Geometric deep learning and combinatorics for virus taxonomy classification

Giuseppe Pagano e Alessia Ture
Dipartimento di Informatica
Università degli Studi di Salerno

9 marzo 2024

Sommario

Questo studio esplora l'applicazione delle Graph Neural Networks (GNN) nella classificazione tassonomica dei virus, sfruttando l'innovativa rappresentazione dei dati genomici attraverso la Frequency Chaos Game Representation (FCGR) e la distanza di Hamming. L' obiettivo principale è stato quello di migliorare l'accuratezza nella classificazione dei virus, affrontando al contempo le sfide computazionali poste dalla vasta diversità e variabilità genetica dei virus. Sono state sviluppate e confrontate due tipologie di grafi, il "Grafo del Caos" e il "Grafo di Overlap", per analizzare le sequenze virali, evidenziando le differenze in termini di precisione, recall, F1 score e efficienza computazionale. I risultati dimostrano che, sebbene il "Grafo del Caos" offra un'alta precisione nella classificazione, richiede tempi di elaborazione notevolmente maggiori, rendendo il "Grafo di Overlap" una soluzione più bilanciata per applicazioni pratiche. Questo lavoro non solo apre nuove prospettive nell'uso delle GNN per lo studio dei virus ma pone anche le basi per ulteriori ricerche sull'ottimizzazione dei metodi di rappresentazione genomica e sull'efficacia delle GNN in bioinformatica.

1 Introduzione

I virus, elementi essenziali e ubiquitari degli ecosistemi terrestri, svolgono ruoli complessi che impattano profondamente sulla salute umana e animale e sull'equilibrio ecologico globale. Nonostante la loro apparente semplicità strutturale, i virus manifestano una diversità genetica e funzionale straordinaria, ponendo significative sfide scientifiche e mediche nella loro comprensione, diagnosi e trattamento. In questo intricato panorama, la classificazione tassonomica dei virus emerge come uno strumento fondamentale per sistematizzare la vasta diversità virale, fornendo un quadro ordinato che facilita l'identificazione, lo studio e la gestione delle infezioni virali.

La tassonomia è la disciplina che si occupa della classificazione gerarchica di elementi viventi o inanimati. L'esempio tipico è la tassonomia biologica, ossia i criteri con cui si classificano gli organismi in una gerarchia di taxa annidati con cui si può per esempio risalire alla loro evoluzione. Con il termine tassonomia, dunque, ci si può riferire sia

alla classificazione gerarchica di concetti, sia al principio stesso della classificazione. Praticamente tutti i concetti, gli oggetti animati e non, i luoghi e gli eventi, possono essere classificati seguendo uno schema tassonomico.

La classificazione dei virus è il processo di denominazione dei virus e di collocazione in un sistema tassonomico simile ai sistemi di classificazione utilizzati per gli organismi cellulari.

I virus sono classificati in base alle caratteristiche fenotipiche, come la morfologia, il tipo di acido nucleico, la modalità di replicazione, gli organismi ospiti e il tipo di malattia che causano. La classificazione tassonomica formale dei virus è responsabilità del sistema dell'International Committee on Taxonomy of Viruses (ICTV).

La tassonomia ha una lunga storia nella biologia, fornendo un sistema per organizzare la diversità della vita in categorie gerarchiche basate su relazioni evolutive, morfologiche e genetiche. Nel contesto dei virus, la tassonomia assume un ruolo critico, non solo per comprendere la loro biologia e evoluzione ma anche per sviluppare strategie efficaci di prevenzione, diagnosi e trattamento delle malattie virali. Tuttavia, la classificazione tassonomica dei virus presenta sfide uniche a causa della loro elevata variabilità genetica e della continua scoperta di nuovi virus, spesso con caratteristiche che sfidano le categorie esistenti.

La classificazione e l'identificazione della tassonomia dei virus sono compiti fondamentali nella virologia, con implicazioni dirette per la comprensione della patogenesi virale, la sorveglianza epidemiologica, lo sviluppo di terapie antivirali e la prevenzione delle malattie infettive. La diversità genetica e la capacità di rapida mutazione dei virus presentano sfide significative per la tassonomia tradizionale, che si basa su caratteristiche fenotipiche e relazioni evolutive. L'espansione dei dati genomici ha offerto nuove opportunità per affrontare queste sfide, consentendo approcci basati sulla genetica per la classificazione e l'analisi filogenetica dei virus.

La tassonomia, che tradizionalmente si basa su caratteristiche fenotipiche e relazioni evolutive, ha trovato nella moderna bioinformatica un alleato prezioso. L'avanzamento delle tecniche di sequenziamento genomico ha rivoluzionato l'approccio alla classificazione virale, permettendo di affrontare la sfida della loro elevata variabilità genetica con metodologie basate sulla genetica. Questo approccio ha ampliato notevolmente la nostra capacità di distinguere e classificare i virus, superando i limiti imposti dai metodi tradizionali.

In questo contesto dinamico, le Graph Neural Networks (GNN) rappresentano un'innovazione promettente, come strumenti avanzati per l'analisi dei dati strutturati in forma di grafi. Le GNN sono particolarmente adatte per catturare le complesse relazioni e i modelli ricorrenti all'interno dei dati genomici dei virus. Questa tecnologia ha il potenziale non solo per migliorare la precisione della classificazione tassonomica dei virus ma anche per offrire nuove intuizioni sulle loro reti evolutive e funzionali, con implicazioni significative per la ricerca virologica e l'epidemiologia. Tuttavia, l'integrazione efficace delle GNN nella tassonomia virale presenta sfide significative. Queste includono la necessità di rappresentazioni grafiche accurate dei dati genomici, la selezione di architetture GNN ottimali in grado di gestire la diversità e la complessità dei virus e l'interpretazione dei

modelli complessi generati dalle reti. Inoltre, la grande variabilità genetica dei virus e la loro continua evoluzione richiedono approcci flessibili e adattabili che possano tenere il passo con il ritmo veloce delle scoperte virali.

Il progetto mira a esplorare il potenziale delle GNN nel rivoluzionare la classificazione tassonomica dei virus attraverso approcci innovativi nella costruzione e nell'analisi dei grafi genomici.

1.1 Classificazione dei Virus

I virus, esseri microscopici con una notevole diversità genetica, sono stati tra i primi organismi a essere sottoposti a sequenziamento del genoma completo [2]. Questo approccio ha portato a rivoluzionari cambiamenti nella virologia, spostando l'attenzione verso la genomica comparativa, un campo che si affida all'analisi computazionale per scoprire pattern evolutivi e funzionali nei genomi virali. Gli strumenti software avanzati, sviluppati per assistere i virologi in queste indagini, hanno facilitato una comprensione più profonda della struttura, funzione ed evoluzione dei virus, arricchendo il nostro sapere sulla biologia virale.

La genomica comparativa ha inoltre giocato un ruolo cruciale nel modellare la tassonomia dei virus, proponendo una struttura gerarchica che integra le nomenclature dei taxa [5]. Questa iniziativa, avviata oltre cinquant'anni fa, ha cercato di organizzare sistematicamente la crescente mole di conoscenze sui virus, un compito affidato al Comitato Internazionale sulla Tassonomia dei Virus (ICTV). L'ICTV ha il compito unico, nel vasto campo della biologia, di prendere decisioni su tutti gli aspetti legati alla classificazione e alla nomenclatura dei virus, stabilendo i criteri per la definizione delle varie categorie tassonomiche [1].

Inizialmente, la classificazione dei virus era limitata a pochi livelli gerarchici, principalmente a causa del ridotto numero di virus conosciuti e del loro elevato tasso di mutazione. Tuttavia, l'espansione della genomica comparativa e l'analisi filogenetica hanno sottolineato la possibilità di costruire una classificazione che rifletta l'evoluzione dei virus, analogamente a quanto fatto per gli organismi cellulari. L'introduzione del sequenziamento genomico ad alto rendimento ha ulteriormente accelerato questa transizione, permettendo una caratterizzazione genomica su larga scala e rivelando la complessa diversità virale presente negli ecosistemi biologici [3].

Questo progresso ha spinto l'ICTV ad adottare la genomica comparativa come principale fondamento per la definizione dei taxa virali, abbracciando una struttura tassonomica che riflette l'albero evolutivo dei virus. Tale approccio non solo ha arricchito la classificazione tassonomica, ma ha anche posto le basi per una comprensione più sistematica e dettagliata della virosfera, offrendo nuove prospettive sulla relazione tra virus e ospiti, sui meccanismi di trasmissione e sulla dinamica evolutiva dei virus nel contesto degli ecosistemi globali.

