UNIVERSITÀ
DI PARMA

Dipartimento di Scienze Matematiche, Fisiche e Informatiche

Corso di Laurea Magistrale in Scienze Informatiche

Studio di mobilità su larga scala della proteina Spike del covid-19 con ricerca locale

Long range mobility of covid-19 Spike protein through local search

Candidato:
Lorenzo Mora

Relatore:
Prof. Alessandro Dal Palù

Correlatori:
Prof. Pietro Cozzini
Federica Agosta

Anno Accademico 2021–2022

Dedica

Indice

Introduzione **1**

1 Background **3**

Background **3**
1.1 Amminoacidi 3
1.1.1 Catena laterale 5
1.2 Proteine 5
1.2.1 Classificazione 6
1.2.2 biochimica 7
1.2.3 Composizione 8
1.2.4 Struttura 8
1.3 Principio di Ramachandran 12
1.4 Ricerca Locale 13
1.4.1 Tipologie di ricerca locale 15
1.4.2 Euristiche 16
1.5 Hydropathic INTeractions HINT 17
1.6 Shortest Path 18
1.6.1 Algoritmo A* 19
1.6.2 Esempio 19
1.7 Algebra e Geometria 21
1.7.1 Spazio vettoriale 21
1.7.2 Prodotto scalare 23
1.7.3 Matrici di rotazione 23
1.7.4 Super Fibonacci Spirals 24

2 Covid **27**

Covid **27**
2.1 Covid-19 27
2.1.1 Storia 28

2.1.2 Caratteristiche genetiche 29
2.1.3 RNA polimerasi 30
2.2 Glicoproteina Spike 31
2.2.1 Struttura della proteina e funzione 32
2.2.2 Scudo di glicani della glicoproteina spike 34
2.3 RBD 35

3 Obbiettivi **37**

4 Progettazione **41**
4.1 Biopython 41
4.2 Strategie di ricerca 41
4.3 Strutture dati a supporto 42
4.4 Approccio Top-Down al codice 43

5 Risultati **47**

Conclusione **49**

Bibliografia **51**

A Appendice di Esempio **55**

Elenco delle figure

1.1 Esempio di amminoacido 4
1.2 Struttura dell'amminoacido 4
1.3 Esempio del grafico di Ramachandran 12
1.4 Uno spazio vettoriale è una collezione di oggetti, chiamati "vettori", che possono essere sommati e riscalati. 22

2.1 Molecola di un coronavirus 29
2.2 Principio del RNA polimerasi 31
2.3 Struttura di massima della glicoproteina spike 33
2.4 Struttura del dominio di legame del recettore SARS-CoV-2, nella conformazione aperta (A) e chiusa (B). (C) rappresenta la struttura legata ad ACE2 35

3.1 La Glicoproteina spike nelle due configurazioni: contraddistinta dal colore magenta troviamo la configurazione chiusa; contraddistinta dal colore ciano troviamo la configurazione aperta. 38
3.2 Nell'immagine dettagliata notiamo i due loop che collegano la parte mobile alla parte fissa colorati rispettivamente di rosso e di blu. 40

Elenco degli algoritmi

1 Generazione di n campioni in $\mathcal{SO}(3)$ 26

Elenco delle tabelle

1.1 Classi di amminoacidi 5

Introduzione

L'introduzione deve contenere un riassunto del lavoro di Tesi. In particolare bisogna esprimere chiaramente e molto sinteticamente: contesto dello studio, motivazioni, contributo e conclusioni. Bisogna quindi fare un sommario dello studio ad alto livello, fornendo le intuizioni senza ricadere in dettagli tecnici. Anche lo stile dovrebbe essere più discorsivo rispetto alle parti tecniche della tesi.

Capitolo 1

Background

In questo capitolo verranno introdotti i concetti di base utili alla comprensione del contesto. Andremo ad introdurre cosa sono le proteine e quali sono i loro componenti principali.

1.1 Amminoacidi

Le informazioni trattate in questa sezione sono prese da [WikipediaAmminoacidi,]. Gli amminoacidi sono una categoria di composti organici che hanno sia il gruppo funzionale amminico (-NH_2), sia quello carbossilico (-COOH). La parola aminoacido deriva quindi proprio dall'unione dei due gruppi funzionali citati prima. Siccome sono presenti contemporaneamente un gruppo acido (carbossilico) e un gruppo basico (amminico), sono definite molecole anfotere. Anfotere sono sostanze chimiche che possono manifestare sia un comportamento acido che uno basico.

In biochimica, ci si riferisce di solito ad un sottogruppo dei seguenti, ovvero gli $L-\alpha-amminoacidi$, ovvero amminoacidi il cui gruppo amminico e carbossilico sono legati allo stesso atomo di carbonio, chiamato appunto α e la loro configurazione è ad L, ovvero il gruppo amminico si troverà sempre alla sinistra del carbonio α. Sono presenti 22 $L-\alpha-amminoacidi$ che costituiscono la struttura delle proteine, anche detti amminoacidi proteinogenici. Oltre hai al gruppo carbosillico e al gruppo aminico, ogni amminoacido si contraddistingue dagli altri per la presenza di un residuo R, conosciuto anche con il nome di catena laterale.

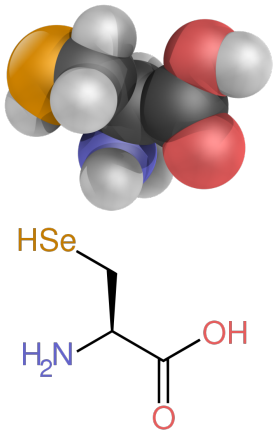

Figura 1.1: Esempio di amminoacido

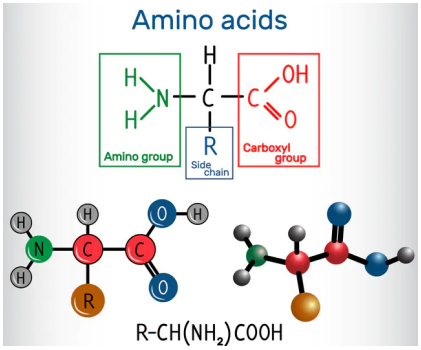

Figura 1.2: Struttura dell'amminoacido

1.1.1 Catena laterale

La catena laterale gioca degli amminoacidi gioca un ruolo importante per la determinazione delle proprietà delle proteine. Esiste una vasta diversità nelle proprietà chimiche delle catene laterali degli amminoacidi, tuttavia essi possono essere raggruppati in 6 classi differenti.

Tipo di catena laterale	`Amminoacidi`
Alifatica	`Glicina, alanina, valina, leucina, isoleucina`
Contenente idrossile o solfuro	`Serina, cisteina, treonina, metionina`
Aromatica	`Fenilalanina, tiroxina, triptofano`
Basica	`Istidina, lisina, arginina`
Acido e la sua ammide	`Acido aspartico, acido glutammico, asparagina, glutammina`
Ciclica	`Prolina`

Tabella 1.1: Classi di amminoacidi

La prolina non può essere inserita in una qualsiasi classe perché è ciclica. La prolina condivide la maggior parte delle proprietà con i gruppi alifatici. La rigidità dell'anello gioca un ruolo cruciale nella struttura delle proteine. Come già detto, gli amminoacidi sono i mattoni di costruzione delle proteine e la metà di questi sono anche essenziali per l'essere umano, poiché non è in grado di produrseli da soli.

Gli amminoacidi in azione combinata o in azione singola sono alla base di molte attività presenti nel nostro corpo.

1.2 Proteine

Le informazioni trattate in questa sezione sono prese da [WikipediaProteine,] A livello chimico, le proteine non sono altro che macromolecole biologiche costituite da catene di amminoacidi legate insieme da un legame peptidico. Il legame peptidico o giunto peptidico è un legame covalente che unisce il gruppo (-NH_2) di un amminoacido con il gruppo (-COOH) di un altro amminoacido. Negli organismi viventi le proteine svolgono innumerevoli funzioni, tra cui la catalisi delle reazioni metaboliche, funzione di sintesi come replicazione del DNA, la risposta a stimoli e il trasporto di molecole da un luogo ad un altro. Le proteine in generale si differiscono nella sequenza degli

amminoacidi, che viene conservata nei geni e che si traduce in un particolare ripiegamento della stessa e una struttura tridimensionale specifica che caratterizza la sua attività.

A differenza di altre macromolecole biologiche come i polisaccaridi e gli acidi nucleici, le proteine sono essenziali negli organismi viventi perché prendono parte a praticamente tutti i processi che avvengono nelle cellule. La maggior parte appartiene alla categoria degli enzimi, che caratterizzano le reazioni biochimiche vitali per il metabolismo degli organismi. Hanno anche funzioni strutturali o meccaniche nei muscoli e che costituiscono il citoscheletro. Alcune sono fondamentali per l'invio di segnali inter ed intracellulari e nella difesa immunitaria. Una volta che sono sintetizzate all'interno del organismo, esistono per un periodo di tempo limitato per poi esser degradate e riciclate attraverso meccanismi cellulari.

Le proteine possono essere purificate da altri componenti cellulari e utilizzando tecniche come: l'ultracentrifugazione; la precipitazione; l'elettroforesi; la cromatografia. L'avvento dell'ingengneria genetica ha portato a nuove tecniche che ne facilitano la purificazione.

Una catena lineare di residui amminoacidi è chiamata polipeptide, ed una proteina generalmente è costituita da uno o più polipeptidi lunghi eventualmente coordinati a gruppi non peptidici, chiamati prostetici o cofattori. I polipeptidi che contengono meno di 20/30 amminoacidi non vengono quasi mai considerati proteine, ma più spesso chiamati peptidi. La sequenza degli amminoacidi in una proteina è definita della sequenza presente nel gene a sua volta codificata nel codice genetico, solitamente ne sono specificati 20 ma possono essere di più in alcuni organismi.

Le proteine che hanno stesso numero e tipo di amminoacidi possono differire per come vengono posti all'interno della struttura, anche una singola variazione può portare ad una variazione nella struttura tridimensionale della macromolecola che può rendere la proteina non funzionale.

1.2.1 Classificazione

Durante l'evoluzione ci sono stati duplicamenti di geni e alterazioni della funzione di una proteina che portano tutt'ora ad avere circa 500 famiglie proteiche. All'interno della stessa famiglia le proteine svolgono funzioni leggermente diverse, ma per quanto riguarda la loro composizione a livello di sequenza di amminoacidi è quasi identica. Chiaramente ci sono anche casi in cui le proteine all'interno della stessa famiglia si differiscono a livello di sequenza di amminoacidi, ma hanno una conformazione tridimensionale molto simile.

Si afferma quindi che nel corso dell'evoluzione si sia più conservata la conformazione tridimensionale, piuttosto che la sequenza. Si può dire che due proteine hanno la stessa struttura generale quando almeno un quarto della loro sequenza amminoacidi corrisponde. Si dice invece che due proteine hanno un qualche grado di parentela, se almeno il 30% degli amminoacidi corrisponde. Alcune proteine possono anche essersi formate per rimescolamento dei domini proteici o per la duplicazione all'interno della proteina stessa con unioni accidentali di DNA.

La classificazione può essere ottenuta grazie alla composizione chimica, alla configurazione molecolare o alla solubilità. Ci sono quindi proteine semplici che sono costituite da soli amminoacidi e proteine coniugate composte dalla proteina semplice e da un gruppo prostetico non proteico.

Tra le proteine semplici vi sono le proteine fibrose, tendenzialmente non solubili nei solventi acquosi e poco attaccabili dagli enzimi proteolitici, inoltre sono presenti le proteine globulari. Mentre per quanto riguarda le proteine coniugate troviamo l'emoglobina, le clorofille e le opsine.

Si possono poi classificare le proteine in base alla funzione che compiono, ci sono le proteine strutturali che sono componenti delle strutture permanenti e che hanno funzione meccanica, poi trovano posto le proteine di trasporto che prendono le sostanza poco idrosolubili e ne consentono il trasporto nell'organismo e poi trovano posto gli enzimi che sono proteine catalitiche. A queste funzioni che abbiamo brevemente descritto si aggiugono la regolazione dell'espressione dei geni, la duplicazione, trascrizione e traduzione del DNA, la regolazione delle reazioni metaboliche, la generazione e la ricezione degli impulsi nervosi.

