spagna e Russia. Le sequenze dei singoli paesi sono state prese seguendo un ordine temporale crescente, in modo tale da poter denotare attraverso il confronto delle mutazioni nel tempo.
- nella seconda parte, invece, effettuiamo un confronto orizzontale tra le ultime sequenze raccolte dei vari paesi scelti, in modo da poter individuare la presenza di eventuali mutazioni nello stato corrente del virus, sempre tenendo come riferimento la prima sequenza presente nella banca dati.

Sequenze

Abbiamo deciso di selezionare alcuni dei paesi con il più alto numero di casi di COVID-19, ossia Stati Uniti (per il quale ci siamo concentrati su uno degli stati a sua volta più colpito, quello di New York), Russia, Spagna e Italia. I casi, al 25 Maggio 2020, sono riportati nella tabella seguente:

Stati Uniti	1,678,477
Russia	353,427
Spagna	235,823
Italia	229,858

Le sequenze analizzate sono state prese dal sito GISAID. Per ogni paese abbiamo preso solo sequenze complete (con più di 29000 basi) e su pazienti tutti di sesso maschile. Abbiamo cercato inoltre di seguire una linea temporale che partisse dalla prima sequenza presente nel database fino all'ultima, con in mezzo sequenze che si distanziavano tra di loro di una settimana.

Tool

Tra i tool messi a disposizione abbiamo scelto di utilizzare Clustal Omega, Kalign e MAFFT.

Abbiamo scelto Clustal Omega poiché sfrutta il modello probabilistico Profilo HMM (Hidden Markov Model), il quale è in grado di incapsulare i cambiamenti evolutivi che si sono verificati in una serie di sequenze correlate ed inoltre questo tool risultava ottimale per sequenze molto lunghe.

Kalign è stato scelto poiché è un tool che si concentra sulle regioni locali e lavora

molto velocemente, motivo per il quale ci sembrava particolarmente adatto per le analisi sui singoli stati che volevamo svolgere.
Infine abbiamo scelto MAFFT poichè attraverso l'algoritmo della trasformata di Fourier veloce ottimizza gli allineamenti in base alle proprietà fisiche, utilizzando sia un allineamento progressivo che iterativo. Anche esso risulta essere ottimale per sequenze di lunghezza medio-grandi.

Analisi verticale

New York (USA)

Russia

Italia

```
'1_hCoV-19/Italy/SPL1/2020|EPI_ISL_412974|2020-01-29/1-29903'11083
1 'G' 'T'
'1_hCoV-19/Italy/SPL1/2020|EPI_ISL_412974|2020-01-29/1-29903'26144
1 'G' 'T'
'2_hCoV-19/Italy/CDG1/2020|EPI_ISL_412973|2020-02-20/1-29903'241 1
'C' 'T'
'2_hCoV-19/Italy/CDG1/2020|EPI_ISL_412973|2020-02-20/1-29903'3037 1
'C' 'T'
'2_hCoV-19/Italy/CDG1/2020|EPI_ISL_412973|2020-02-20/1-29903'14408
1 'C' 'T'
'2_hCoV-19/Italy/CDG1/2020|EPI_ISL_412973|2020-02-20/1-29903'23403
1 'A' 'G'
'2_hCoV-19/Italy/CDG1/2020|EPI_ISL_412973|2020-02-20/1-29903'29867
2 'TG' 'NN'
'3_hCoV-19/Italy/FVG-ICGEB_S9/2020|EPI_ISL_417423|2020-03-01/1-29891'0
1 '-' 'G'
'3_hCoV-19/Italy/FVG-ICGEB_S9/2020|EPI_ISL_417423|2020-03-01/1-29891'241
1 'C' 'T'
'3_hCoV-19/Italy/FVG-ICGEB_S9/2020|EPI_ISL_417423|2020-03-01/1-29891'3037
1 'C' 'T'
'3_hCoV-19/Italy/FVG-ICGEB_S9/2020|EPI_ISL_417423|2020-03-01/1-29891'14408
1 'C' 'T'
'3_hCoV-19/Italy/FVG-ICGEB_S9/2020|EPI_ISL_417423|2020-03-01/1-29891'23403
1 'A' 'G'
'3_hCoV-19/Italy/FVG-ICGEB_S9/2020|EPI_ISL_417423|2020-03-01/1-29891'29879
13 'AAAAAAAAAAAAAA' '-----'
'4_hCoV-19/Italy/TE4880/2020|EPI_ISL_418256|2020-03-14/1-29898'42 1
'T' 'Y'
'4_hCoV-19/Italy/TE4880/2020|EPI_ISL_418256|2020-03-14/1-29898'44 2
'CG' 'NN'
'4_hCoV-19/Italy/TE4880/2020|EPI_ISL_418256|2020-03-14/1-29898'241
1 'C' 'T'
'4_hCoV-19/Italy/TE4880/2020|EPI_ISL_418256|2020-03-14/1-29898'3037
1 'C' 'T'
```