Gli approcci computazionali alla tassonomia dei virus, come delineato in [5], si basano su una varietà di metodi che riflettono la complessità e la diversità del mondo virale. Questi metodi includono:

1. **Modelli di Variazione Sequenziale:** Questi approcci si concentrano sulle differenze nelle sequenze di nucleotidi o aminoacidi, cercando modelli e variazioni specifiche che possano distinguere i virus tra loro. La variazione sequenziale può offrire intuizioni sulla funzione e l'evoluzione dei virus, nonché sulla loro relazione tassonomica.

2. **Omologia di Contenuto Genetico o Proteico:** Questa metodologia si basa sul confronto del contenuto genetico o proteico dei virus per identificare somiglianze e differenze significative. L'omologia, o la somiglianza genetica dovuta alla discendenza da un antenato comune, è un concetto chiave in questo approccio, che può aiutare a definire le relazioni evolutive tra i virus.

3. **Filogenia:** Gli alberi filogenetici sono costruiti per rappresentare le relazioni evolutive tra i virus, basandosi su sequenze genetiche o proteiche. La filogenia fornisce un quadro visuale di come i virus si siano evoluti nel tempo e come siano collegati tra loro, supportando la classificazione tassonomica.

4. **Distanza Genetica a Coppie:** Questo approccio misura la distanza genetica, o il grado di differenza, tra coppie di sequenze virali. Le distanze genetiche possono essere utilizzate per raggruppare i virus in categorie tassonomiche, basate sulla loro somiglianza genetica.

Nonostante la diversità di questi approcci, spesso vengono combinati per ottenere una comprensione più completa della tassonomia dei virus. Tuttavia, la scelta delle metriche può essere influenzata da fattori pratici, come il tipo di virus studiato o le specificità dei dati disponibili.

Un metodo relativamente semplice si basa sull'analisi statistica delle sequenze, come l'abbondanza di particolari combinazioni di nucleotidi (ad esempio, il contenuto di G+C) o il conteggio di tutti i possibili oligonucleotidi o oligopeptidi di una certa lunghezza (metodi K-mer). Questi metodi possono essere utilizzati per classificare i virus, ma possono essere influenzati da fattori come il campionamento irregolare dei virus, la ricombinazione, la convergenza evolutiva e i tassi di evoluzione specifici per lignaggio o regione. Inoltre, l'interpretazione delle differenze rilevate attraverso i metodi K-mer può essere complessa e talvolta non immediatamente intuitiva.

2 Background

Questa sezione fornisce una panoramica delle conoscenze preliminari necessarie per comprendere gli approcci utilizzati in questo studio. In particolare, si concentra sulla Frequency Chaos Game Representation (FCGR) e sulla Distanza di Hamming, due concetti chiave utilizzati nella costruzione dei grafi per l'analisi con Graph Neural Networks (GNN).

2.1 Frequency Chaos Game Representation (FCGR)

Chaos Game Representation (CGR) è una metodologia innovativa che trasforma sequenze lineari di DNA in rappresentazioni grafiche bidimensionali. Ogni nucleotide (Adenina, Timina, Citosina e Guanina) è associato a un vertice di un quadrato e la sequenza viene percorsa per generare un pattern unico. Questa trasformazione non solo offre una nuova prospettiva visuale sulla composizione nucleotidica, ma rivela anche modelli nascosti all'interno delle sequenze genetiche, che possono essere difficilmente percepibili attraverso metodi di rappresentazione lineare.

Jeffrey [4], nel 1990, ha introdotto l'applicazione della CGR al DNA, aprendo la strada a vasti campi di applicazione nella bioinformatica. CGR, come mostrato nell'articolo [7] ha dimostrato di possedere proprietà uniche che la rendono particolarmente adatta per l'analisi genomica:

- Ogni sequenza genetica viene mappata a un pattern unico, permettendo l'identificazione visuale di specifiche caratteristiche genomiche.
- La CGR rappresenta tutte le possibili sequenze a qualsiasi lunghezza in uno spazio bidimensionale o tridimensionale, offrendo una visualizzazione completa delle informazioni genetiche.
- La capacità della CGR di codificare l'intera sequenza in una singola coordinata finale la rende uno strumento efficace per il confronto di sequenze e l'analisi filogenetica.

Fondamentalmente, l'intero insieme di frequenze delle parole trovate in una data sequenza genomica può essere visualizzato sotto forma di una singola immagine in cui ogni pixel è associato a una parola specifica, come mostrato nella figura 1.

Per via delle sue proprietà, il CGR è già stato utilizzato, ad esempio, nel confronto delle sequenze senza allineamento, nella filogenesi e come codifica per l'apprendimento automatico, e ha anche un enorme potenziale per applicazioni future in bioinformatica. [7]

Oltre a quelle proprietà già menzionate, il CGR ha alcune proprietà uniche aggiuntive. Il CGR è una rappresentazione di tutte le possibili sequenze in qualsiasi lunghezza in uno spazio continuo. Può essere considerato come una generalizzazione di un modello di Markov (lo stato successivo dipende dallo stato corrente), e la sequenza completa può essere ricostruita esclusivamente dalle ultime coordinate del CGR.

La figura 3 mostra tre esempi di modelli CGR per diverse sequenze genomiche di organismi rispetto a una sequenza casuale. Questo modello è l'attrattore della sequenza, cioè i punti o le aree a cui un sistema dinamico converge. Il CGR è stato anche usato per esaminare visivamente la qualità dei generatori di numeri casuali. Per i generatori di numeri casuali di alta qualità, non emergeranno modelli visibili nel CGR [7].

Mentre l'approccio CGR originale di Jeffrey è stato sviluppato per il DNA, diversi approcci per le proteine sono stati sviluppati in seguito. Dopo che l'algoritmo originale

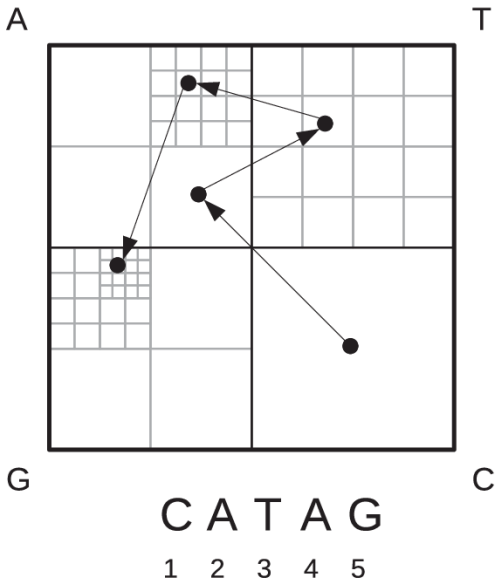

Figura 1: Esempio di costruzione di CGR per la sequenza CATAG

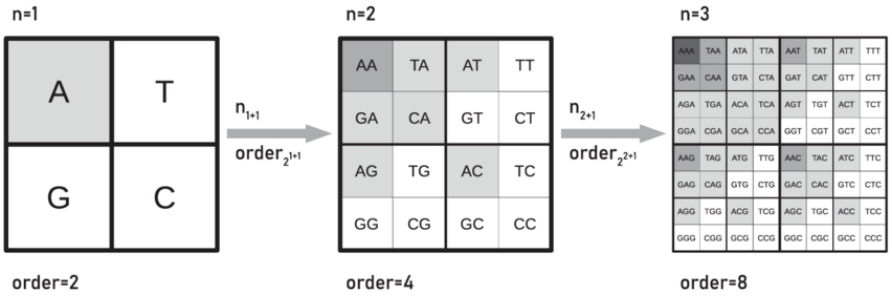

Figura 2: Matrice FCGR al crescere di n

basato su un triangolo di Sierpinski è stato adattato in un quadrato per prendere in considerazione i quattro nucleotidi del DNA da Jeffrey, sono stati fatti sforzi per estendere l'algoritmo per i 20 aminoacidi per costruire CGR di proteine.

La Frequency Chaos Game Representation (FCGR) estende la CGR incorporando la frequenza dei k-mers, sottosequenze di lunghezza k presenti nel DNA. Questo arricchimento permette di convertire le sequenze di DNA in matrici di frequenza, dove ogni cella rappresenta la frequenza relativa di un k-mer specifico all'interno della sequenza. Le matrici FCGR offrono una rappresentazione densa e informativa delle sequenze, evidenziando la presenza, l'assenza o la rarità di certi k-mers, che possono indicare regioni genomiche di particolare interesse biologico o evolutivo.