1.2.2 biochimica

La stra grande maggioranza delle proteine sono costituite da polimeri lineari combinando 20 diversi $L - \alpha - amminoacidi$. Tutti gli aminoacidi che possono essere impiegati nella costruzione di proteine hanno una struttura comune, un carbonio α con un gruppo amminico, un gruppo carbosillico e una catena laterale variabile a seconda dell'aminoacido, l'unica che si distingue è la prolina che contiene un anello insolito al gruppo amminico. Le catene laterali degli amminoacidi hanno ognuna la loro struttura e le loro proprietà chimiche, l'effetto combinato delle catene laterali determina la struttura tridimensionale e la reattività chimica della proteina. Una volta collegati tramite legame peptidico gli amminoacidi sono chiamati residui e la serie di carbonio, azoto e atomi d'ossigeno è nota come catena principale o backbone.

Il legame peptidico ha due forme di risonanza che contribuiscono al doppio legame e non permettono la rotazione attorno al suo asse, in modo tale che i carbonio α dei vari residui siano pressoche complanari. Ci sono però altri due angoli nel legame peptidico che ne determinano la forma assunta.

1.2.3 Composizione

Uno dei dogmi fondamentali della biologia è che ad ogni struttura tridimensionale di una proteina via sia associata una specifica funzione biochimica. In questo modo le proteine possono essere classificate in due famiglie: le proteine globulari e le proteine a struttura estesa o fibrosa. Questa separazione riflette anche una separazione funzionale:

- le proteine estese o fibrose svolgono funzioni biomeccaniche quindi fornendo sostegno strutturale;
- le proteine globulari sono coinvolte in molteplici e specifiche funzioni biologiche e sono di fondamentale importanza per l'economia cellulare.

Una proteina è formata da uno o più polipeptidi, che sono molecole con più di 10 unità di amminoacidi, eventualmente accompagnati o legati da uno o più gruppi prostetici. Una proteina attiva può esistere solo in soluzioni salina diluita e la sua struttura dipendera esclusivamente dalle caratteristiche chimico-fisiche della soluzione in cui è inserita, che può anche determinare modifiche strutturali e alterare le sue propieta funzionali.

La molecola proteica risulta costituita da atomi di carbonio, ossigeno, idrogeno e azoto, può contenere anche zolfo e, talvolta, fosforo e/o metalli.

1.2.4 Struttura

Mettiamo un sunto sensato delle sottosottosezioni....

Ripiegamento

Prendendo una proteina, ovvero una macromolecola formata da decina di migliaia di atomi, potrebbe assumere un numero di ripiegamenti elevati; tuttavia, non è cosi perché ci sono considerazioni fisiche che limitano la maggior parte dei ripiegamenti. Gli atomi non possono sovrapporsi e il loro comportamento è da immaginarsi come delle semplici sfere, con un certo raggio detto raggio di van der Waals. Ciascun amminoacido contribuisce alla formazione della catena con tre possibili legami:

- legame peptidico (C-N) tra il carbonio di un amminoacido e l'azoto di un amminoacido adiacente;

- legame convezionalmente chiamato Cα-C che è presente nei due carboni della catena principale del singolo amminoacido;

- legame Cα-N all'interno dello stesso amminoacido.

Il legame peptidico è planare e non consente alcuna rotazione, mentre gli altri due legami possono ruotare e definiscono due angoli: l'angolo di rotazione del legame Cα-C è detto ψ; l'angolo di rotazione del legame Cα-N è φ. La conformazione degli atomi che fanno parte della catena principale è determinata dagli angoli descritti in precedenza. Non è pero possibile ruotare come si vuole questi angoli dato che non sono possibili collisioni steriche tra gli amminoacidi. Ramachandran come vedremo nella prossima sezione nel dettaglio ha individuato e rappresentato in un grafico le coppie di angoli di rotazione a seconda delle coppie di atomi. Dal grafico si può vedere che le proteine assumono due grandi tipologie di conformazione: l'$\alpha - elica$ e il $\beta - foglietto$.

Tra gli atomi che sono all'interno di una proteina si stabiliscono dei legami che possono essere covalenti o non covalenti, i legami non covalenti sono sicuramente meno potenti, tuttavia il numero all'internoo di una proteina li rende fondamentali per comprendere il ripiegamento. Ci sono tre tipi di legami non covalenti:

- il legame idrogeno, che si effettua tra un atomo di ossigeno e uno vicino ad idrogeno;

- le attrazzioni elettrostatiche che avvengono tra gruppi laterali con cariche periferiche opposte;

- le attrazzioni di van der Waals si verificano tra dipoli molecola istantanei indotti, tra dipoli permanenti o tra un dipolo permanente e uno corrisponde indotto, nelle quali entrano in gioco forze diverse.

Vanno aggiunte alle precendenti interazioni la tendenza dei gruppo di amminoacidi idrofobici ad avvicinarsi e unirsi tra loro, formando delle tasche idrofobiche che sono però lontane dai legami idrogeno sempre presenti in un ambiente acquoso. Tendenzialmente questi gruppi sono posti all'interno della proteina, mentre i suoi amminoacidi idrofili (polari e con carica) saranno tendenzialmente all'esterno, poiché essa si trova tipicamente in un

ambiente acquoso. Di solito la proteina tende poi ad assumere la struttura tridimensionale che ha la più bassa energia libera.

Le conformazioni più comuni

L'$\alpha - elica$ e il $\beta - foglietto$ sono le conformazioni più comuni riscontrabili nelle catene polipeptidiche di una proteina. Una singola proteina può prevedere sia $\alpha - elica$ che $\beta - foglietto$ in numero variabile.

L'$\alpha - elica$ è la più comune nelle proteine, in particolare nei recettori cellulari, e si possono trovare più $\alpha - elica$ per singola proteina. L'elica è una delle conformazioni più favorevoli perché riduce al minimo l'energia libera. Essa si forma quando una catena polipeptidica si ripiega su se stessa con formazione di legami idrogeno tra un legame peptidico e il quarto successivo, nel dettaglio tra il gruppo chetonico C=O dell'uno e il gruppo N-H dell'altro, e il legame è tra O e H. Tutti i gruppi amminici di un'elica sono rivolti verso l'N-terminale della proteina, tutti quelli chetonici verso il C-terminale, così l'elica assume parziale carica positiva all'N-terminale e parziale carica negativa al C-terminale.

Il $\beta - foglio$ pieghettato è la seconda conformazione più comune nelle proteine, si trova maggiormente in alcuni enzimi e nelle proteine coinvolte nella difesa immunitaria. Esso consiste in numerose catene polipeptidiche che si dispongono l'una adiacente all'altra, collegate in una struttura continua da brevi sequenze ad U.

Livelli di organizzazione

All'interno della proteina si possono distinguere vari livelli di organizzazione, che possono essere tre o quattro a seconda della tipologia della stessa. I livelli d'organizzazione sono:

- la struttura primaria è formata dalla sequenza specifica di amminoacidi, dalla catena peptidica e il numero delle catene determina anche il ripiegamento della stessa;

- la struttura secondaria consiste nella conformazione delle catene ($\alpha - elica$, $\beta - foglietto$, etc..). All'interno di una singola proteina vi può essere una combinazione una combinazione di varie tipologie di sequenze;

- il dominio è un'unità globulare o fibrosa formata da catene polipeptidiche ripiegate in più regioni compatte, costituiscono divisioni della

struttura terziaria, ha la caratteristica di ripiegarsi più o meno indipendentemente rispetto al resto della proteina. Molte delle proteine più complesse sono aggregazioni modulari di numerosi dominii proteici;

- la struttura terziaria, che dal punto di vista termodinamo è quella ccon la più bassa energia libera, è rappresentata dalla configurazione tridimensionale completa che la catena polipeptidica assume nell'ambiente in cui si trova;

- la struttura quaternaria deriva dall'associazione di due o più unità polipeptidiche, unite tra loro da legami deboli.

Le proteine contenenti una parte non polipeptidica sono anche dette coniugate. Due proteine possono essere definite isoforme se a parità di struttura primaria ci sono differenze in uno degli altri livelli di struttura. Quando si dice denaturare una proteina significa distruggerne la sua conformazione spaziale, anche se viene mantenuta intatta la struttura primaria, essa non è più in grado di esplicare la sua funzione.

Proteine complesse

Le proteine prese singolarmente sono si complesse, ma all'intermo degli organismi possono aggregarsi ad altre proteine identiche e non creando cosi dei complessi proteici. Tutto ciò avviene grazie ai legami non covalenti che permettono ad una proteina di assumere una determinata configurazione. Nella proteina sono presenti una o più zone capaci di interazioni covalenti e vengono chiamate sidi di legame. Quando le proteine si uniscono insieme in un complesso proteico, le singole unità vengono chiamate subunità proteiche.

Ci sono complessi proteici formati da molte subunità che permettono di realizzare filamenti, poiché in un polo possiedono un sito di legame e dall'altro una struttura proteica complementare allo stesso sito. Vi sono poi proteine la cui funzione è resa possibile proprio dalla loro struttura poco caratterizzabile e quasi casuale; esse principalmente hanno molte funzioni all'interno della cellula. La loro caratteristica è quella di avere ridondanza di amminoacidi e bassa presenza di amminoacidi idrofobici.

Ci possono essere casi in cui la proteina è esposta ad un alto livello di degradazione, esse sono quindi stabilizzate da legami di solfuro, che permettono di mantenere la sua conformazione.

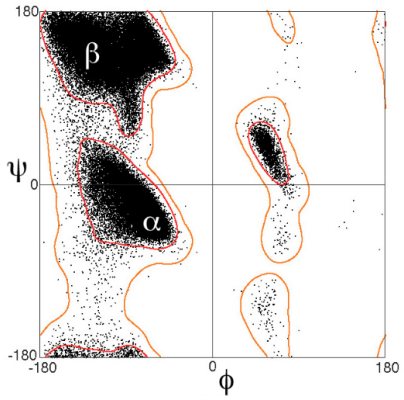

Figura 1.3: Esempio del grafico di Ramachandran

1.3 Principio di Ramachandran

Le informazioni trattate in questa sezione sono prese da [Martz,] e da [Proteopedia,]. Come detto nella sezione precedente Phi φ e Psi ψ determinano la conformazione di un polipeptide. Il principio di Ramachandran ci dice quali $\alpha - elica$, $\beta - foglietto$ e spire sono le conformazioni più probabili da adottare per una catena polipeptidica, poiché la maggior parte delle configurazioni sono impossibili a causa delle collisioni steriche tra gli atomi.

Il principio di Ramachandran ha stabilito che solo alcune configurazioni di angoli di torsione sono stabili e presenti naturalmente nelle proteine. Queste configurazioni stabili sono descritte come regioni favorevoli o regioni accettabili nel grafico di Ramachandran. Le configurazioni non favorevoli o non accettabili sono associate con strutture instabili o anomale.

Il principio di Ramachandran viene utilizzato nella biologia strutturale per verificare la qualità delle predizioni di struttura proteica e per identificare eventuali errore nella modellizzazione. Viene anche utilizzato nella progettazione di proteine artificiali e nella modifica della struttura delle proteine esistenti per migliorare le proprietà terapeutiche o funzionali.

Non tutte le coppie di angoli teoricamente possibili sono realmente ottenibili all'interno delle strutture peptidiche in quanto impedite dalla presenza di ingombri sterici fra i gruppi delle catene laterali dei residui amminoacidici; le zone del grafico in cui tali contatti sterici sfavorevoli non sono presenti possono essere delimitate da linee di contorno, e possono essere associate ai motivi secondari assunti dalla catena. La conformazione complessiva del peptide è quindi definita assegnando i valori a ciascuna coppia di angoli ψi, φi per ogni amminoacido.

Nel caso della glicina, le zone di conformazioni possibili nel grafico pre-

sentano una simmetria centrale e sono molto più ampie rispetto a quelle degli altri amminoacidi grazie alla simmetria del residuo ed alle piccole dimensioni dell'atomo di idrogeno legato in posizione α. Gli altri amminoacidi presentano invece un diagramma asimmetrico, con un insieme di zone indicanti le conformazioni ammesse meno esteso, e la cui ampiezza dipende soprattutto dalla presenza di gruppi stericamente ingombranti in posizione β. La presenza di catene laterali, anche lunghe, che non siano ramificate al carbonio β, come nel caso della leucina, ha invece una scarsa influenza e riduce solo di poco il dominio delle conformazioni permesse.

Un caso a parte si ha per la prolina, la cui struttura rigida rende possibili solo poche conformazioni in una porzione limitata del grafico.

1.4 Ricerca Locale

La ricerca locale è una tecnica euristica di ottimizzazione e risoluzione dei problemi che mira a trovare una soluzione che sia vicina alla soluzione attuale e la migliori. Funziona apportando piccole modifiche alla soluzione corrente e valutando i risultati, con l'obiettivo di trovare una soluzione ottimale in uno spazio di ricerca limitato. La ricerca locale viene spesso utilizzata nei problemi di ottimizzazione combinatoria, come il problema del commesso viaggiatore, e nell'apprendimento automatico, dove può essere utilizzata per ottimizzare algoritmi complessi come le reti neurali. L'idea chiave alla base della ricerca locale è evitare di rimanere bloccati in soluzioni non ottimali apportando una serie di piccole modifiche informate alla soluzione corrente fino a quando non viene trovata una soluzione ottimale.