'7_hCoV-19/Italy/TE13858/2020|EPI_ISL_435152|2020-04-09/1-29878'29701
1 'G' 'T'
'7_hCoV-19/Italy/TE13858/2020|EPI_ISL_435152|2020-04-09/1-29878'29878
25 'AAAAAAAAAAAAAAAAAAAAAA' ,-----,
'8_hCoV-19/Italy/TE26643/2020|EPI_ISL_436729|2020-04-27/1-29873'1 9
'ATTAAAGGT' 'NNNNNNNNNN'
'8_hCoV-19/Italy/TE26643/2020|EPI_ISL_436729|2020-04-27/1-29873'31
1 'A' 'G'
'8_hCoV-19/Italy/TE26643/2020|EPI_ISL_436729|2020-04-27/1-29873'241
1 'C' 'T'
'8_hCoV-19/Italy/TE26643/2020|EPI_ISL_436729|2020-04-27/1-29873'3037
1 'C' 'T'
'8_hCoV-19/Italy/TE26643/2020|EPI_ISL_436729|2020-04-27/1-29873'3045
1 'C' 'T'
'8_hCoV-19/Italy/TE26643/2020|EPI_ISL_436729|2020-04-27/1-29873'6449
1 'C' 'T'
'8_hCoV-19/Italy/TE26643/2020|EPI_ISL_436729|2020-04-27/1-29873'6863
1 'A' 'M'
'8_hCoV-19/Italy/TE26643/2020|EPI_ISL_436729|2020-04-27/1-29873'6866
1 'A' 'W'
'8_hCoV-19/Italy/TE26643/2020|EPI_ISL_436729|2020-04-27/1-29873'6869
1 'A' 'W'
'8_hCoV-19/Italy/TE26643/2020|EPI_ISL_436729|2020-04-27/1-29873'14408
1 'C' 'T'
'8_hCoV-19/Italy/TE26643/2020|EPI_ISL_436729|2020-04-27/1-29873'19677
1 'G' 'R'
'8_hCoV-19/Italy/TE26643/2020|EPI_ISL_436729|2020-04-27/1-29873'21627
1 'C' 'Y'
'8_hCoV-19/Italy/TE26643/2020|EPI_ISL_436729|2020-04-27/1-29873'23403
1 'A' 'G'
'8_hCoV-19/Italy/TE26643/2020|EPI_ISL_436729|2020-04-27/1-29873'25459
1 'G' 'K'
'8_hCoV-19/Italy/TE26643/2020|EPI_ISL_436729|2020-04-27/1-29873'28881
3 'GGG' 'AAC'
'8_hCoV-19/Italy/TE26643/2020|EPI_ISL_436729|2020-04-27/1-29873'29873
30 'AAAAAAAAAAAAAAAAAAAAAA' ,-----,

Spagna