Le frequenze delle parole trovate in una sequenza vengono visualizzate in un'immagine quadrata, con la posizione di una determinata parola scelta secondo una procedura

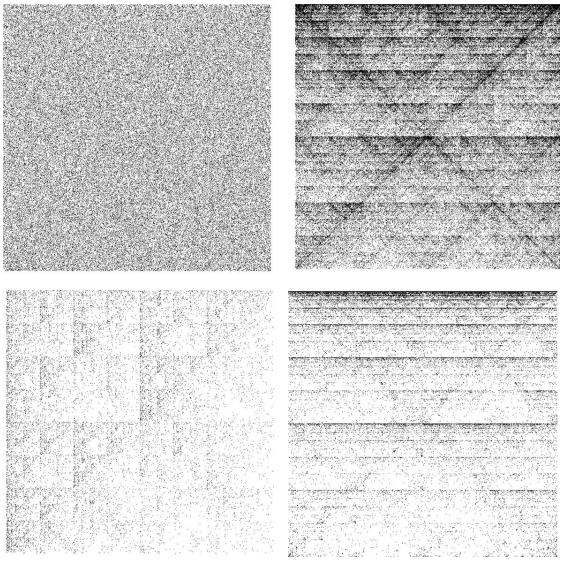

Figura 3: Creazione di immagini a partire dalla matrice FCGR

ricorsiva. Pertanto, l'immagine è divisa in quattro quadranti in cui vengono raccolte le sequenze che terminano con la base appropriata. Questo dà la composizione di base della sequenza. Ogni quadrante è successivamente diviso in quattro sotto quadranti, ciascuno contenente sequenze che terminano con un dato di nucleotide , in modo tale che le sequenze differiscono solo nel primo letterale nei sotto quadranti adiacenti. Le frequenze delle parole sono visualizzate dall'intensità di ogni pixel [6]. Il processo iterativo che porta alla formazione della matrice FCGR data la grandezza del k-mer è rappresentato nella figura 2

Nel contesto del progetto, la FCGR è stata utilizzata per trasformare le sequenze virali in una forma che è facilmente elaborabile dalle Graph Neural Networks (GNN). Questo approccio permette di:

- Visualizzare e confrontare complesse sequenze virali, facilitando l'identificazione di somiglianze e differenze a livello genomica.
- Costruire grafi informativi dove i nodi rappresentano k-mers significativi e i loro collegamenti riflettono le relazioni strutturali e funzionali tra le sequenze.

La Frequency Chaos Game Representation (FCGR) rappresenta un notevole progresso nel campo della bioinformatica, specialmente per l'analisi delle sequenze di DNA, inclusi quelli dei virus. Questo metodo fornisce una visualizzazione unica e compatta che trasforma sequenze lineari in rappresentazioni grafiche, evidenziando modelli e caratteristiche che potrebbero non essere immediatamente evidenti con approcci tradizionali. Tuttavia, come ogni metodologia, la FCGR porta con sé una serie di sfide e limitazioni che necessitano di attenzione per massimizzare il suo potenziale e l'affidabilità dei risultati ottenuti.

L'approccio Frequency Chaos Game Representation (FCGR) si basa sull'analisi dei k-mer, che comporta l'identificazione e la quantificazione di tutte le possibili sotto sequenze di lunghezza k in una sequenza di DNA. Sebbene questa tecnica fornisca un'abbondante quantità di informazioni, presenta diverse sfide, specialmente con l'aumentare della lunghezza di k, che possono complicare la sua implementazione a causa dell'esponenziale crescita del numero di possibili combinazioni di k-mer.

Una delle principali difficoltà nel conteggio dei k-mer è la gestione dell'ingente volume di dati generato, particolarmente evidente quando si lavora con genomi di grandi dimensioni o dati metagenomici complessi. La necessità di elaborare un numero esponenziale di combinazioni possibili richiede notevoli risorse computazionali e di archiviazione. L'aumento della grandezza dei k-mer di un unità implica un aumento stimato del carico computazionale tra il 300% ed il 400%, raggiungendo anche con valori di k considerabili modesti tempi necessari lunghi settimane.

Un ulteriore ostacolo è rappresentato dagli errori di sequenziamento, che possono generare k-mer inaccurati non presenti nel genoma originale. Questo problema richiede di distinguere accuratamente tra k-mer reali ed errori di sequenziamento, un compito che può diventare particolarmente arduo in presenza di alti tassi di errore.

La ricombinazione genetica, frequente in molti virus, introduce un'altra sfida, poiché può modificare sostanzialmente la composizione genomica e, di conseguenza, la distribuzione dei k-mer, complicando l'analisi e l'interpretazione dei dati.

Per affrontare l'elaborazione su larga scala dei k-mer, sono spesso necessarie strategie di computazione parallela e di ottimizzazione della memoria. L'adozione di algoritmi efficienti dal punto di vista della memoria e la capacità di sfruttare le architetture di calcolo parallelo sono cruciali per minimizzare i tempi di calcolo e rendere fattibile l'analisi dei k-mer su set di dati genomici estesi, anche l'uso di filtri di Bloom si è rivelato una strategia efficace. I filtri di Bloom sono strutture dati probabilistiche che consentono di memorizzare e interrogare la presenza di elementi in un set con un'elevata efficienza di spazio e tempo, al costo di una possibilità controllabile di falsi positivi. Questa caratteristica li rende particolarmente adatti per applicazioni in cui lo spazio di archiviazione è un vincolo critico e dove un piccolo tasso di errore è accettabile. L'uso dei filtri di Bloom nell'elaborazione dei k-mer non solo migliora l'efficienza della memoria ma può anche facilitare la parallelizzazione delle computazioni. Poiché i filtri di Bloom supportano operazioni di inserimento e query indipendenti, è possibile distribuire il processo di conteggio dei k-mer su più unità di elaborazione parallele, accelerando ulteriormente l'analisi. [8]

Oltre a queste sfide tecniche, la metodologia FCGR deve considerare la possibile presenza di sequenze ripetitive e la sovrarappresentazione di determinati k-mer, che possono influenzare l'equilibrio della matrice FCGR e, di conseguenza, l'interpretazione dei risultati. Inoltre, la sensibilità del metodo FCGR a variazioni casuali e rumore nelle sequenze di DNA può rendere più complessa l'identificazione di pattern significativi e la distinzione tra caratteristiche biologicamente rilevanti e artefatti metodologici.

Infine, l'interpretazione dei dati prodotti dalla FCGR rappresenta una sfida notevole. Sebbene le rappresentazioni grafiche offrano una panoramica dettagliata delle sequenze,

trasformare queste informazioni in conoscenze biologicamente significative non è sempre immediato e richiede una profonda comprensione sia della metodologia sia del contesto biologico. È fondamentale integrare le rappresentazioni FCGR con altre fonti di dati per acquisire una comprensione olistica e approfondita delle sequenze analizzate.

2.2 Distanza di Hamming

La Distanza di Hamming è una metrica che quantifica la differenza tra due stringhe di uguale lunghezza, misurando il numero minimo di sostituzioni necessarie per trasformare una stringa nell'altra. Nel contesto della bioinformatica, questa distanza è particolarmente utile per confrontare sequenze genetiche, offrendo una misura della loro diversità o somiglianza a livello molecolare. é concettualmente semplice e computazionalmente efficiente, permettendo di calcolare rapidamente le distanze tra un grande numero di sequenze. Questa efficienza è cruciale nel trattare i vasti set di dati genetici tipici della bioinformatica, dove è comune confrontare migliaia o anche milioni di sequenze.

La Distanza di Hamming è fondamentale in bioinformatica per confrontare sequenze genetiche, quantificando le differenze tra stringhe di DNA o RNA di uguale lunghezza. Essa misura il numero minimo di cambiamenti necessari per trasformare una sequenza nell'altra, fornendo una misura diretta della loro somiglianza o divergenza. Questa metrica si rivela utile nelle analisi filogenetiche, nell'identificazione di varianti genetiche e nella valutazione della variabilità genetica all'interno delle popolazioni [9].

2.3 Il Grafo del Caos

Il "Grafo del Caos" emerge come un'innovativa fusione tra la Frequency Chaos Game Representation (FCGR) e la Distanza di Hamming, offrendo una struttura dati ricca e multidimensionale per analisi avanzate con le Graph Neural Networks (GNN). Questo approccio sfrutta la rappresentazione visiva e la frequenza dei k-mers all'interno delle sequenze virali, insieme alla loro somiglianza genetica, per costruire un grafo che rifletta le intricate relazioni tra diverse sequenze. Di seguito è riportata una descrizione dettagliata e migliorata dei passaggi chiave nella costruzione di un Grafo del Caos:

1. **Costruzione della Matrice FCGR:** Ogni sequenza virale viene trasformata in una matrice FCGR, rappresentando visivamente la frequenza dei k-mers all'interno della sequenza.

2. **Selezione dei Nodi:** I k-mers rappresentati nelle matrici FCGR con frequenza maggiore di zero vengono selezionati come nodi del grafo, enfatizzando le sotto sequenze genetiche significative.