La ricerca locale è un sottocampo di:

- Metaheuristic: è una procedura o euristica di livello superiore progettata per trovare, generare o selezionare un'euristica che può fornire una soluzione sufficientemente buona a un problema di ottimizzazione, in particolari con informazioni incomplete o limitate capacità di calcolo;

- Stochastic optimization: sono metodi di ottimizzazione che generano e utilizzano variabili casuali; I metodi di ottimizzazione stocastica generalizzano metodi deterministici per problemi deterministici;

- Mathematical optimization: è la selezione di un elemento migliore rispetto ad un qualche criterio, da un insieme di alternative disponibili; Nell'approccio più generale, un problema di ottimizzazione consiste nel

> massimizzare o minimizzare una funzione reale scegliendo sistematicamente i valori di input all'interno di un insieme consentito e calcolando il valore della funzione.

La maggior parte dei problemi può essere formulata in termini di spazio di ricerca e target in diversi modi. Ad esempio, per il problema del commesso viaggiatore una soluzione può essere un percorso che tocca tutte le città e l'obiettivo è trovare il percorso più breve. Ma una soluzione può anche essere un percorso, che può anche contenere un ciclo.

Un algoritmo di ricerca locale parte da una soluzione candidata e quindi si sposta iterativamente verso una soluzione vicina; un quartiere è l'insieme di tutte le possibili soluzioni che differiscono dalla soluzione attuale per la minima misura possibile. Ciò richiede la definizione di una relazione di vicinato nello spazio di ricerca. Ad esempio, l'intorno della copertura del vertice è un'altra copertura del vertice che differisce solo per un nodo. Per la soddisfacibilità booleana, i vicini di un'assegnazione booleana sono quelli che hanno una singola variabile in uno stato opposto. Lo stesso problema può avere più quartieri distinti definiti su di esso; l'ottimizzazione locale con quartieri che implicano la modifica di fino a k componenti della soluzione viene spesso definita k-opt.

Tipicamente, ogni soluzione candidata ha più di una soluzione vicina; la scelta di quale selezionare viene presa utilizzando solo le informazioni sulle soluzioni nelle vicinanze dell'assegnazione corrente, da cui il nome ricerca locale. Quando la scelta della soluzione del vicino è fatta prendendo quella che massimizza localmente il criterio, cioè: una ricerca avida, la metaeuristica prende il nome di hill climbing. Quando non sono presenti vicini in miglioramento, la ricerca locale è bloccata in un punto localmente ottimale. Questo problema di ottimo locale può essere risolto utilizzando i riavvii (ricerca locale ripetuta con diverse condizioni iniziali), la randomizzazione o schemi più complessi basati su iterazioni, come la ricerca locale iterata, sulla memoria, come l'ottimizzazione della ricerca reattiva, su modifiche stocastiche senza memoria, come la simulated annealing.

La ricerca locale non fornisce una garanzia che una determinata soluzione sia ottimale. La ricerca può terminare dopo un determinato periodo di tempo o quando la migliore soluzione trovata fino a quel momento non è migliorata in un determinato numero di passaggi. La ricerca locale è un algoritmo anytime: può restituire una soluzione valida anche se viene interrotta in qualsiasi momento dopo aver trovato la prima soluzione valida. La ricerca locale è in genere un'approssimazione o un algoritmo incompleto, poiché la ricerca potrebbe interrompersi anche se la migliore soluzione corrente trovata non è ottimale. Ciò può accadere anche se la terminazione avviene perché la mi-

gliore soluzione attuale non può essere migliorata, poiché la soluzione ottima può trovarsi lontano dall'intorno delle soluzioni attraversate dall'algoritmo.

Schuurman & Southey propongono tre misure di efficacia per la ricerca locale (profondità, mobilità e copertura):

- Profondità: il costo dell'attuale (migliore) soluzione;
- Mobilità: la capacità di spostarsi rapidamente in diverse aree dello spazio di ricerca (mantenendo bassi i costi);
- Copertura: quanto sistematicamente la ricerca copre lo spazio di ricerca, la distanza massima tra qualsiasi incarico inesplorato e tutti gli incarichi visitati.

Ipotizzano che gli algoritmi di ricerca locale funzionino bene, non perché abbiano una certa comprensione dello spazio di ricerca, ma perché si spostano rapidamente in regioni promettenti ed esplorano lo spazio di ricerca a basse profondità nel modo più rapido, ampio e sistematico possibile.

1.4.1 Tipologie di ricerca locale

La ricerca locale si basa sul presupposto che il miglioramento di una soluzione che si trovi nelle immediate vicinanze di quest'ultima (vicinato). Data una qualsiasi soluzione euristica (nodo corrente) la ricerca locale verifica l'esistenza di soluzioni migliori nell'insieme delle soluzioni più vicine (nodi vicini). Ad esempio, un algoritmo di ricerca locale può essere utilizzato per risolvere il problema del commesso viaggiatore. Le tecniche di ricerca locale consentono di raggiungere risultati sia nella ricerca delle soluzioni a un problema (problem solving) e sia nell'ottimizzazione. Appartengono alla categoria degli algoritmi di ricerca locale i seguenti algoritmi:

- Ricerca hill climbing: L'algoritmo di ricerca hill climbing è una tecnica di ricerca locale in cui il nodo corrente è sostituito con il nodo migliore;
- Simulated annealing: L'algoritmo di Simulated annealing o Annealing simulato (SA) è un algoritmo probabilistico-metaeuristico di ottimizzazione della ricerca in uno spazio di ricerca di grandi dimensioni. Il suo nome prende spunto dal processo metallurgico che consente di riaggregare la materia fusa dei metalli in base a determinate caratteristiche

comuni. Nell'algoritmo di simulated annealing un processo di selezione consente l'eventuale sostituzione della soluzione del nodo corrente con quella di un nodo selezionato casualmente da una lista di soluzioni candidate;

- Beam Search. La beam search è una ricerca locale basata su tecniche euristiche. La beam search (ricerca di fascio) esplora un fascio limitato di nodi promettenti (soluzioni parziali) dello spazio di ricerca e li analizza con una ricerca locale. Utilizzando le tecniche euristiche l'algoritmo beam search consente di riconoscere quando una soluzione parziale è anche una soluzione completa.

1.4.2 Euristiche

Un algoritmo di tipo Ricerca Locale appartengono alla classe delle euristiche di miglioramento. Sono euristici quegli algoritmi che non garantiscono di fornire la soluzione ottima. Ovvero, data una soluzione iniziale ammissibile s, essa è migliorata attraverso successive trasformazioni di s.

Le euristiche sono quindi strategie che aiutano a trovare una soluzione ottimale in una ricerca locale. Ecco alcune delle eurisiche più comuni utilizzate nella ricerca locale:

- First-Improvement: Questa euristica cerca sempre la prima soluzione migliore rispetto alla soluzione corrente. Il vantaggio di questa euristica è che è molto veloce, poiché interrompe la ricerca non appena viene trovata una soluzione migliore. Tuttavia, il suo limite è che potrebbe non essere in grado di trovare la soluzione ottimale, poiché potrebbe fermarsi presto;

- Best-Improvement: Questa euristica cerca sempre la migliore soluzione possibile rispetto alla soluzione corrente. Il vantaggio di questa euristica è che è molto precisa, poiché cerca sempre la soluzione migliore. Tuttavia, il suo limite è che potrebbe essere molto lenta, poiché deve valutare tutte le soluzioni possibili prima di scegliere la migliore;

- Stochastic Hill Climbing: Questa euristica cerca una soluzione migliore casualmente. Il vantaggio di questa euristica è che è molto flessibile, poiché cerca soluzioni in modo casuale. Tuttavia, il suo limite è che potrebbe essere impreciso, poiché potrebbe scegliere soluzioni subottimali;

- Simulated Annealing: Questa euristica cerca una soluzione migliore basata sulla probabilità. Il vantaggio di questa euristica è che è molto precisa, poiché cerca soluzioni basate sulla probabilità. Tuttavia, il suo limite è che potrebbe essere molto lenta, poiché deve valutare molte soluzioni prima di scegliere la soluzione migliore;

- Variable Neighborhood Search: Questa euristica cerca soluzioni migliori cambiando continuamente il vicinato della soluzione corrente. Il vantaggio di questa euristica è che è molto flessibile, poiché cerca soluzioni in molti vicinati diversi. Tuttavia, il suo limite è che potrebbe essere molto lenta, poiché deve valutare molte soluzioni prima di scegliere la soluzione migliore;

- Greedy Algorithm: Questa euristica seleziona sempre la soluzione che sembra essere la migliore in un determinato momento. Il vantaggio di questa euristica è che è molto veloce, poiché seleziona sempre la soluzione migliore. Tuttavia, il suo limite è che potrebbe non essere in grado di trovare la soluzione ottimale, poiché seleziona sempre la soluzione migliore in un dato momento, senza preoccuparsi delle conseguenze future. Pertanto, è importante utilizzare questo algoritmo con cautela e verificare che la soluzione ottenuta soddisfi i requisiti del problema.

1.5 Hydropathic INTeractions HINT

Le informazioni trattate in questa sezione sono prese da [Federica Agosta,] e da [Wikipedia,]. La valutazione della stabilità intramolecolare di una proteina gioca un ruolo fondamentale nella comprensione del loro comportamento e nel meccanismo di azione. Piccole alterazioni strutturali possono impattare l'attività biologica e di conseguenza la modulazione farmacologica. L'analisi della struttura tridimensionale della proteina e la risultante stabilità permettono di predirre un possibile meccanismo di attivazione e rivelano nuove strategie per scoprire nuovi farmaci. Ogni modifica apportata alla struttura di una proteina porta quasi sicuramente ad una variazione del suo meccanismo di azione, e valutare la stabilità di una proteina mutata può essere molto costoso in termini di tempo. La stabilità delle proteine mutate è influenzata dall'interazione intramolecolare, ovvero sono forze che permettono di creare e tenere insieme una struttura. Nel tentativo di analizzare la stabilità proteica, sono stati sviluppati molti metodi che svariano in diversi approcci dai meccanismi molecolari basati su force field fino a tecniche di machine learning. Tuttavia, questa pratica risulta essere molto costosa, specialmen-

te quando sono presenti un elevato numero di mutazioni. Uno degli aspetti fondamentali su cui è basato HINT è dovuta al fatto che la stabilità delle proteine mutate è dimostrato essere influenzata dalle interazioni intramolecolari. Le strutture native delle proteine sono relativamente stabili poiché si formano come risultato di un equilibrio tra le varie forze non covalenti a cui sono soggette: i legami idrogeno, i legami ionici, le forze di Wan der Walls e le interazioni idrofobiche. In generale le interazioni chimiche deboli sono attrazzioni tra atomi appartenenti o alla stessa molecola (intramolecolari) o a molecole diverse (intermolecolari). A differenza dei legami covalenti che legano gli atomi tra loro, le interazioni chimiche deboli non sono forti abbastanza per legare tra loro atomi isolati. Per questo le interazioni chimiche deboli si formano e si rompono in continuazione alla temperatura fisiologica dell'organismo. a meno che, cumulandosi in gran numero, esse non diano collettivamente stabilità alle strutture che contribuiscono a generare.

HINT è stato sviluppato sulla base dei lavoro svolto da (indica il personaggio) che ha esteso il metodo fragment per la predizione dei $\log P\_o/v$ (coefficiente di partizione per il trasferimento di soluti in 1-ottanolo/acqua), cio permette di quantificare le interazioni idrofobiche e polari tra o all'interno dei molecole. La funzione di energia intramolecolare di HINT può essere calcolata rapidamente dalla struttura in termini di somma di punteggi d'interazione atomo-atomo. Il punteggio intramolecolare viene calcolato come somma delle interazioni idrofobiche e polari tra tutte le coppie di atomi considerando l'area accessibile al solvente. Le interazioni idrofobiche si verificano quando le molecole non polari si avvicinano tra di loro in un solvente polare, come l'acqua. Le molecole non polari tendono a respingersi reciprocamente e ad attirarsi con le molecole del solvente. Il processo anche noto come esclusione dei solventi, porta alla formazioni di molecole non polari all'interno del solvente. Le molecole non polari quindi cercano di organizzarsi in modo tale da minimizzare il contatto con il solvente e massimizzare le interazioni tra di loro.

