Bioinformatica

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

SARS-CoV-2

Il SARS-CoV-2 fa parte della famiglia dei coronavirus, virus noti per causare malattie che vanno dal comune raffreddore a malattie più gravi come la Sindrome respiratoria mediorientale (MERS) e la Sindrome respiratoria acuta grave respiratoria (SARS).
Nel Dicembre 2019 a Wuhan, in Cina, è stato isolato un nuovo virus appartenente a questa famiglia, denominato SARS-CoV-2. La sequenza virale ha un'omologia del 76% con il virus che causò la pandemia nel 2002, risultando così molto simile.
I Coronavirus hanno morfologia rotondeggiante e dimensioni di 100-150nm di diametro. Partendo dallo strato più esterno andando verso l'interno troviamo le seguenti componenti: - Glicoproteina S ("spike"), sono le protezioni sulle superfici del virus, lunghe circa 20nm e somigliano ad una corona che circonda il virone. Questa proteina è una degli elementi che varia rispetto al vecchio virus, infatti essa determina la specificità del virus per cellule epiteliali del tratto respiratorio, modelli 3D suggeriscono che sia in grado di legarsi al recettore ACE2 espresso dalle cellule dei capillari dei polmoni;
- Proteina M, proteina membrana che attraversa il rivestimento (envelope) interagendo all'interno del virone con il complesso RNA-proteina;
- Dimero emagglutinina-esterasi (HE): questa proteina del rivestimento, più piccola della glicoproteina S, svolge una funzione importante durante la fase di rilascio del virus all'interno della cellula ospite,
- Proteina E: l'espressione di questa proteina aiuta la glicoproteina S (e quindi il virus) ad attaccarsi alla membrana della cellula bersaglio;
- RNA e proteina N: il genoma dei Coronavirus è costituito da un singolo filamento di RNA a polarità positiva di grande taglia (da 27 a 32 kb nei diversi virus); non sono noti virus a RNA di taglia maggiore. L'RNA dà origine a 7 proteine virali ed è associato alla proteina N, che ne aumenta la stabilità.

Obiettivo del progetto

L'obiettivo del nostro progetto è diviso in due parti:
- nella prima parte vogliamo eseguire un confronto verticale fra la sequenza di riferimento (quindi la prima sequenza che è presente nella banca dati raccolta a Wuhan) e le sequenze prese rispettivamente da Italia, New York, Spagna e Russia. Le sequenze dei vari paesi sono state prese in ordine temporale crescente, in modo che attraverso il confronto potessimo denotare delle mutazioni nel tempo.

-nella seconda parte, invece, abbiamo voluto eseguire un confronto orizzontale tra le utlimissime sequenze raccolte dei vari paesi in modo da poter visualizzare la presenza di eventuali mutazioni, sia con la sequenza di riferimento sia tra di loro.

Scelta delle sequenze

Le seqeunze sono state prese dal sito Gsaid. Per ogni paese abbiamo preso solo sequenze complete e su pazienti tutti di genere maschile, inoltre abbiamo cercato di seguire una linea temporale che partisse dalla prima sequenza presente nel database all'ultima, con in mezzo sequenze che si distanziavano tra di loro di una settimana.
Per quanto riguarda, invece, la scelta dei paesi Italia, NY, Spagna e Russia è stata dettata dal fatto che risultavano essere tra i paesi più colpiti dal virus e quindi li abbiamo ritenuti interessanti da analizzare.

Scelta dei tool

Tra i tool messi a disposizione abbiamo scelto di utilizzare. Clustal Omega, Kalign e MAFFT.
Abbiamo scelto Clustal Omega poichè 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. Per fare questo il modello va a studiare le informazioni specifiche di ogni amminoacido conservato nella colonna di allineamento. Inoltre questo tool risultava utile per sequenze molto lunghe.
Kalign è stato scelto poichè è un tool che si concentra sulle regioni locali e lavoro molto velocemtne. Questo ci sembrava perfetto per il tipo di analisi sui precisi stati che volevamo svolgere.
Infine abbiamo scelto MAFFT poichè attraverso l'algoritmo della trasformata di Fourier veloce ottimizza gli allineamenti proteici in base alle proprietà fisiche degli aminoacidi, utilizzando sia un allineamento progressivo che iterativo. Anche esso risulta essere ottimale per sequenze di lunghezza medio-grandi.

Risultati ottenuti