3. **Connessione dei Nodi:** I nodi vengono collegati basandosi sulla Distanza di Hamming, stabilendo legami tra k-mers simili e riflettendo la loro vicinanza genetica o funzionale.

I virus sono noti per la loro elevata tasso di mutazione e per i fenomeni di ricombinazione. La Distanza di Hamming è particolarmente utile in questo contesto perché può catturare le variazioni puntiformi che spesso caratterizzano l'evoluzione virale, fornendo una misura quantitativa della divergenza genetica tra i ceppi virali. Integrando la Distanza di Hamming nella costruzione del grafo, si possono sfruttare le potenti capacità delle GNN per l'analisi di dati strutturati come grafi. Le GNN possono apprendere rappresentazioni complesse dei dati e identificare pattern nascosti nelle relazioni tra sequenze, offrendo nuove intuizioni nella biologia virale e nelle interazioni ospite-patogeno.

La Figura 7 illustra un esempio concreto di come il Grafo del Caos viene costruito a partire da una sequenza virale di esempio, con una distanza di Hamming fissata a 2 e k-mers di lunghezza 4. Questa rappresentazione grafica non solo facilita l'analisi visuale delle relazioni tra sequenze, ma apre anche la porta a metodologie di analisi computazionale avanzate, permettendo l'esplorazione di pattern complessi e la predizione di nuove caratteristiche virali mediante l'impiego di GNN.

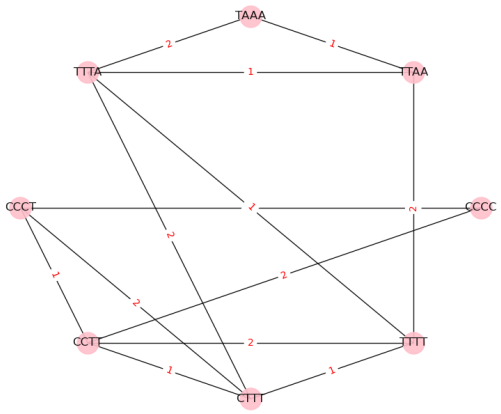

Figura 4: Costruzione del grafo del caos su una sequenza d'esempio

Il Grafo del Caos, con la sua capacità di sintetizzare e rappresentare complessi insiemi di dati genomici, si pone come uno strumento potente e versatile nel campo della bioinformatica e della virologia computazionale, promettendo nuove frontiere nell'analisi e nella comprensione delle malattie virali.

2.4 Il Grafo di Overlap

Il grafo di overlap si distingue come strumento essenziale per disvelare le complesse relazioni tra sequenze di DNA, RNA o peptidi, facilitando la comprensione della disposizione originale delle sequenze all'interno del genoma. Attraverso la rappresentazione di sovrapposizioni significative tra le sequenze, questo grafo fornisce una mappa dettagliata

che aiuta a decifrare l'organizzazione genetica e le interconnessioni funzionali tra diversi frammenti di sequenze virali.

La costruzione di un grafo di overlap segue questi passaggi essenziali:

1. **Identificazione dei Nodi:** Ogni sequenza o frammento di sequenza viene considerato come un nodo nel grafo.

2. **Determinazione degli Archi:** Per ogni coppia di nodi, si esamina la presenza di una sovrapposizione significativa tra la fine di una sequenza e l'inizio dell'altra. Se una tale sovrapposizione esiste e supera una soglia di lunghezza predefinita, si aggiunge un arco tra i due nodi corrispondenti.

3. **Ponderazione degli Archi**: Gli archi possono essere ponderati in base alla lunghezza della sovrapposizione

Nella figura 6 è presente un esempio della costruzione del grafo su una sequenza d'esempio con lunghezza dei k-mer pari a 5 e tolleranza di sovrapposizione maggiore o uguale a 4.

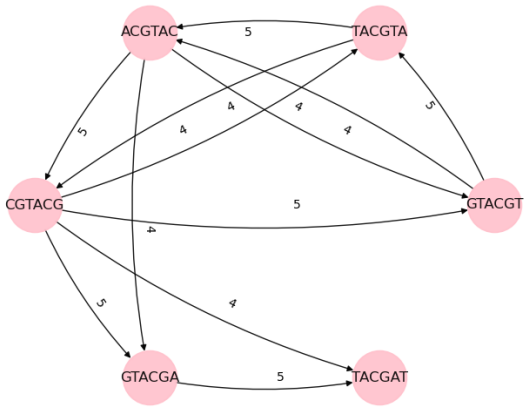

Figura 5: Costruzione del grafo di overlap su una sequenza d'esempio

Il grafo di overlap non è soltanto un'astrazione teorica; esso si rivela uno strumento potente per svariati ambiti di applicazione nella ricerca virale, inclusa la ricostruzione di genomi e l'analisi comparativa di sequenze virali. Grazie alla sua capacità di mappare dettagliatamente le sovrapposizioni, il grafo di overlap facilita:

- La Ricostruzione Accurata dei genomi, permettendo agli scienziati di assemblare sequenze genomiche complete a partire da frammenti isolati, un'operazione cruciale per comprendere la struttura e la funzione dei virus.

- Analisi Comparative approfondite, che aiutano a identificare somiglianze e divergenze tra differenti ceppi virali, offrendo spunti vitali per lo studio della loro evoluzione e della loro diffusione.

2.5 Il Grafo di de Bruijn

Il Grafo di de Bruijn gioca un ruolo pivotale nell'analisi computazionale di sequenze genetiche, specialmente nell'era del big data genetico. Tradizionalmente impiegato nell'assemblaggio di genomi, questo grafo trova una nuova e promettente applicazione nell'ambito della classificazione di read virali tramite Graph Neural Networks (GNN). Caratterizzato dalla sua capacità di decomporre complesse sequenze genomiche in k-meri e di rappresentarne le sovrapposizioni, il Grafo di de Bruijn sintetizza efficacemente l'informazione contenuta nelle sequenze virali, trasformandola in una struttura facilmente analizzabile da modelli basati su GNN. Questi ultimi, sfruttando la struttura del grafo, sono in grado di identificare e apprendere relazioni complesse tra i k-meri, consentendo una classificazione accurata e dettagliata dei read virali in base a caratteristiche genomiche distintive. L'impiego del Grafo di de Bruijn nell'ambito delle GNN offre diversi vantaggi, tra cui:

- **Efficienza Computazionale**: Riducendo la complessità delle sequenze genomiche a una struttura di grafo, si facilitano operazioni computazionali che altrimenti sarebbero proibitive a causa della grandezza del genoma
- **Rilevamento di Pattern Genomici**: La natura del grafo permette di evidenziare sovrapposizioni e ripetizioni all'interno delle sequenze, elementi chiave nella classificazione e nell'analisi filogenetica dei virus.
- **Applicabilità Versatile:** Oltre alla classificazione, il Grafo di de Bruijn trova applicazione in numerosi altri ambiti della virologia computazionale, inclusi l'identificazione di nuovi patogeni e lo studio della loro evoluzione.

Tuttavia, l'utilizzo di questa struttura presenta anche delle sfide:

- **Gestione della Complessità**: La costruzione e l'analisi di grafi di de Bruijn per sequenze genomiche estese richiedono risorse computazionali significative, specialmente per quanto riguarda la memoria e il tempo di elaborazione.
- **Sensibilità agli Errori di Sequenziamento**: Le inaccurattezze nei dati di sequenziamento possono portare alla formazione di archi errati nel grafo, influenzando la qualità dell'analisi.
- **Interpretazione dei Risultati**: Mentre le GNN possono offrire predizioni accurate, l'interpretazione biologica di questi risultati può essere non immediata, richiedendo un'analisi approfondita e una valida comprensione del contesto genomico.

L'integrazione del Grafo di de Bruijn con le GNN rappresenta una frontiera innovativa nella classificazione di read virali. Questa sinergia tra bioinformatica e apprendimento automatico non solo potenzia la nostra capacità di interpretare le immense quantità di dati genomici generati dalle moderne tecnologie di sequenziamento, ma apre anche la strada a nuove scoperte nella ricerca virologica.

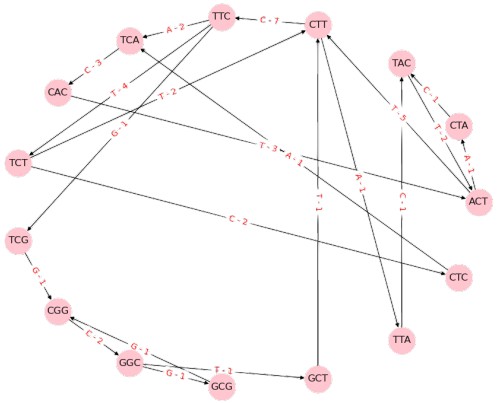

Figura 6: Costruzione del grafo di de Bruijn su una sequenza d'esempio

3 Materiali e Metodi

3.1 Dataset

I dati per la creazione dei dataset provengono dal **NCBI** (National Center for Biotechnology Information), che tramite una API chiamata Entrez permette il download di sequenze in formato fasta dalla proprie banche dati [10].
Ogni file in formato fasta viene diviso nei singoli record che lo compongono ed una query per ogni singolo record è fatta all'API Entrez. per ottenerne la tassonomia. I record, ai quali è stata aggiunta la tassonomia come label, sono poi trasformati in un grande dataframe pandas che viene successivamente diviso in dataset di train, validation e test.