Il punteggio energetico di HINT intramolecolare consente quindi di calcolare tutte le interazioni idropatiche (idrofobiche e polari) che si verificano nella molecola e ne consentono di valutare la stabilità termodinamica. Dato il suo livello di sensibilità nell'individuare piccole differenze di energie verrà poi utilizzato per stimare la stabilità della Glicoproteina spike del SARS-CoV-2.

1.6 Shortest Path

Lo shortest path è un importante concetto nell'ambito dell'informatica cosi come nell'ingegneria e nella matematica applicata. Essa si riferisce al per-

corso più breve tra due punti in un grafo, ovvero insieme di nodi collegati da archi.

Esso è molto importante per esempio nel campo della robotica, viene utilizzato per pianificare il percorso del robot, consentendogli di evitare ostacoli e raggiungere il suo obiettivo in modo efficiente. Possiamo trovare lo stesso concetto applicato nelle reti di telecomunicazioni, poiché viene utilizzato per instradare il traffico tra due nodi garantendo una comunicazione efficiente e affidabile. Viene anche usato in algoritmi di routing per cercare di minimizzare il tempo di ritardo. Chiaramente esso si adatta bene nel campo della logistica dove è necessario ottimizzare il trasporto delle merci o delle persone. Infine, trova applicazione anche nel campo della progettazione di circuiti elettronici.

1.6.1 Algoritmo A*

L'algoritmo A* è spesso utilizzato nella risoluzione del problema dello shortest path, ovvero per trovare il percorso più breve tra due nodi in un grafo pesato. L'algoritmo A* utilizza una funzione di stima euristica per guidare la ricerca verso il percorso più breve, permettendo di ottenere un'efficace combinazione di completezza e velocità nell'individuare la soluzione.

Nella versione dello shortest path risolto con l'algoritmo A*, ogni nodo del grafo viene valutato in base alla sua distanza dal nodo di partenza, la stima della distanza dal nodo di arrivo e il costo totale del percorso, ovvero la somma della distanza dal nodo di partenza e della stima della distanza dal nodo di arrivo. In questo modo, l'algoritmo A* tiene traccia del percorso più breve scoperto fino a quel momento e lo utilizza per determinare quale nodo espandere successivamente.

L'algoritmo A* può essere ulteriormente ottimizzato attraverso l'utilizzo di tecniche come la memorizzazione della distanza minima calcolata per ogni nodo, in modo da ridurre il numero di espansioni necessarie. Inoltre, l'algoritmo A* può essere utilizzato con successo in grafi di grandi dimensioni, grazie alla sua capacità di selezionare le espansioni più promettenti e di escludere le aree meno promettenti del grafo.

1.6.2 Esempio

In questo esempio, la funzione A_star prende in input la cella di partenza 'start', la cella di arrivo 'goal' e una rappresentazione del labirinto come 'graph'. La funzione A_star esplora le celle adiacenti alla cella corrente, calcola il costo del percorso per ogni cella e aggiunge le celle alla coda di priorità. La coda di priorità è ordinata in base al costo totale, che è la

somma del costo del percorso e della stima euristica. La funzione A_star restituisce il percorso più breve dal punto di partenza al punto di arrivo.

```
    # Definizione della funzione di costo
    def cost(current, next):
      if current[0] == next[0] or current[1] == next[1]:
        return 1
      else:
        return math.sqrt(2)

    # Definizione della funzione euristica
    def heuristic(current, goal):
      return math.sqrt((current[0]-goal[0])**2 + (current
    [1]-goal[1])**2)

    # Definizione dell'algoritmo A*
    def A_star(start, goal, graph):
      queue = []
      heapq.heappush(queue, (0, start, []))
      visited = set()

      while queue:
        _, current, path = heapq.heappop(queue)
        if current == goal:
        return path + [current]
        if current in visited:
          continue
        visited.add(current)
        for neighbor in graph[current]:
          if neighbor in visited:
            continue
          new_cost = cost(current, neighbor)
          total_cost = new_cost + heuristic(neighbor, goal)
          heapq.heappush(queue, (total_cost, neighbor, path
    + [current]))
      return None
```

Codice 1.1: Algoritmo A*

1.7 Algebra e Geometria

L'algebra e la geometria sono due aree fondamentali della matematica che sono strettamente correlate e si influenzano a vicenda. La geometria si occupa dello studio delle figure geometriche nello spazio, come punti, linee, piani, poligoni, solidi e curve.

L'algebra, come accennato in precedenza, si occupa dello studio delle proprietà delle operazioni aritmetiche e delle relazioni tra le variabili. L'algebra fornisce uno strumento matematico molto potente per la risoluzione di problemi e l'espressione di relazioni matematiche in modo astratto.

L'algebra e la geometria sono strettamente correlate, poiché l'algebra può essere utilizzata per risolvere problemi geometrici, come ad esempio la determinazione di equazioni di rette e di superfici, la risoluzione di equazioni di secondo grado che descrivono curve, e la risoluzione di problemi di trigonometria.

D'altra parte, la geometria può essere utilizzata per visualizzare le relazioni algebriche, ad esempio rappresentando graficamente le funzioni algebriche o le equazioni differenziali.

1.7.1 Spazio vettoriale

Gli spazi vettoriali sono una struttura algebrica complessa, composta da: un campo, in cui gli elementi sono detti scalari; un insieme, i cui elementi sono detti vettori; due operazioni binarie, la somma e la moltiplicazione scalare con determinate proprietà.

Per ogni numero naturale $n \geq 1$ consideriamo l'insieme $\mathbf{R}^n$ delle n-uple ordinate di numeri reali:

$$\mathbf{R}^n = \{v = (x_1, .., x_n) | x_i \in \mathbf{R}\}$$

.

Viene chiamato $\mathbf{R}^n$ lo spazio n-dimensionale e i suoi elementi vettori o punti. Presi due vettori

$$v = (x_1, .., x_n) e w = (y_1, .., y_n)$$

si possono sommare componente per componente per ottenere un nuovo vettore:

$$v + w = (x_1 + y_1, .., x_n + y_n)$$

.

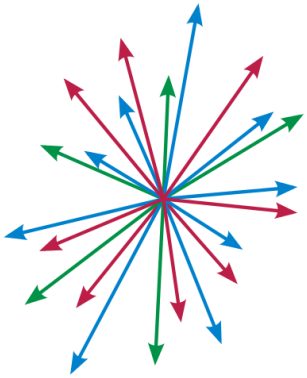

Figura 1.4: Uno spazio vettoriale è una collezione di oggetti, chiamati "vettori", che possono essere sommati e riscalati.

Un numero $c \in \mathbf{R}$ si può moltiplicare con un vettore $v = (x_1, .., x_n)$ in modo da ottenere un nuovo vettore:

$$cv = (cx_1, .., cx_n)$$

.

Le operazioni di somma e moltiplicazione per scalare sono anche chiamate operazioni di spazio vettoriale.

Per la somma di vettori $v, w, u \in \mathbf{R}$ vengono verificate le seguenti proprietà:

- $v + w = w + v$;
- $(v + w) + u = v + (w + u)$;
- $v + \mathcal{O} = v, v + (-v) = \mathcal{O}$, dove $\mathcal{O} = (0, .., 0)$.

Per la moltiplicazione per scalare di $a, b \in \mathbf{R}$ con $v, w \in \mathbf{R}^n$ valgono invece:

- $a(bv) = (ab)v$;
- $a(v + w) = av + aw, (a + b)v = av + bv$;
- $1v = v$.

Si osserva che le operazioni di somma e di moltiplicazione per scalare si possono definire non solo nell'insieme dei numeri reali o delle n-uple di numeri, ma anche in altri insiemi (di funzioni, di matrici, ...). Un insieme con somma e moltiplicazione per scalare che verificano tutte le proprietà precedentemente descritte si dice spazio vettoriale.

1.7.2 Prodotto scalare

Il prodotto scalare di $v = (x_1, .., x_n) e w = (y_1, .., y_n)$ è il numero

$$\langle v, w \rangle := x_1 y_1 + x_2 y_2 + ... + x_n y_n = \sum_{i=1}^{n} x_i y_i$$

talvolta viene anche denotato come $v \cdot w$.

Ci sono importanti proprietà alla base del prodotto scalare, ovvero per $v, w, u \in \mathbf{R}^n e c \in \mathbf{R}$ valgono:

- $\langle v, w \rangle = \langle w, v \rangle$;
- $\langle v, w + u \rangle = \langle v, w \rangle + \langle v, u \rangle$;
- $\langle cv, w \rangle = c\langle v, w \rangle = \langle v, cw \rangle$;
- $\langle v, v \rangle \geq 0$, e $\langle v, v \rangle = 0 \iff v = \mathcal{O}$ (positività di $\langle , \rangle$).

Nel piano cartesiano il prodotto scalare permette di definire e trattare la nozione geometrica di lunghezza di un vettore. Tale concetto può essere esteso ad uno spazio vettoriale di dimensione arbitraria introducendo un concetto analogo: la norma.

1.7.3 Matrici di rotazione

In matematica, e in particolare in geometria, una rotazione è una trasformazione del piano o dello spazio euclideo che sposta gli oggetti in modo rigido e che lascia fisso almeno un punto, nel caso del piano, o una retta, nel caso dello spazio. I punti che restano fissi nella trasformazione formano più in generale un sottospazio: quando questo insieme è un punto o una retta, si chiama rispettivamente il centro e l'asse della rotazione.

Più precisamente, una rotazione è una isometria di uno spazio euclideo che ne preserva l'orientazione, ed è descritta da una matrice ortogonale speciale. Qualunque sia il numero delle dimensioni dello spazio di rotazione, gli elementi della rotazione sono:

- il verso (orario-antiorario);
- l'ampiezza dell'angolo di rotazione;
- l centro di rotazione (il punto attorno a cui avviene il movimento rotatorio).

Nel nostro caso ci concentriamo nello spazio a 3 dimensioni; in questo caso la rotazione è determinata da un asse, dato da una retta r passante per l'origine, e da un angolo θ di rotazione. Per evitare ambiguità, si fissa una direzione dell'asse, e si considera la rotazione di angolo θ effettuata in senso antiorario rispetto all'asse orientato. Senza cambiare base, la rotazione di un angolo θ intorno ad un asse determinato dal versore (x, y, z) (ossia un vettore di modulo unitario) è descritta dalla matrice seguente:

$$\begin{bmatrix} x^2 + (1 - x^2)\cos(\theta) & xy(1 - \cos(\theta)) - z\sin\theta & xz(1 - \cos\theta) + y\sin\theta \\ xy(1 - \cos(\theta)) - z\sin\theta & y^2 + (1 - y^2)\cos(\theta) & yz(1 - \cos\theta) - x\sin\theta \\ xz(1 - \cos(\theta)) - y\sin\theta & yz(1 - \cos\theta) + x\sin\theta & z^2 + (1 - z^2)\cos(\theta) \end{bmatrix}$$

Ponendo $(x, y, z) = (1, 0, 0)$ oppure $(x, y, z) = (0, 1, 0)$ oppure $(x, y, z) = (0, 0, 1)$ si ottiene rispettivamente la rotazione attorno all'asse x, all'asse y e all'asse z. Tale matrice è stata ottenuta scrivendo la matrice associata alla trasformazione lineare (rispetto alle basi canoniche nel dominio e codominio) della formula di Rodrigues.

1.7.4 Super Fibonacci Spirals

Le informazioni di questa sezione sono state recuperate da [Alexa, 2022]. Le super spirali di Fibonacci sono un'estensione delle spirali di Fibonacci che consentono la generazione rapida di un arbitrario ma fisso numero di orientamenti 3D. Una valutazione completa rispetto ad altri metodi mostra che gli insiemi di orientamenti generati hanno una bassa discrepanza, componenti spuri minimi nello spettro di potenza e volumi Voronoi quasi identici. L'ottimizzazione degli insiemi per discrepanze basse è difficile. Inoltre, l'ottimizzazione $\mathcal{SO}(3)$ è scomoda a causa della geometria sottostante. [1]

Lo strumento principale per usare il campionamento di Fibonacci su $\mathcal{S}^3$ è una mappatura che preserva il volume di un cilindro solido in $\mathcal{R}^3$ alla 3-sfera.

[1] $\mathcal{SO}(3)$ è il gruppo delle rotazioni tridimensionali che preservano la norma e il prodotto vettoriale, e che hanno determinante 1. Sono tutte le possibili rotazioni attorno ad un punto nello spazio tridimensionale, senza alcuna traslazione.