3.2 Costruzione dei grafi

Per la costruzione dei grafi viene seguito un approccio ad hoc per ogni tipo di grafo che si vuole costruire. Nel caso dei grafi di overlap ogni nodo è identificato come un kmer di lunghezza k, le cui feature sono le basi azotate che compongono il kmer stesso in rappresentazione one hot, Gli archi collegano invece nodi con un overlap significativo (di lunghezza superiore ad una soglia predefinita) ed hanno come feature la lunghezza dell'overlap. Nel caso dei grafi del caos invece i nodi sono tutti gli elementi della matrice fcgr con un valore diverso da 0 mantenendo anche in questo caso le basi azotate che compongono il kmer stesso in rappresentazione one hot come feature. Gli archi invece collegano nodi abbastanza simili, dove la somiglianza è codificata come una distanza di hamming inferiore ad una determinata soglia. Indipendentemente dal tipo di grafo utilizzato, il risultato finale, rappresentato da una serie di grafi pytorch geometric vengono salvati infinite in memoria tramite Pickle.

Tabella 1: Iperparametri usati per l'addestramento del modello.

hidden	embedding	embedding_mlp	layers
256	64	128	1

3.3 Architettura della GNN

Il modello utilizzato è stato il modello presentato nel paper [11], che utilizza una serie di layer che applicano DiffPool per l'estrazione di feature gerarchiche a partire dal grafo, alternandoli a layer di convoluzione e a layer GRAPHSAGE per la creazione delle matrici di embedding. Dopo ogni layer GRAPHSAGE viene applicata batch normalization.

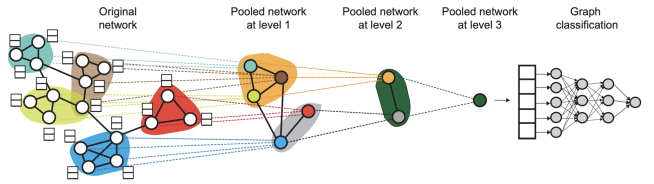

Figure 1: High-level illustration of our proposed method DIFFPOOL. At each hierarchical layer, we run a GNN model to obtain embeddings of nodes. We then use these learned embeddings to cluster nodes together and run another GNN layer on this coarsened graph. This whole process is repeated for L layers and we use the final output representation to classify the graph.

Figura 7: Funzionamento del metodo DiffPool

3.4 Training del Modello

Per il training del modello si è deciso di operare su read di grandezza 250 con overlap di grandezza 200, la batch size utilizzata è stata di 512 ed è stata effettuata una ricerca comprensiva per la ricerca del parametro k ottimale. La funzione di loss prescelta è stata la CrossEntropyLoss e sono stati provati 2 diversi ottimizzatori: **Adagrad** e **AdamW** con AdamW che si è rivelato la scelta superiore delle due. Il limite massimo delle epoche è stato fissato a 1000, con un early-stopping dopo 30 epoche senza miglioramenti.

4 Analisi dei Risultati

L'indagine condotta sulle performance dei grafi di overlap, del caos e di de Bruijn nel contesto della classificazione di read virali rivela insights cruciali relativi al bilanciamento tra precisione delle classificazioni e costi computazionali. L'analisi dei risultati mostra che con l'aumento della lunghezza del k-mer nel grafo del caos, la precisione migliora significativamente, passando dal 54% con un k-mer di lunghezza 5 al 67% con una lunghezza di 7, e ulteriormente al 73% con k-mer di lunghezza 9. Questo incremento, tuttavia, comporta un notevole aumento del tempo di elaborazione, passando da 56 secondi a 619 secondi con k-mer di lunghezza 7, e raggiungendo 9629 secondi

per k-mer di lunghezza 9. Questo suggerisce un compromesso tra precisione e efficienza computazionale, dove l'uso di k-mer più lunghi migliora la specificità ma richiede tempi di elaborazione significativamente maggiori.

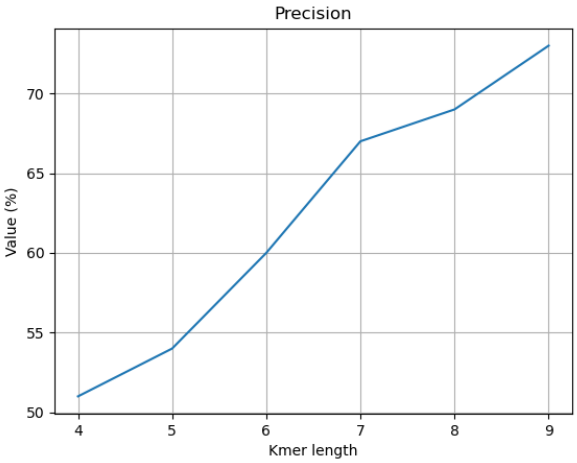

Figura 8: Precisione al crescere della lunghezza del k-mer, per il grafo del caos

Analogamente, per quanto riguarda la recall, l'accuratezza e l'F1 Score, si registra un incremento progressivo, con valori che salgono dal 54% al 73%, seguendo la stessa tendenza osservata per la precisione. Questa correlazione diretta tra la lunghezza del k-mer e le performance del modello suggerisce che l'accuratezza della classificazione può essere sensibilmente migliorata estendendo la lunghezza del k-mer, sebbene ciò comporti un aumento esponenziale dei tempi di elaborazione.

Basandoci sull'adattamento esponenziale dei dati, il tempo previsto per l'elaborazione di un k-mer di lunghezza 10 è di circa 9.26 ore. Per un k-mer di lunghezza 15, il tempo previsto aumenta significativamente, raggiungendo circa a 6132.17 ore, equivalente a più di 255 giorni. Questo evidenzia come il tempo di elaborazione aumenti esponenzialmente con l'aumentare della lunghezza del k-mer, sottolineando l'importanza di considerare l'efficienza computazionale nella scelta della lunghezza del k-mer per l'analisi.

L'analisi dei risultati ottenuti dal grafo di overlap mostra un andamento interessante rispetto al tempo di elaborazione e alle metriche di valutazione quali precisione, recall e F1 score. Il tempo necessario per il calcolo aumenta in modo relativamente lineare con la lunghezza del k-mer, come mostrato nella Figura 13. Questo comportamento suggerisce che il grafo di overlap potrebbe essere computazionalmente più gestibile rispetto al grafo del caos, specialmente per k-mer di lunghezza maggiore.

Anche il per il grado di de Bruijn si osserva un andamento simile al grafo di overlap.

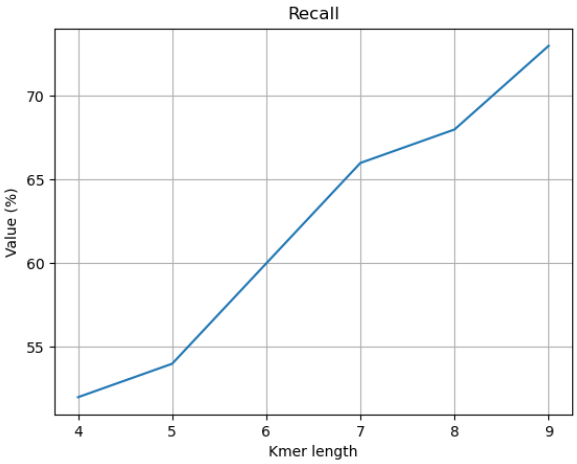

Figura 9: Recall al crescere della lunghezza del k-mer, per il grafo del caos

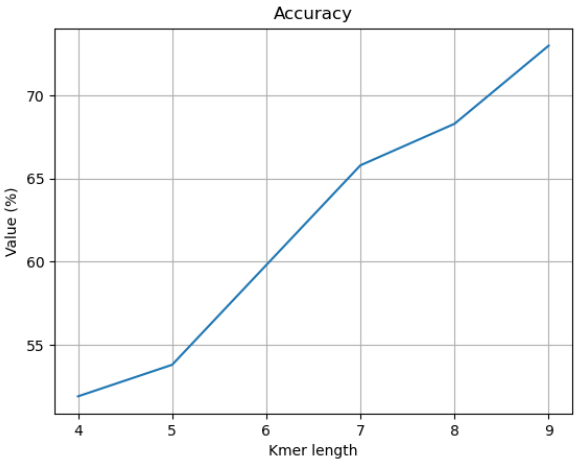

Figura 10: Accuratezza al crescere della lunghezza del k-mer, per il grafo del caos