Considerando il cilindro $(h, y = (y_0, y_1))| - \pi < h \leq \pi, y^T y \leq 1$. Allora

$$x(h, y) = \begin{pmatrix} z \cos h \\ z \sin h \\ y_0 \\ y_1 \end{pmatrix}, z = \sqrt{1 - y^T y}$$

mappa punti nel cilindro della sfera $\mathcal{R}^4$. La mappatura inversa di $x = (x_0, x_1, x_2, x_3)$ è data da $(h, y_0, y_1) = (\arctan 2(x_1, x_0), x_2, x_3)$. Questo mostra che la mappatura è una biezione tra l'interno relativo del cilindro e la sfera senza equatore $x_0 = x_1 = 0$. La linea $-\pi < h \leq \pi, y^T y \leq 1$ sulla superficie del cilindro sono mappate ai punti $(0, 0, y_0, y_1)$ sull'equatore della sfera. All'interno del cilindro $y^T y < 1$ dove la mappatura è biettiva, possiamo calcolare la Jacobians come:

$$\mathcal{J}_{h,j} = \begin{pmatrix} -z \sin h - \frac{y_0}{z} \cos h - \frac{y_1}{z} \cos h \\ z \cos h - \frac{y_0}{z} \sin h - \frac{y_1}{z} \sin h \\ 010 \\ 001 \end{pmatrix}$$

Usando quindi Jacobians possiamo analizzare il cambio di volume e reclamare che la mappatura $x(h, y) = (h, y)| - \pi < h \leq \pi, y^T y < 1 \mapsto \mathcal{S}^3 \subset \mathcal{R}^4$ preserva il volume.

Si può dimostrare ciò che viene reclamato in precedenza come $det(x(h, y), J_(h, y))$, perché $x(h, y)$ è ortogonale alla tangente al piano e ha lunghezza unitaria. Sviluppando si ottiene

$$\begin{aligned} &+z \cos h(z \cos h) - z \sin h(z \sin h) \\ &+y_0(y_0 \sin^2 h + y_0 \cos^2 h) \\ &-y_1(y_1 \cos^2 h - y_1 \sin^2 h) \\ &= (1 - y^T y)(\cos^2 h + \sin^2 h) + y^T y = 1 \end{aligned}$$

.

Data la mappatura, l'idea è quella di utilizzare il campionamento di Fibonacci 2 volte per generare punti sul cilindro: la prima lungo l'asse principale h del cilindro; la seconda sul cerchio (y_0, y_1) ortogonale all'asse. Usando due differenti costanti ϕ e ψ, per i due campionamenti otteniamo:

$$y(t) = \left(\sqrt{t} \sin \frac{2\pi n t}{\phi}, \sqrt{t} \cos \frac{2\pi n t}{\phi} \right)$$

$$z(t) = \left(\frac{nt}{\psi} - \left\lfloor \frac{nt}{\psi} \right\rfloor, \sqrt{t} \sin \frac{2\pi n t}{\phi}, \sqrt{t} \cos \frac{2\pi n t}{\phi} \right)^T$$

.

Inserendo questo campione nella mappatura della 3-sfera otteniamo la seguente semplice curva che esibisce la simmetria aspettata:

$$w(t) = \begin{pmatrix} \sqrt{t}\sin\frac{2\pi nt}{\phi} \\ \sqrt{t}\cos\frac{2\pi nt}{\phi} \\ \sqrt{t-1}\sin\frac{2\pi nt}{\psi} \\ \sqrt{t-1}\cos\frac{2\pi nt}{\psi} \end{pmatrix}$$

.

Il campionamento di questa curva a valori regolari di t_i è un metodo naturale per la generazione di campioni di orientamento. L'implementazione algoritmica è la seguente 1.

Algoritmo 1 Generazione di n campioni in $\mathcal{SO}(3)$

function SUPER-FIBONACCI(n, ϕ, ψ)
  for $i \in 0, .., n-1$ **do**
    $s \leftarrow i + \frac{1}{2}$
    $t \leftarrow \frac{s}{n}$, $d \leftarrow 2\pi s$
    $r \leftarrow \sqrt{t}$, $R \leftarrow \sqrt{1-t}$
    $\alpha \leftarrow \frac{d}{\phi}$, $\beta \leftarrow \frac{d}{\psi}$
    $q_i \leftarrow (r\sin\alpha, r\cos\alpha, R\sin\beta, R\cos\beta)$

Questo algoritmo mostra che un insieme di kn campioni contiene l'insieme generato per n campioni o, più generalmente, insieme con m e n campioni condividono ogni k-esimo campione, dove k è il minimo comune diviso tra m e n. Giocano un ruolo fondamentale anche gli altri due parametri di questo algoritmo, ovvero ϕ e ψ che devono essere irrazionali, ma anche la loro relazione è importante. Non ci sono però teorie matematiche a supporto quindi è necessario adattarli alla situazione d'uso.

Capitolo 2

Covid

In questo capitolo verrà introdotto il virus covid, la funzione della proteina spike e vedremo una conformazione della stessa.

2.1 Covid-19

Le informazioni in questa sezione sono state prese da [MALIK,]. I coronavirus sono stati scoperti negli anni 60 e da allora i coronavirus sugli umani sono stati identificati a partire dal SARS-CoV nel 2002. La pandemia da Covid-19 è causata da un nuovo ceppo virale chiamato SARS-CoV-2, un virus a singola elica di RNA della famiglia *coronaviridae*. Le specie patogene individuate nel tempo sono SARS-CoV, MERS-CoV e appunto SARS-CoV-2 di ordine nidovirale della famiglia coronaviridae e sotto famiglia ortocoronavirinae, e tutte hanno un aspetto a forma di corona solare. Dei 4 generi di coronavirus $(\alpha, \beta, \gamma, \delta)$ fa parte dei $\beta - CoV$ e mostra molte somiglianze con 2 coronavirus derivati dai pipistrelli.

SARS-CoV e MERS-CoV hanno avuto origine nei pipistrelli, e sembra che sia così anche per SARS-CoV-2. Era gia stata dimostrata in precedenza la possibilità che un host intermedio facilitasse l'emergere del virus negli umani; dopodiché il contagio tra uomo e uomo può avvenire attraverso contatti ravvicinati con goccioline respiratorie, oppure con con diretto con infetti e o per contatto con oggetti e superfici contaminate.

Il genoma del virus contiene 4 strutture proteiche: la proteina spike; la membrana; l'envelope; la nucleocapside. La proteina spike media l'attaccamento del virus ai recettori dell'ospite, ma la approfondiremo nella prossima sezione. La membrana è la proteina più abbondante e definisce la forma dell'involucro virale. La proteina envelope è la più piccola proteina appartenente alla struttura virilica, partecipa all'assemblaggio virale e al gemogliamento.

La proteina nucleocapside è l'unica che si lega al genoma del RNA ed è coinvolta nel assemblaggio virale e nel germogliamento.

La replicazione del virus inizia con l'attaccamento e l'ingresso; l'attaccamento del virus alla cellula ospite avviene tra la proteina spike e il recettore specifico. Una volta che si viene a creare il legame, il virus entra nel citosol della cellula ospite. Dopodiché avviene la traduzione del gene per la replicazione dal RNA genomico e quindi poi la traduzione e l'assemblaggio dei processi di replicazione virale. Una volta conclusa questa fase avviene l'incapsulamento che porta alla formazione del virus maturo. Una volta completato il processo di assemblaggio viene trasportato sulla superficie cellulare e rilasciato per esocitosi.

Prima dell'epidemia del SARS-CoV erano già stati indivuduati due tipologie di coronavirus nell'uomo che però erano visti come le cause del raffreddore. Con la comparsa nel 2012 di MERS-CoV e l'attuale SARS-CoV-2 è necessario studiare e comprendere appieno quelle che sono le proprietà e le caratteristiche di questo virus.

2.1.1 Storia

Le informazioni in questa sezione sono state prese da [Tinelli,] e da [di Sanità EpiCentro L'epidemiologia per la sanità pubblica,]. Ha avuto inizio il 31 Dicembre del 2019 quando l'OMS (Organizzazione Mondiale della Sanità) è stata informata di casi di polmonite di eziologia sconosciuta nella città di Wuhan provincia di Hubei Cina. Il nuovo corona virus è stato ufficialmente annunciato il 7 gennaio del 2020 e tre giorni dopo è stata resa pubblica la sequenza genomica. Sono state poi rilasciate altre sequenze genomiche, le quali tutte suggerivano la presenza di un virus strettamente legato al SARS-CoV.

L'11 Febbraio del 2020 l'OMS ha definito la nuova polmonite indotta da coronavirus come malattia da corona virus 2019. Allo stesso tempo la Commissione internazionale di classificazione dei virus ha annunciato che il virus nominato provvisoriamente come 2019-nCoV veniva nominato come grave sindrome respiratoria acuta SARS-CoV2. Dopo che il patogeno è stato valutato sulla base della filogenesi, della tassonomia e della pratica consolidata, è stato definito un forte legame con il precedente SARS-CoV.

L'inizio della pandemia è avvenuto quindi a Wuhan in Cina. In Italia si sviluppo poi un focolaio autoctono che poi si è diffuso progressivamente in tutto il paese e in particolare nelle regioni del nord. Successivamente il virus si espanse in Europa e nel resto del mondo. L'OMS dichiarò l'inizio della pandemia l'11 Marzo del 2020, è poi storia di ogni giorno della pandemia che

ha raggiunto milioni di persone. Si ritiene che comunque il tasso di mortalità del virus sia di circa il 3.5%.

2.1.2 Caratteristiche genetiche

I coronavirus sono sferici con un diametro di circa 125nm con punte a forma di clava che sporgono dalla superficie del virus che danno l'aspetto di una corona solare.

All'interno dell'involucro troviamo una simmetria elicoidale dei nucleocapsidi, che in realtà è molto rara tra i virus a RNA a senso positivo. I CoV sono classificati nell'ordine Nidovirales, famiglia Coronaviriade e sotto famiglia Orthocoronavirinae. Nidovirales è un ordine di virus a RNA a singolo filamento positivo. Essi si distinguono da altri virus a RNA per la loro lunghezza e complessità del genoma e per la presenza di una struttura a corona di proteine sulla superficie virale. Il loro genoma contiene diversi geni che codificano proteine non strutturali coinvolte nella replicazione e nella trascrizione del virus nonché geni che codificano le proteine strutturali che costituiscono il virus stesso. Gli Nidovirales includono quattro famiglie virali:

- Coronaviridae,
- Arteriviridae,
- Roniviridae,
- Mesoniviridae.

La famiglia Coronaviridae è una famiglia di virus a RNA a singolo filamento positivo che causano malattie respiratorie e gastroenteriti in animali e

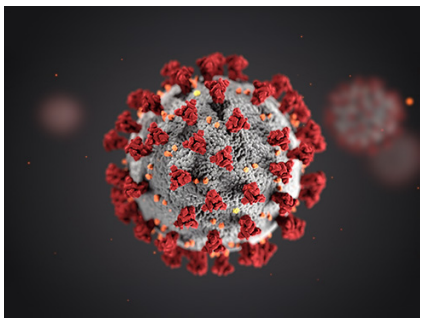

Figura 2.1: Molecola di un coronavirus

in alcuni casi anche nell'uomo. I coronavirus sono stati scoperti per la prima volta negli anni '60, e devono il loro nome alla loro forma a corona, che deriva dalle protuberanze sulla loro superficie virale. La famiglia Coronaviridae è suddivisa in quattro generi:

- Alphacoronavirus,
- Betacoronavirus,
- Gammacoronavirus,
- Deltacoronavirus.

Un'analisi filogenetica ha inserito SARS-CoV2 sotto il sottogenere Sarbecovirus del genere Betacoronavirus.

Le 4 proteine strutturali, citate in precedenza, sono richieste dalla maggior parte dei CoV per produrre una particella virale strutturalmente completa suggerendo che alcuni Cov possono codificare proteine aggiuntive con funzioni in sovrapposizione compensative.

2.1.3 RNA polimerasi

Le informazioni in questa sezione sono prese da [Raffaele,] La fase di ingresso del virus viene approfondita nella prossima sezione, ora pensiamo a cosa succede quando è all'interno. Una volta all'interno della parte acquosa della cellula ospite, chiamata citosol, il virus si "scompone" rilasciando il suo contenuto; un insieme di proteine e materiale genetico. L'uomo conserva l'informazione genetica, ovvero quella che ci permette di costruire le cellule, nelle molecole di DNA, il virus utilizza una singola molecola di RNA. L'RNA è presente anche nell'uomo, ma che principalmente viene utilizzato per la costruzione delle proteine.

Una volta che il coronavirus ha infettato una cellula, il suo scopo è quello di far si che si costruiscano nuove copie dello stesso in modo da avere una discendenza in grado di espandere la specie. Tutto quello che è presente nella singola cellula infetta deve essere "copiato" e assemblato per formare nuovi vironi. Per ottenere la "copia" è necessario replicare il materiale genetico "RNA" del virus stesso. Il meccanismo che permette di far ciò si chiama RNA polimerasi, ovvero un enzima che è in grado di formare lunghe catene di RNA. In modo specifico l'RNA polimerasi di SARS-CoV-2 si chiama RdRP (RNA polimerasi RNA-dipendente).

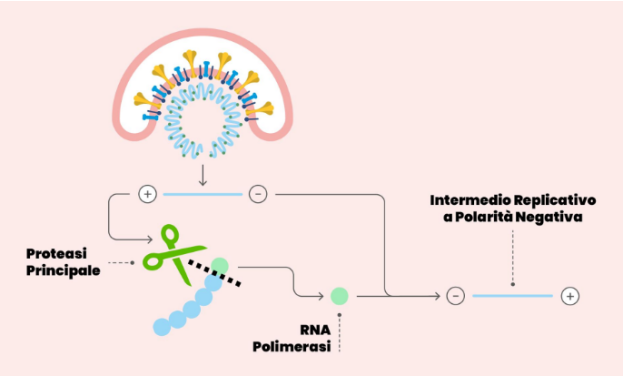

Figura 2.2: Principio del RNA polimerasi

Esso viene però definito "distratto", infatti ogni volta che effettua la "copia" del materiale genetico originale commette degli errori, di conseguenza la copia presenta delle differenze. Le seguenti differenze non sono altro che mutazioni che si ripercuotono sulla struttura delle proteine del virus e sulla loro funzione. Non a caso si sono sentite varie varianti presenti come Delta, Omicron, etc... Quello che succede è che l'RNA polimerasi ha fatto una copia del genoma virale e ha cambiato una base di RNA. In ogni caso le mutazioni hanno conseguenze sul virus, possono essere neutre, vantaggiose o svantaggiose. Una mutazione può iniziare a circolare se essa da un vantaggio per il suo ciclo vitale, questo può significare essere più abile nell'infettare, ma provocare meno danni nell'organismo ospite, dato che un virus altamente letale è destinato a scomparire per mancanza di ospiti.

2.2 Glicoproteina Spike

Le informazioni in questo sezione sono state prese da [Duan Liangwei,] La glicoproteina spike svolge un ruolo essenziale nell'attaccamento, nella fusione e nell'ingresso del virus nella cellula ospite. Una caratteristica dei coronavirus è quella di accedere alle cellule ospiti e poi dare inizio all'infezione attraverso la fusione delle membrane virali alle cellule. La fusione della membrana viene mediata dalla membrana di tipo 1 della glicoproteina spike e dal recettore affine della cellula ospite. Essendo in una posizione superficiale nella struttura del virus, ciò la rende un bersaglio diretto per le risposte immunitarie dell'ospite rendendola anche il principale bersaglio degli anticorpi. Data la sua importanza nella replicazione e fusione virale è al centro della maggior parte delle strategie vaccinali e degli interventi terapeutici.

La glicoproteina Spike viene sintetizzata come precursore di una poliproteina sul reticolo endoplasmatico ruvido (RER). Il precursore non processato ospita una sequenza segnale del reticolo endoplasmatico (ER) situato nel terminale N, che indirizza la glicoproteina alla membrana RER. Durante la sintesi vengono aggiunte catene laterali di oligosaccaridi ad alto contenuto di mannosio. Poco dopo la sintesi i monomeri della glicoproteina trimerizzano, il che può facilitare il trasporto dall'ER al complesso di Golgi. Il complesso di Golgi è un organulo di composizione lipo-proteica con una delicata struttura nella cellula in posizione paranucleare che si occupa di rielabolare, selezionare ed esportare i prodotti del reticolo endoplasmatico. All'interno del complesso di Golgi la glicoproteina spike viene scissa proteoliticamente dalla furina cellulare o da proteasi simili in S1 e S2. La subunità di superficie S1, che attacca il virus al recettore della superficie della cellula ospite e la subunità S2 che media la fusione delle membrane cellulari alla cellula ospite. Anche dopo la fase di scissione le subunità S1 e S2 rimangono associate attraverso interazioni non covalenti in uno stato di profusione metastabile. La scissione è però necessaria per l'infettività virale ed è anche necessaria per un efficacie infezione delle cellule polmonari. Un segnale di recupero dell'ER costituito da un motivo conservato KxHxx assicura che la proteina matura si accumuli vicino al compartimento intermedio di Golgi dove guidata dall'interazione con la proteina di membrana (M) partecipano all'assemblaggio delle particelle virali. Una frazione delle proteine mature viaggia attraverso via secretoria fino alla membrana plasmatica, dove può mediare la fusione di cellule infette con cellule non infette per formare cellule giganti multinucleate.

2.2.1 Struttura della proteina e funzione

Come accenato nei precedenti paragrafi la glicoproteina spike svolge un ruolo fondamentale nell'infezione virale e nella patogenesi. Essa è un trimero fortemente glicosilato. La subunità S1 è composta da 672 amminoacidi ed organizzata in 4 domini: un dominio N-terminale; un dominio C-terminale, noto anche come dominio di legame; due sottodomini SD1 e SD2. Mentre la subunità S2 è composta da 588 amminoacidi e contiene un peptide di fusione idrofobica N-terminale, due ripetizioni eptade, un dominio transmembrana e una coda citoplasmatica.

Come una tipica proteina di fusione di classe I la glicoproteina spike condivide caratteristiche strutturali topologiche e meccaniche comuni con altre proteina di fusione di classe I come la glicoproteina dell'involucro dell'HIV e l'emoagglutinina del virus dell'influenza. Essa è una macchina coformazionale che media l'ingresso virale riorganizzando da uno stato non unliganded metastabile, attraverso uno stato intermedio ad uno stato post fusione stabi-

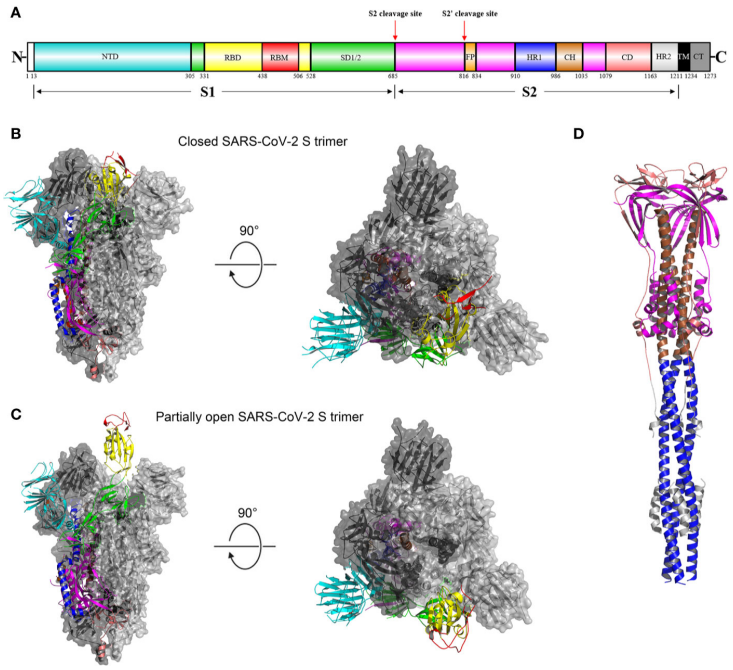

Figura 2.3: Struttura di massima della glicoproteina spike

le. Da quando è stata resa pubblica la struttura sono state scoperte numerose strutture per i frammenti di trimero della glicoproteina spike negli stati pre e post fusione.

L'architettura di massima dell'ectodominio pre fusione della spike è stabilizzato da due mutazioni consecutive della prolina in due conformazioni determinate dalla microscopia crioelettronica a singola particella è un trimero con una sezione trasversale triangolare. La differenza strutturale tra queste due conformazioni risiede solo nella posizione di uno dei tre RDB. Quando tutti e tre gli RBD sono nella posizione giù il trimero risultante di ectodominio S assume una configurazione chiusa, in cui la superficie di legame del recettore dell'RBD S1 è sepolta tra i protomeri e non può essere accessibile dal suo recettore (Fig. **??**B). Il trimero di ectodominio S con un singolo RBD nella posizione "up" assume una conformazione parzialmente aperta e rappresenta lo stato funzionale poiché la superficie di legame del recettore del RBD "up" può essere completamente esposta (Fig. **??**C). La subunità S1 riposa mentre il trimero S2 stabilizzano quest'ultimo nella conformazione di pre fusione. Quando il trimero di ectodomionio adotta una conformazione parzialmente aperta l'RBD nella posizione "su" abolirà i contatti con la subunità S2 di un protomero adiacente, destabilizzando la conformazione parzialmente aperta. Ciò sarà vantaggioso per la dissociazione e faciliterà i riagganciamenti subiti

per mediare l'ingresso virale.

Le strutture di pre fusione del coronavirus umano senza due mutazioni consecutive della prolina rivelano solo una conformazione completamente chiusa. In particolare è noto che la pre fusione trimerica risiede principalmente in una configurazione chiusa che è conformazionalmente mascherata per eludere le neutralizzazioni mediate dagli anticorpi. Si può quindi pensare che le glicoproteine spike del covid-19 native su virioni maturi e infettivi che condividano una simile caratteristica di mascheramento conformazionale, nascondendo la superficie di legame del recettore.

2.2.2 Scudo di glicani della glicoproteina spike

Come nominato in precedenza la proteina spike del SARS-CoV-2 è fortemente circondata da glicani N-legati che sporgono dalla superficie del trimero. Sono stati incontrati fino a 22 glicani N-legati che probabilmente svolgono un ruolo importante nel ripiegamento delle proteine e nell'invasione immunitaria dell'ospite come scudo glicano. Dei 22 potenziali disponibili per la glicosilazione, 14 vengono identificati come prevalentemente occupati da glicani di tipo complesso. I restanti invece risultano dominati da glicani di tipo oligomannosio che sono diversi da quelli fondati sulle glicoproteine dell'ospite. Per glicosilazione si intende una modifica della struttura della proteina da parte del complesso di Goigi, durante o in seguito ad un processo di sintesi proteica. Essa avviene per più motivi, uno dei quali è il raggiungimento del ripiegamento corretto, la può proteggere dall'attaccco di proteasi e aumenta la solubilità della molecola che viene dunque stabilizzata in tutti gli aspetti. Si può anche affermare che l'affinità di legame tra la proteina spike del SARS-CoV-2 e ACE2 non dipendono dalla glicosalinzione della stessa.

Quando i glicani specifici sono mappati sulla struttura di pre fusione dell'ectodominio della spike del SARS-CoV-2 il modello ha mostrato livelli sostanzialmente più elevati di superficie priva di glicani. Questo ci porta alla considerazione che la proteina spike del SARS-CoV-2 è ricoperta da uno scudo meno denso e quindi risulta essere una buona notizia per i potenziali vaccini.

Nel caso di SARS-CoV-2, più recentemente è stato dimostrato che un potente anticorpo neutralizzante sia contro SARS-CoV che SARS-CoV-2, S309, riconosce un epitopo RBD contenente glicano altamente conservato. Queste osservazioni suggeriscono che le frazioni di carboidrati potrebberò essere immunogeniche ed evidenziano la necessità per gli immunogeni di mostrare i glicani importanti per il riconoscimento degli anticorpi neutralizzanti. Di conseguenza anche in questo caso è diventa fondamentale la ricerca in questo campo per i vaccini.

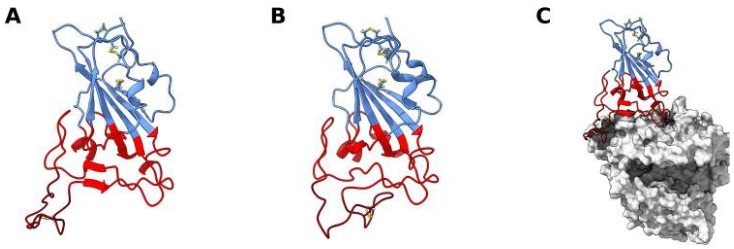

Figura 2.4: Struttura del dominio di legame del recettore SARS-CoV-2, nella conformazione aperta (A) e chiusa (B). (C) rappresenta la struttura legata ad ACE2

2.3 RBD

Le informazioni in questa sezione sono state prese da [Valério M,]. Diverse linee di ricerca hanno stabilito che l'enzima di conversione dell'angiotesina 2 (ACE2) è un recettore di ingresso per SARS-CoV-2. Interazioni dettagliate tra il SARS-CoV-2 RBD e il suo recettore sono state rivelate da diverse strutture in ACE2. Come detto nella sezione precedente le subunità S1 e S2 sono responsabili del legame del recettore e della fusione della membrana. La subunità S1 è costituita da un dominio N-terminale e un dominio di legame o RBD. Nello stato di pre fusione, la proteina S esiste come omotrimero e subisce grandi cambiamenti conformazionali per controllare l'esposizione e l'accessibilità del RBD.Tutto questo avviene mediante un meccanismo di "su" e "giù", la differenza sta nel rendere accessibile o inaccessibile il recettore. La struttura del nucleo RBD quando è lagata ad ACE2 è costituita da un foglio β antiparallelo a cinque filamenti intrecciati con eliche e anelli di collegamento corti.