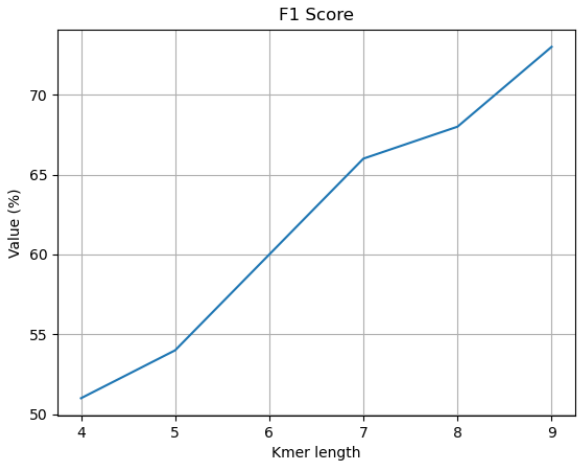

Figura 11: F1 Score al crescere della lunghezza del k-mer, per il grafo del caos

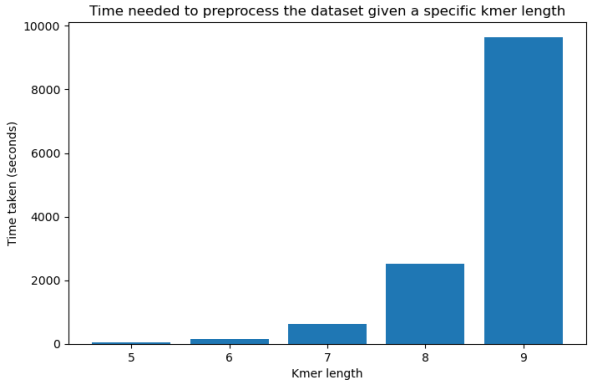

Figura 12: Tempo necessario al crescere della lunghezza del k-mer, per il grafo del Caos

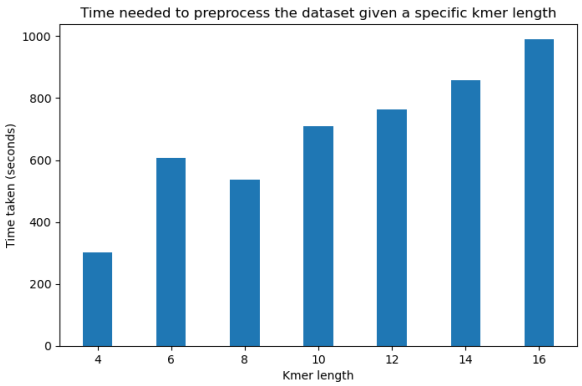

Figura 13: Tempo necessario al crescere della lunghezza del k-mer, per il grafo di Overlap

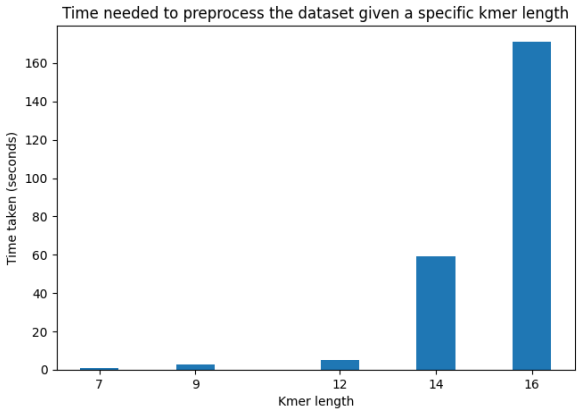

Figura 14: Tempo necessario al crescere della lunghezza del k-mer, per il grafo di de Bruijn

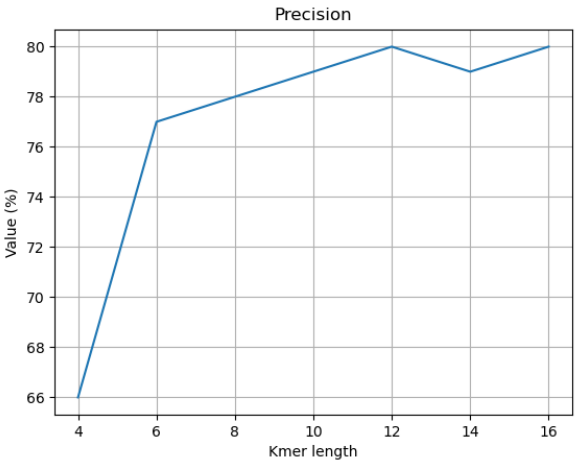

Figura 15: Precision al crescere della lunghezza del k-mer, per il grafo di Overlap

La precisione del grado di overlap , come illustrato nella Figura 15, aumenta notevolmente passando da k-mer di lunghezza 4 a 6, raggiungendo il 77%. Tuttavia, l'incremento della precisione rallenta significativamente per lunghezze di k-mer superiori a 6, con un picco di precisione dell'80% a k-mer di lunghezza 12, per poi osservare una leggera decrescita seguita da un ritorno all'80% con k-mer di lunghezza 16. Questo suggerisce che, dopo un certo punto, aumentare la lunghezza del k-mer non comporta miglioramenti significativi della precisione.

La recall e l'F1 score del grafo di overlap seguono un andamento simile alla precisione, con un picco all'80% per k-mer di lunghezza 12, una leggera decrescita e poi un ritorno all'80% con k-mer di lunghezza 16. Questo indica che la capacità del grafo di overlap di identificare correttamente i k-mer rilevanti raggiunge un punto di saturazione, dopo il quale non si osservano miglioramenti significativi.

Interessante è anche l'andamento delle performance del grafo di de Bruijn (20, 19, **??**, 21), che mostra come, nonostante un iniziale miglioramento, le metriche di precisione, recall, F1 score e accuratezza tendano a stabilizzarsi o addirittura a diminuire leggermente con l'aumento della lunghezza del k-mer oltre un certo punto. Questo fenomeno può riflettere una saturazione nella capacità del modello di trarre vantaggio da informazioni aggiuntive fornite da k-mer più lunghi, suggerendo l'esistenza di un punto ottimale di lunghezza del k-mer che massimizza l'efficacia della classificazione senza incorrere in costi computazionali proibitivi.

I diversi ordini virali possono avere sequenze genomiche molto diverse, che codificano per proteine strutturali e funzionali uniche. Le variazioni nella sequenza genomica possono influenzare la capacità del modello di identificare le caratteristiche distintive di ciascun ordine virale. Ad esempio, alcuni ordini virali potrebbero avere regioni altamen-

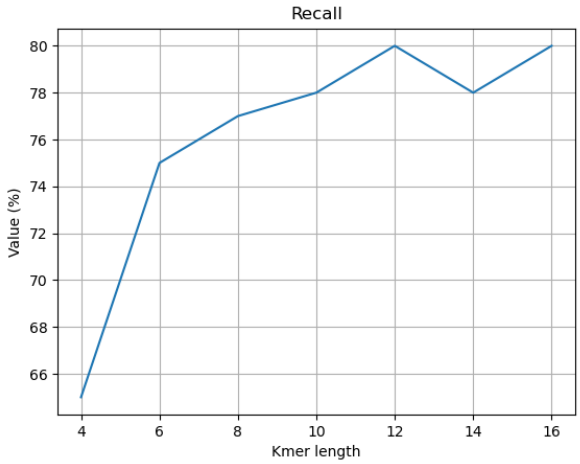

Figura 16: Recall al crescere della lunghezza del k-mer, per il grafo di Overlap

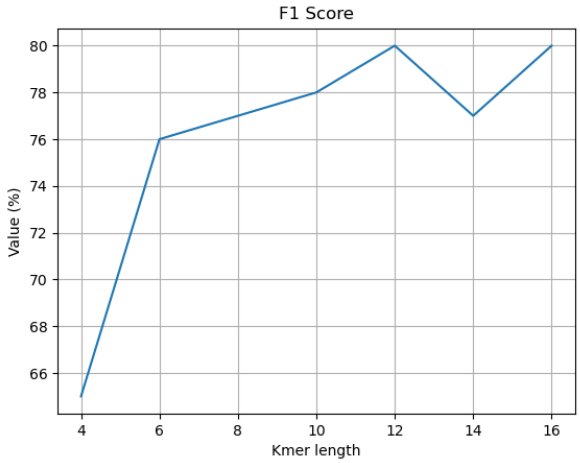

Figura 17: F1 Score al crescere della lunghezza del k-mer, per il grafo di Overlap

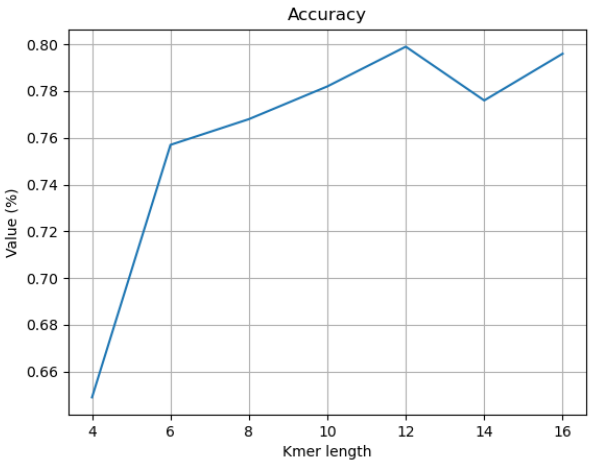

Figura 18: Accuracy al crescere della lunghezza del k-mer, per il grafo di Overlap

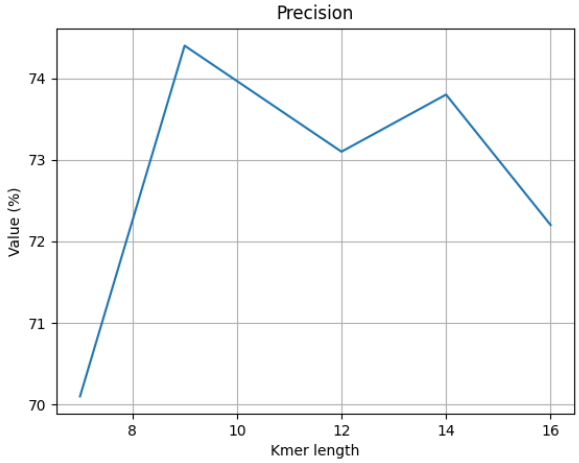

Figura 19: Precisione al crescere della lunghezza del k-mer, per il grafo di de Bruijn

te conservate nel genoma che facilitano la loro identificazione, mentre altri potrebbero avere regioni più variabili che rendono la classificazione più difficile.

Nelle figure 23, 24 e 25 sono riportate le metriche per ciscun ordine virale. Si osserva che in generale il grafo di de Bruijn performa in modo peggiore

In termini di tempo di elaborazione, il grafo di overlap mostra un incremento relativamente lineare, indicando una gestibilità computazionale maggiore rispetto al grafo del caos, specialmente per k-mer di lunghezza maggiore. Questo rende il grafo di overlap una scelta potenzialmente più vantaggiosa in scenari pratici dove il tempo di elaborazione è un fattore critico.

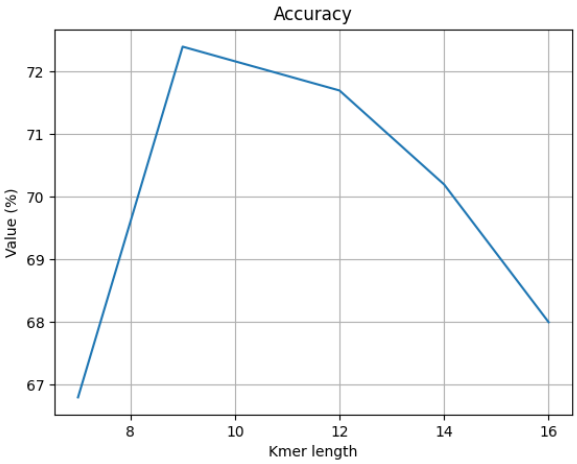

Figura 20: Accuratezza al crescere della lunghezza del k-mer, per il grafo di de Bruijn

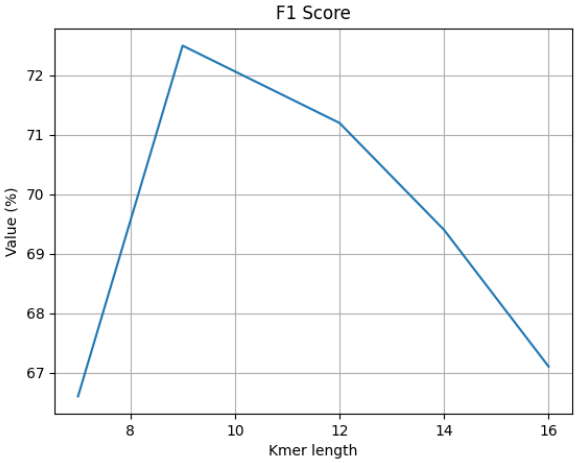

Figura 21: F1 Score al crescere della lunghezza del k-mer, per il grafo di de Bruijn

Complessivamente, questi risultati evidenziano l'importanza di valutare attentamente la lunghezza del k-mer in relazione agli obiettivi specifici dell'analisi e alle risorse disponibili. Mentre il grafo del caos può offrire un'alta precisione, il tempo di elaborazione necessario e l'efficienza computazionale devono essere considerati. Il grafo di overlap, con il suo compromesso tra precisione, recall e tempo di elaborazione, potrebbe essere una scelta più equilibrata per applicazioni pratiche in bioinformatica, specialmente quando si lavora con set di dati di grandi dimensioni o complessi.

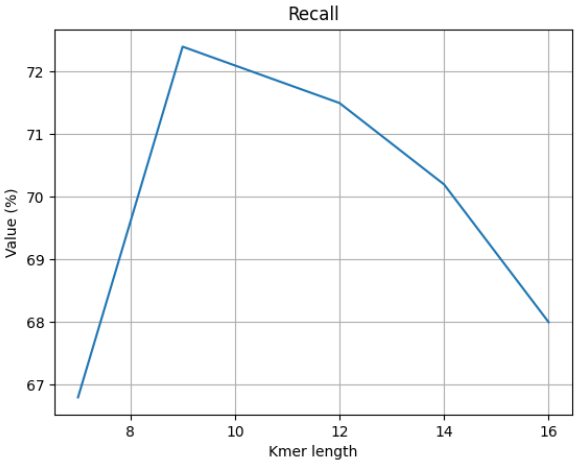

Figura 22: Recall al crescere della lunghezza del k-mer, per il grafo di de Bruijn

5 Conclusioni

Le conclusioni tratte da questo studio evidenziano una dinamica fondamentale nell'ambito della bioinformatica e della classificazione virale attraverso l'utilizzo di grafi, ponendo particolare enfasi sul trade-off tra precisione analitica e efficienza computazionale. L'incremento della lunghezza dei k-mer nel grafo del caos, sebbene conduca a un miglioramento sostanziale della precisione dei risultati, si accompagna a un notevole aumento dei tempi di elaborazione. Questo solleva un quesito essenziale: qual è il giusto compromesso tra l'accuratezza dei risultati ottenuti e la praticità del loro processo di elaborazione? La risoluzione di tale dilemma varia a seconda delle priorità specifiche di ciascun progetto di ricerca e delle risorse computazionali disponibili.

Parallelamente, il grafo di overlap emerge come una soluzione promettente per chi cerca un bilanciamento ottimale tra accuratezza dei risultati e sostenibilità del carico computazionale. Grazie alla sua capacità di gestire un incremento più lineare dei tempi di elaborazione al variare della lunghezza dei k-mer, questo approccio si configura come particolarmente vantaggioso per implementazioni pratiche dove la velocità di analisi rappresenta un fattore critico. Le performance in termini di precisione e F1 Score, pur mantenendo livelli di accuratezza competitivi, evidenziano come il grafo di overlap possa rappresentare un compromesso efficace tra dettaglio analitico e agilità computazionale.

Nell'analisi comparativa, come illustrato nella 26, si osserva che il grafo di overlap mostra valori leggermente superiori in termini di precisione, recall e F1 Score quando la lunghezza del k-mer è impostata a nove. Questo contrappone il grafo di overlap al grafo del caos e al grafo di de Bruijn, i quali, nonostante presentino performance paragonabili tra loro, non raggiungono l'efficacia del primo in termini di bilanciamento tra le diverse metriche considerate.