Questa struttura centrale del foglio β è ulteriormente stabilizzata da 3 legami di di solfuro. Tra i filamenti centrali c'è una regione estesa contenente 2 filamenti β corti, le eliche e gli anelli. Questa regione è il motivo legante il recettore (RBM) che contiene la maggior parte dei residui responsabili dell'interazione con ACE2. Quando complessato con ACE2, l'RBM si ripiega in una superficie concava che ospita l'α-elica N-terinale di ACE2. E' proprio in questa superficie che diversi residui di RBM stabiliscono interazioni specifiche e non specifiche con i residui di ACE2. Dai dati disponibili riguardo alla struttura sembrerebbe che la struttura centrale sia abbastanza stabile, mentre l'RBM risulta molto dinamico e non definito strutturalmente, a meno che non sia legato ad altre proteine come ACE2.

Durante il corso della pandemia sono state segnalate un numero significativo di mutazioni naturali della proteina Spike. Molte delle mutazioni sono

state identificate nel RBD, alcune delle quali hanno dato origine a varianti virali. Si ritiene che molte di queste mutazioni RBD aumentino l'affinità di legame per ACE2 o riescono ad ingannare in modo migliore gli anticorpi monoclonali.

I vaccini che sono stati elaborati nel corso della pandemia vanno proprio ad agire in questa zona tra RBD e ACE2, cercando di impedire che avvenga il contatto e che quindi impedire di conseguenza l'ingresso del virus all'interno dell'ospite.

Capitolo 3

Obbiettivi

Come spiegato nel precedente capitolo il caso di studio su cui ci focalizziamo e incentriamo il nostro lavoro è la glicoproteina spike del SARS-CoV-2. Focalizziamo la nostra attenzione su questa proteina per studiarne il comportamento attraverso l'uso di tecniche di ricerca locale, avvicinandosi molto alle tecniche di dinamica e modellistica molecolare. Queste due tecniche introdotte si basano su metodi teorici o tecniche computazionali utilizzate per simulare il comportamento delle molecole. Queste tecniche vengono utilizzate per studiare la dinamica di evoluzione nel tempo di un sistema chimico e svolgere il processo mediante l'utilizzo della tecnologia permette l'applicazione della modellistica a sistemi relativamente complessi. Il livello di dettaglio che si può ottenere utilizzando questi sistemi è il livello atomistico; il più piccolo livello di informazione rappresentato dagli atomi individuali (o piccoli gruppi di atomi). Queste tecniche vengono utilizzate per svariati compiti dal folding proteico, la catalisi enzimatica, la stabilità delle proteine, i cambiamenti conformazionali associati alla funzione biomolecolare e il riconoscimento molecolare delle proteine. Le tecniche descritte in precedenza hanno però delle problematiche a partire dalle risorse necessarie a compiere questo tipo di compiti, il tempo computazionale necessario e il fatto che si possa lavorare in modo completo solo su piccole strutture oppure è necessario limitare di molto lo spazio di lavoro.

Ricollegandoci alla Glicoproteina spike del SARS-CoV-2 prendiamo in considerazione entrambe le configurazioni sia quello in stato chiuso che quella in stato aperto; esse differiscono per lo stato in cui si trova la parte mobile: nella chiusa si trova in uno stato di pre fusione; nello stato aperto è pronta a legarsi ad un recettore. Possiamo vedere realmente la differenza delle due configurazioni guardando la Fig. 3.1.

L'obbiettivo di questo framework lavorando sulle due configurazione della glicoproteina spike del SARS-CoV-2 è quello di trovare uno shortest path tra

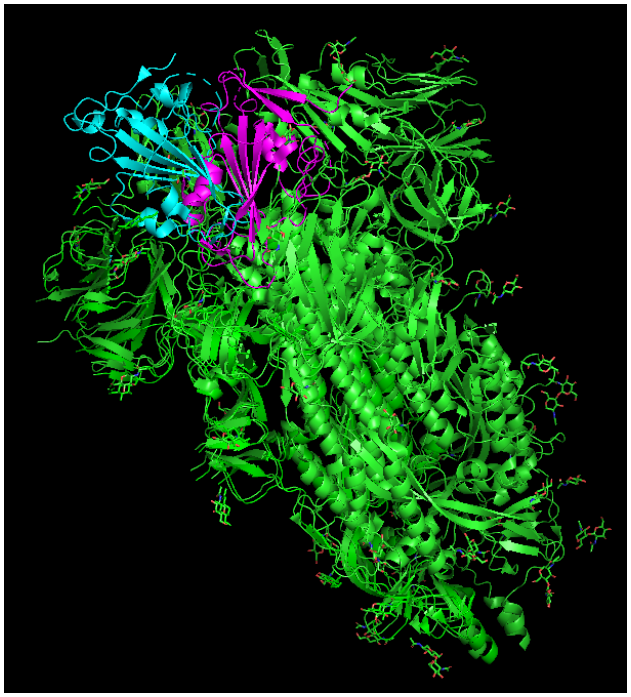

Figura 3.1: La Glicoproteina spike nelle due configurazioni: contraddistinta dal colore magenta troviamo la configurazione chiusa; contraddistinta dal colore ciano troviamo la configurazione aperta.

le due configurazioni utilizzando due principi differenti per il costo del singolo movimento.

- calcolo del costo mediante la sola geometria, ovvero somma delle componenti di traslazione e angolo di rotazione

- calcolo del costo mediante non solo la geometria, ma l'utilizzo della funzione d'energia per minimizzare le variazioni di energia.

Il percorso più breve porta con se la necessità di ottenere delle configurazioni intermedie per raggiungere l'obbiettivo preposto e questo ci porta a dover affrontare il problema della fattibilità della connessione tra parte mobile nella nuova configurazione e la parte fissa [Fig. 3.2]. In questo caso si va effettivamente ad agire sugli amminoacidi che fanno parte dei due loop che collegano appunto le parti e in questo caso possiamo affermare che il framework proposto si occupa non solo del movimento della backbone (catena principale), ma tiene in considerazione anche la presenza delle catene laterali. Le catene laterali non vengono coinvolte attivamente nel movimento, ma viene considerato l'ingombro nei confronti e nel rispetto delle proprietà chimico-fisiche. Il framework quindi muove la catena principale rispettando la struttura dell'amminoacido e poi nel rispetto della struttura si preoccupa che la catena laterale non effettui clash con nessun amminoacido nelle vicinanze.

Il motivo per cui è stato effettuato questo lavoro è quello di proporre un framework che fosse in grado di fornire un metodo per lo studio del movimento di una proteina in modo più ampio rispettando comunque tutte le regole necessarie in questo tipo di movimenti. Questo framework consente quindi di effettuare studi di più larga scala rispetto a quanto fornito dai metodi di dinamica molecolare e giocando con i parametri che tengono in considerazione il clash possiamo fornire uno strumento con complessità computazionale minore.

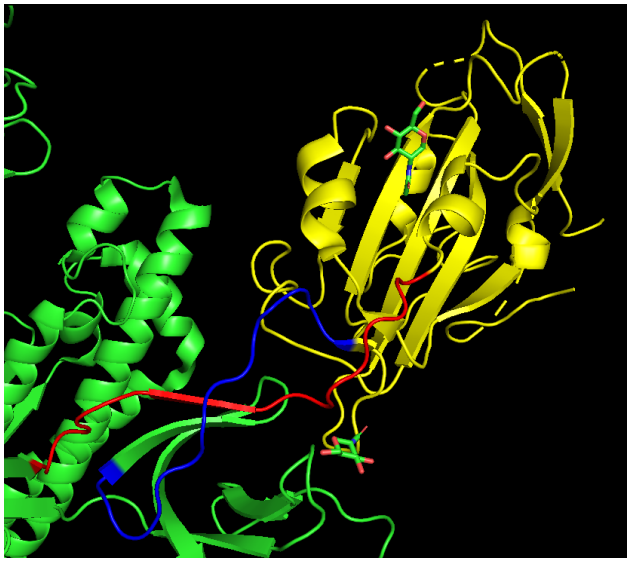

Figura 3.2: Nell'immagine dettagliata notiamo i due loop che collegano la parte mobile alla parte fissa colorati rispettivamente di rosso e di blu.

Capitolo 4

Progettazione

In questo capitolo viene affrontata la parte di implementazione e progettazione, vedremo le strategie di ricerca scelte, l'organizzazione delle strutture dati a supporto dell'esecuzione e poi avremo un approccio top-down verso il codice scritto. Introduco anche la libreria di supporto che ci ha permesso di svolgere il lavoro sul python.

4.1 Biopython

Il progetto Biopython è una libreria opensource [Jeff Chang,] di strumenti python per la biologia computazionale e la bioinformatica. Essa contiene classi per rappresentare sequenze biologiche e annotazioni di sequenze ed è in grado di leggere e scrivere file provenienti da diversi formati. Questa libreria mi ha permesso di utilizzare il parser per i file pdb (Protein Data Bank) al cui interno sono codificate le informazioni riguardanti gli atomi come nome dell'atomo posizione all'interno dello spazio tridimensionale ed altre informazioni. Oltre a fornirmi un parser, al suo interno sono presenti dei moduli che permettono di generare oggetti come le catene, i residui e gli atomi stessi. Questo da la possibilità di chiamare metodi specifici.

4.2 Strategie di ricerca

All'interno di questo framework è stato necessario usare tecniche di ricerca locale nella fase in cui si cerca di andare a ristabilire il collegamento dei due loop alla parte mobile una volta effettuata una rotazione/traslazione su di essa. In questo caso la tecnica utilizzata è la hill-climbing, in cui si cerca di trovare il massimo (o il minimo) di una funzione di costo attraverso la ricerca di soluzioni vicine a quella corrente. In questo caso specifico, la funzione di

ricerca cerca di ottimizzare la conformazione di un loop di collegamento alla volta attraverso la rotazione di alcuni angoli torsionali. Si parte da una configurazione iniziale del loop rappresentata mediante ad una lista di residui, e cerco di migliorarla ruotando uno dei due angoli torsionali a disposizione, partendo da un residuo specificato. L'angolo torsionale viene scelto casualmente tra $\phi o \psi$. Una volta applicata la rotazione viene restituita la nuova conformazione della proteina e si procede in questo modo fin tanto che non viene raggiunto il risultato desiderato.

Viene utilizzata anche una tecnica di ricerca per guidare il processo che porta al completamento del percorso più breve tra le due configurazioni. La strategia di ricerca, in questo caso, è una forma di shortest path con una coda di priorità implementata tramite heap, albero binario ordinato. Essa viene utilizzata per selezionare il nodo con il costo totale (cioè il costo del percorso finora più la stima del costo rimanente) più basso in ogni iterazione. In questo modo si cerca di espandere i nodi che hanno costo totale più basso per primi.

4.3 Strutture dati a supporto

Le strutture dati sono importanti all'interno di un algoritmo, non solo per permettere di conservare informazioni, ma incidono anche sul tempo di esecuzione del singolo programma. Organizzare quindi l'accesso alle strutture dati nel modo migliore permette di risparmiare tempo d'esecuzione oltre a semplificare nella maggior parte dei casi la comprensione della stessa.

Il linguaggio di programmazione scelto, python, necessità di studiare nel modo corretto come accedere alle strutture dati per evitare di trovarsi con dati inconsistenti poiché sono frutto di più modifiche. Per questo motivo e per quanto detto in precedenza io ho previsto 8 strutture di dati importanti: due dedicate a contenere la parte mobile e i loop, due dedicate al salvataggio delle posizioni acquisite da parte mobile e loop durante il cammino da una configurazione all'altra, due dedicate al salvataggio temporaneo e al ripristino in caso di clash delle coordinate dei loop pre e post rotazione e poi due dedicate alla gestione dei clash.