Questo studio sottolinea l'importanza di una scelta informata riguardo la lunghez-

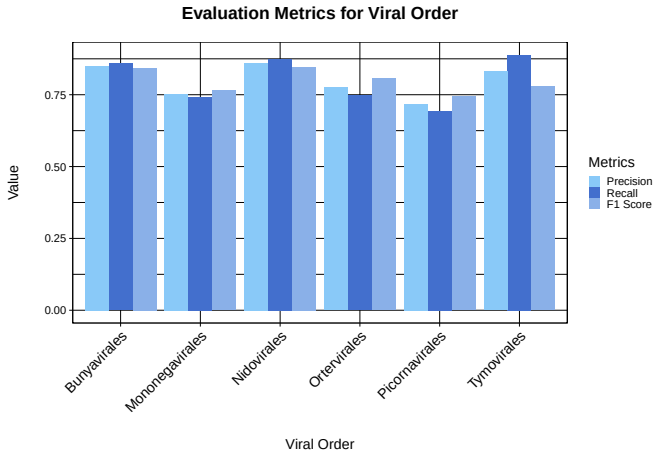

Figura 23: Metriche per ordine virale, per il grafo di Overlap

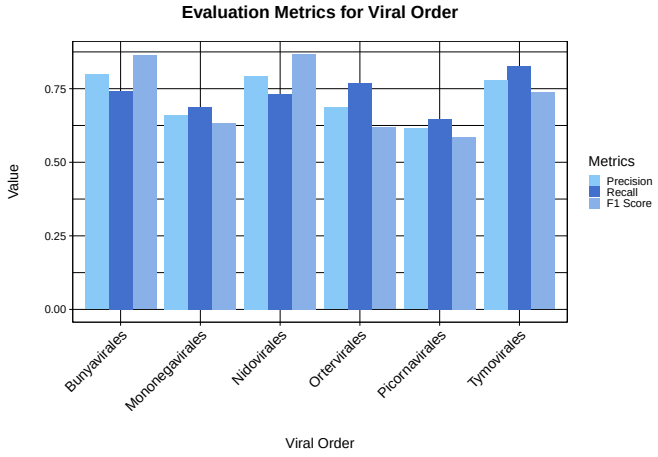

Figura 24: Metriche per ordine virale, per il grafo del Caos

za dei k-mer e il tipo di grafo da utilizzare in funzione degli obiettivi specifici e delle limitazioni computazionali di ogni progetto di ricerca. La comprensione approfondita del trade-off tra precisione e tempo di elaborazione può guidare i ricercatori nella selezione delle strategie analitiche più adatte, ottimizzando così l'efficacia delle indagini bioinformatiche.

Guardando al futuro, si rende necessario un ulteriore sviluppo di metodologie e algoritmi in grado di ridurre il divario tra accuratezza analitica e efficienza computazionale. L'esplorazione di nuove strutture di dati, insieme all'impiego di tecnologie di calcolo avanzate, potrebbe offrire soluzioni innovative per superare le sfide attuali, ampliando le potenzialità della bioinformatica nella ricerca virale e oltre.

In conclusione, questo studio non solo fornisce spunti critici sulle strategie di classifi-

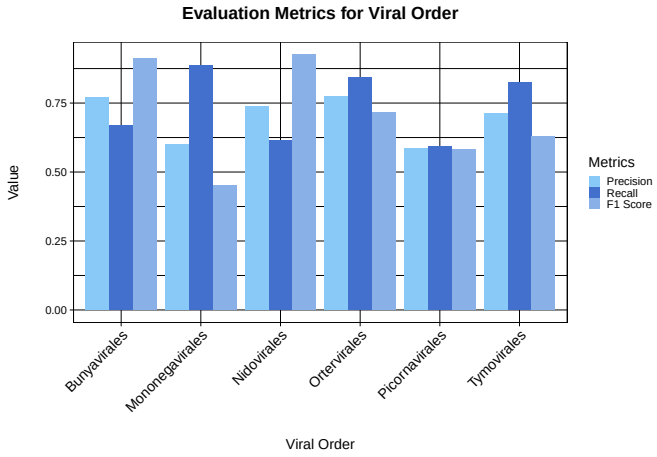

Figura 25: Metriche per ordine virale, per il grafo di de Bruijn

cazione virale basate su grafi, ma apre anche la strada a future ricerche volte a esplorare e sfruttare il pieno potenziale di queste tecniche nel contesto della biologia computazionale.

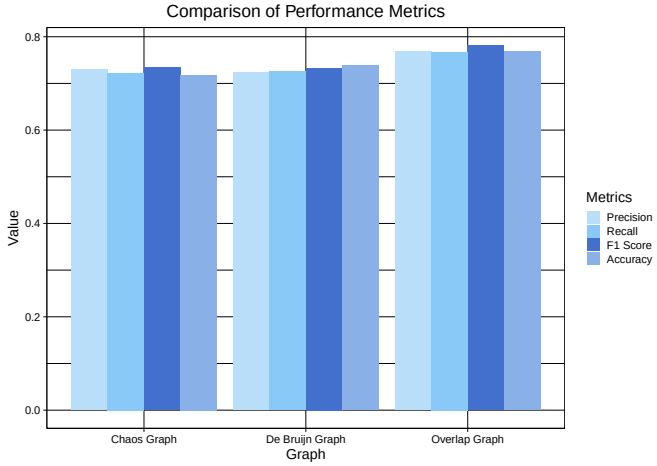

Figura 26: Confronto delle metriche per k = 9

6 Sviluppi Futuri

I risultati raggiunti finora pongono solide basi per un'ampia gamma di sviluppi futuri nel campo della bioinformatica e dello studio dei virus. Uno dei principali ambiti di interesse è l'ottimizzazione della Matrice FCGR. Esiste un'importante opportunità di migliorare il processo di costruzione di questa matrice, rendendola più efficiente in ter-

mini di computazione e più gestibile per set di dati di grandi dimensioni. Ciò potrebbe tradursi in algoritmi più veloci e meno onerosi dal punto di vista della memoria, oltre a metodi raffinati per ridurre il rumore, migliorando così la precisione nell'identificazione dei k-mer.

Un'altra strada promettente è l'approfondimento delle analisi riguardanti i grafi del caos e di overlap. Approfondire l'esplorazione di questi grafi potrebbe rivelare pattern e relazioni complesse tra le sequenze virali che finora sono rimaste inesplorate. Integrare ulteriori dati, come informazioni funzionali o evolutive, potrebbe arricchire significativamente l'analisi, offrendo una comprensione più dettagliata delle dinamiche virali.

7 Data availability

- Codice del progetto
- FCGR: `https://github.com/AlgoLab/complexCGR`

8 Siti web consultati

- `https://it.wikipedia.org/wiki/Tassonomia`
- `https://en.wikipedia.org/wiki/Virus_classification`
- **UGformer:** `https://github.com/diningphil/gnn-comparison`

Riferimenti bibliografici

[1] M.J. Adams, E.J. Lefkowitz, A.M.Q. King, et al. 50 years of the international committee on taxonomy of viruses: progress and prospects. *Archives of Virology*, 162:1441–1446, 2017.

[2] J. Rodney Brister, Danso Ako-adjei, Yiming Bao, and Olga Blinkova. NCBI Viral Genomes Resource. *Nucleic Acids Research*, 43(D1):D571–D577, 11 2014.

[3] Robert A. Edwards and Forest Rohwer. Viral metagenomics. *Nature Reviews Microbiology*, 3(6):504–510, 06 2005.

[4] H. J. Jeffrey. Chaos game representation of gene structure. *Nucleic Acids Research*, 18(8):2163–2170, 1990.

[5] Jens H. Kuhn. Virus taxonomy. In Dennis H. Bamford and Mark Zuckerman, editors, *Encyclopedia of Virology (Fourth Edition)*, pages 28–37. Academic Press, Oxford, fourth edition edition, 2021.

[6] Hannah Löchel, Marius Welzel, Georges Hattab, Anne-Christin Hauschild, and Dominik Heider. Fractal construction of constrained code words for dna storage systems. *Nucleic Acids Research*, 50, 12 2021.

[7] Hannah Franziska Löchel and Dominik Heider. Chaos game representation and its applications in bioinformatics. *Computational and Structural Biotechnology Journal*, 19:6263–6271, 2021.

[8] Páll Melsted and Jonathan K Pritchard. Efficient counting of k-mers in dna sequences using a bloom filter. *BMC Bioinformatics*, 12:333, 2011.

[9] Aluísio Pinheiro, Hildete Prisco Pinheiro, and Pranab Kumar Sen. The use of hamming distance in bioinformatics. volume 28 of *Handbook of Statistics*, pages 129–162. Elsevier, 2012.

[10] Eric W Sayers, Jeffrey Beck, Evan E Bolton, Devon Bourexis, James R Brister, Kathi Canese, Donald C Comeau, Kathryn Funk, Sunghwan Kim, William Klimke, Aron Marchler-Bauer, Melissa Landrum, Stacy Lathrop, Zhiyong Lu, Thomas L Madden, Nuala O'Leary, Lon Phan, Sanjida H Rangwala, Valerie A Schneider, Yuri Skripchenko, Jiyao Wang, Jian Ye, Barton W Trawick, Kim D Pruitt, and Stephen T Sherry. Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, 49(D1):D10–D17, January 2021.

[11] Rex Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L. Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling, 2019.