Le strutture dati sono state quasi tutte sviluppate come dei dizionari ad accesso chiave valore. Per quanto riguarda la struttura dati dedicata al controllo della presenza dei clash, la struttura contiene tuple di coordinate tridimensionali e per ciascuna tupla è contenuto la catena il residuo e l'atomo che si trovano in quella posizione. Questo è stato possibile mediante una discretizzazione di tutti gli atomi all'interno di una matrice tridimensionale. Chiaramente il passo di campionamento può essere impostato in qualunque

modo, di default è impostato a 5 Ångström. [1] Per quanto riguarda le strutture dati per mantenere salvate le coordinate dei loop e della parte mobile anch'esse sono organizzate come dizionari la cui chiave indicizza l'insieme delle varie coordinate. Le chiavi che indicizzano la struttura dati sono delle tuple contenenti la tripletta di traslazione effettuata più un numero in più che codifica il grado di rotazione della parte mobile. Attraverso metodi di get e set possiamo settarli nelle variabili coinvolte nel processo evitando problemi di condivisione della memoria. Abbiamo anche due dizionari dedicati al salvataggio temporaneo e al ripristino delle coordinate nel caso la rotazione non sia buona. Questi dizionari sono solamente per i due loop all'interno della procedura che si occupa di farli convergere alla parte mobile. La struttura dedicata ai loop e alla parte mobile invece sono due liste che a loro volta contengono residui a cui applicare le varie rotazioni. Voglio comunque porre enfasi sullo studio necessario per progettare la singola struttura che mi ha concesso non solo di risolvere il problema della condivisione della memoria, ma anche di velocizzare il processo di accesso alla struttura dati e quindi portare più velocità nell'esecuzione della ricerca locale.

4.4 Approccio Top-Down al codice

Ora vediamo quello che è lo speudo-codice che permette di compiere il lavoro all'algoritmo.

[1] Ångström, sono un'unità di misura, non SI, pari a $10^{-}10m$ è molto utilizzata nell'ambito molecolare.

```
1     parser = PDBParser(PERMISSIVE = True, QUIET = True)
2     covidchiusa = parser.getstructure(titleclosed)
3     covidaperta = parser.getstructure(titleopen)
4     calcoloparallelepipedo(covidchiusa, covidaperta)
5     loop = identifyloop(listacatene, catenainteresse,
      costantiA)
6     coil.append(loop)
7     loop = identifyloop(listacatene, catenainteresse,
      costantiB)
8     loopr = loop[::-1]
9     coil.append(loopr)
10
11    coildictcoordinates = [{"loop": x, "residui":[{"residuo
      ":y, "coordinate": []} for y in range(len(coil[coilname.
      index(x)]))]} for x in coilname]
12
13    for loop,idx in zip(coil, range(len(coil))):
14      salviamo(coildictcoordinates, loop)
15      salviamo(coildictrestorecoordinates, loop)
16
17    ptmob = partemobile(listacatene, catenainteresse)
18    storedptmob = salviamo(ptmob)
19    if (key not in posizioniacquisiteptmob):
20      insert(posizioniacquisite, key)
21
22    residuipartefissa = partefissa(listacatene,
      catenainteresse)
23    coordinate = salviamo(coildictcoordinates)
24    if (key not in posizioniacquisiteloop):
25      insert(posizioniacquisiteloop, key)
26
27
```

Codice 4.1: Fase d'inizializzazione del framework

Come si può vedere all'interno di questo primo spezzone di pseudo-codice si vanno ad inizializzare le componenti principali del sistema. Innanzitutto viene creato un parser che permette ha il compito di parsare il file pdb in ingresso. Dopodiché si effettua la lettura sia della configurazione chiusa che di quella aperta, questo è necessario perché come si vede nel prossimo passo dobbiamo generare un parallelepipedo che contenga entrambe le parti mobili.

```
1     def calcoloparallelepipedo(chiusa, aperta):
2     global variabili globali
3     x = []
4     y = []
5     z = []
6     mobilclosedpart = partemobile(listacatene,
      catenainteresse)
7     mobilopenedpart = partemobile(listacatene,
      catenainteresse)
8     model_open = aperta.get_models()
9
10    for mobil in [mobilclosedpart, mobilopenedpart]:
11      for residue in mobil:
12        for atom in residue.getatom():
13          diviamo le coordinate nei gli array degli assi
      cartesiani corrispondenti
14
15    for coordinata in [x, y, z]:
16      for indicatore in [max, min]:
17        assegnamo alla variabile globale del
      parallelepipedo la coordinata corretta
18
19
```

Codice 4.2: Calcolo del parallelepipedo

Il parallelepipedo che contiene entrambe le parti mobili è necessario per restringere il campo di ricerca, questo perché trovandoci in spazio aperto è fondamentale limitare lo spazio di ricerca ad una precisa zona altrimenti il processo non terminerebbe mai. Come detto nel capitolo 3 si lavora sulla configurazione chiusa e si tenta di convergere nella configurazione aperta, quindi tutto ciò che viene eseguito di seguito avviene sulla configurazione chiusa; la configurazione aperta è quindi solo necessaria per capire dove ci dirigiamo con la ricerca. Procedendo si passa ad individuare i loop, che mettono in collegamento la parte mobile con la parte fissa. I due loop sono simili, ma diversi nella direzione da loro presa, ovvero il loop-A si dirige dalla parte fissa alla parte mobile, mentre il loop-B lavora al contrario quindi per rendere omogeneo il comportamento del programma nei confronti dei due loop è necessario capovolgere il secondo. Questo non comporta nessun altra modifica, è però necessario ricordarsi di capovolgere il loop nel momento in cui si effettua il processo di scrittura file. Le costanti sul posizionamento dei loop sono state fornite insieme alle configurazioni ed è importante sottolineare che è stato preso un residuo in più in capo e in coda al loop, per assicurarsi

che ogni modifica si adattasse bene alla struttura a cui si deve attaccare. Si intravede l'organizzazione di una delle strutture citate in 4.3, essa è necessaria per salvare temporaneamente le coordinate dei loop, in modo tale che l'effetto di una rotazione può essere testato prima di diventare definitivo. Dopodiché si passa alla ricerca della parte mobile e della parte, in questo caso non sono necessarie le costanti, ma per identificarle andiamo ad esclusione rispetto a quelle utilizzate per i loop. Si intravedono anche le strutture dati dove sono conservati tutti gli step che portano da una configurazione all'altra.

```
1     if len(sys.argv) != 5 and len(sys.argv) > 1:
2       print("Devi passare cinque parametri allo script!")
3       sys.exit()
4     else:
5       iterazionimassime = 15000
6       printpdbtimeout = 25000
7       step = 5
8       if (len(sys.argv) == 5):
9         print("RICERCA SOLUZIONE SPECIFICA")
10      else:
11        if (len(sys.argv) == 1):
12          print("RICERCA DI POTENZIALI NUOVE SOLUZIONI")
13          costo, listachiavi = shortestpath((0,0,0,0))
14          for chiave in listachiavi:
15            Recuperiamo le coordinate e le stampiamo in un
    file di output
16
```

Codice 4.3: Modalità di esecuzione

Il framework può essere utilizzato in due modalità come si può vedere dal codice. In una modalità si può ricercare una soluzione specifica, ovvero è possibile dare un angolo di rotazione alla parte mobile una traslazione e il programma nel numero di iterazioni consentite cercherà la convergenza dei loop. Nell'altra modalità si va invece a cercare lo shortest-path tra le due configurazioni. Una volta terminata la ricerca del percorso vengono ricaricate tutte le posizioni acquisite e per ognuna di esse viene creato un file pdb apposito. Terminato il processo grazie all'uso di un altro script a disposizione è possibile generare un video per avere una prova grafica di cosa è accaduto durante la ricerca.

Capitolo 5

Risultati

Conclusione

Conclusione che riassume il lavoro svolto ed eventuali lavori futuri.

Bibliografia

[Alexa, 2022] Alexa, M. (2022). Super-fibonacci spirals: Fast, low-discrepancy sampling of so(3). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8291–8300.

[di Sanità EpiCentro L'epidemiologia per la sanità pubblica,] di Sanità EpiCentro L'epidemiologia per la sanità pubblica, I. S. Tutto sulla pandemia di sars-cov-2. Technical report.

[Duan Liangwei,] Duan Liangwei, Zheng Qianqian, Z. H. N. Y. L. Y. W. H. The sars-cov-2 spike glycoprotein biosynthesis, structure, function, and antigenicity: Implications for the design of spike-based vaccine immunogens.

[Federica Agosta,] Federica Agosta, Glen E. Kellogg, P. C. From oncoproteins to spike proteins: the evaluation of intramolecular stability using hydropathic force field.

[Jeff Chang,] Jeff Chang, Brad Chapman, I. F. T. H. M. d. H. P. C. T. A. E. T. B. W. *Biopython Tutorial and Cookbook*. Open Bioinformatics Foundation (OBF), Lyon, France.

[MALIK,] MALIK, Y. A. Properties of coronavirus and sars-cov-2.

[Martz,] Martz, E. The ramachandran principle. Technical report.

[Proteopedia,] Proteopedia. The ramachandran principle phi and psi angles in proteins. Technical report.

[Raffaele,] Raffaele, U. V.-S. S. Rna polimerasi, la "fotocopiatrice distratta" di sars-cov-2. Technical report.

[Tinelli,] Tinelli, M. La storia del sars-cov-2.

[Valério M,] Valério M, Borges-Araújo L, M. M. L. D. S. C. Sars-cov-2 variants impact rbd conformational dynamics and ace2 accessibility.

[Wikipedia,] Wikipedia. Forza intramolecolare. Technical report.

[WikipediaAmminoacidi,] WikipediaAmminoacidi. Amminoacidi. Technical report, Wikipedia.

[WikipediaProteine,] WikipediaProteine. Proteine. Technical report, Wikipedia.

Ringraziamenti

Innanzitutto voglio ringraziare il Professor Alessandro Dal palù che mi ha assistito durante tutto questo percorso e mi ha permesso di raggiungere questo traguardo. La ringrazio non solo per la cordialità e la pazienza che mi ha dimostrato, ma anche per avermi concesso la possibilità di mettere in pratica ciò che studiato in questi anni lavorando al progetto "Carriere studenti". Voglio ringraziare anche il Professor Pietro Cozzini e Federica Agosta senza i quali sarei ancora a cercare di capire com'è fatta la catena laterale di un amminoacido.

Un grazie immenso alla mia famiglia perché se tutto ciò sta avvenendo è anche grazie a voi. Mamma scusami per tutte le volte che ti risposto male e trattato peggio forse non sono molto bravo a gestire la pressione, però senza di te non ce l'avrei fatta. Papa se sono un minimo determinato a raggiungere i miei obbiettivi è grazie a te che mi hai insegnato a non mollare mai e non dare nulla per scontato, spero di continuare cosi e non mollare mai come fai tu. Let'z ce ne sarebbero troppe da dire, ma per qualsiasi cosa ti servirà ci sarò sempre, in fin dei conti siamo fratelli.

Alexa, da un anno e mezzo a questa parte i nostri cammini si sono intrecciati e fino ad ora tra alti e bassi è stato bellissimo; speriamo di poter condividere ancora tante altre avventure insieme. Grazie per essermi stata accanto sempre anche se so che non è stato facile.

Grazie a tutti i miei amici, quelle persone che ci sono sempre e che sempre ci saranno. Voglio però citare alcune persone in particolare con cui sono più legato nell'ultimo periodo, nonostante siate tutti molto importanti. Partendo da questi otto scappati di casa il cui unico obbiettivo è vandalizzarmi casa se salto un uscita. Ci sono però anche tanti bei momenti e notti magiche che non si possono dimenticare. Fede anche a te va un ringraziamento speciale per avermi ascoltato e tante volte risolto problemi che dal mio punto di vista sembravano insormontabili, la mia collega preferita. Marti, cosa possiamo dire, la mia seconda sorella aggiunta che ha passato un intero pomeriggio con me a guardare i legami covalenti, per non parlare di quella bellissima paperella gialla. Un ringraziamento speciale va anche a tutte quelle persone

con cui ho condiviso questo percorso sia la triennale che la magistrale, abbiamo condiviso tanto tempo insieme e tanta sofferenza, teniamoci in contatto perché siamo un bel gruppo. Ci sono poi gli amici di una vita per cui non ci sono parole.

Voglio ringraziare anche la mia classe di inglese che mi ha ascoltato durante tutte le lezioni e un grazie anche alla mia favorite teacher Kim che mi sta insegnando l'inglese con tanta pazienza perché diciamocelo non sono uno studente modello e si sa i compiti di inglese non vorrebbe farli nessuno.

Grazie a tutti.

Appendice A

Appendice di Esempio

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tri-

stique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.