UNIVERSITÀ DEGLI STUDI DI SALERNO

DIPARTIMENTO DI INFORMATICA

Esame Strumenti formali per la Bioinformatica

LA PANGENOMICA

Candidati:

Martina Giugliano

Mat. 0522501909

Jacopo de Dominicis

Mat. 0522501892

ANNO ACCADEMICO 2024/2025

Contents

Chapter 1

Introduzione

La pangenomica rappresenta un paradigma emergente e rivoluzionario nell'ambito della genomica, volto a superare i limiti imposti dal tradizionale approccio basato sullo studio di un singolo genoma di riferimento. Questo concetto è nato per analizzare e comparare l'insieme di tutti i geni presenti in una popolazione o in un gruppo di organismi appartenenti alla stessa specie, creando un panorama genomico complessivo ed inclusivo. L'idea alla base è quella di superare le carenze del genoma di riferimento, il quale non riesce a cogliere pienamente la complessità e la variabilità presente all'interno di una specie. La pangenomica suddivide il genoma in due componenti principali: il genoma core, che comprende i geni condivisi da tutti gli individui della specie ed i geni variabili, a loro volta suddivisi in geni shell e geni cloud. Questa suddivisione consente di comprendere meglio la plasticità genetica e di esplorare le relazioni tra genotipo e fenotipo con una precisione maggiore rispetto agli approcci tradizionali. L'importanza della pangenomica è particolarmente evidente in ambiti come l'agricoltura, dove lo studio del pangenoma di colture come il riso o il mais permette di identificare geni utili per il miglioramento delle varietà coltivate, aumentando la resa, la resistenza a malattie e la tolleranza a stress ambientali. Inoltre, nella medicina, la pangenomica offre nuove prospettive per comprendere la variabilità genetica umana, individuare i fattori genetici alla base delle malattie e sviluppare terapie personalizzate. Un aspetto centrale della

pangenomica è rappresentato dall'uso di tecnologie avanzate di sequenziamento del DNA, che consentono di generare enormi quantità di dati genomici. Questi dati vengono poi analizzati mediante strumenti bioinformatici in grado di identificare somiglianze e differenze tra i genomi e di costruire pangenomi che possano essere visualizzati e interrogati. Tuttavia, il grande volume di dati prodotto pone anche sfide significative, tra cui la necessità di infrastrutture computazionali potenti e di metodologie di analisi che possano gestire efficacemente la complessità e la scala di tali dataset. La pangenomica computazionale si sta rapidamente affermando come un campo di ricerca di fondamentale importanza, in grado di rivoluzionare la nostra comprensione della diversità genetica e di trasformare l'approccio degli informatici alle sfide legate all'analisi delle sequenze biologiche. Negli ultimi decenni, il progresso in settori come la combinatoria, la teoria dei grafi e lo sviluppo di strutture dati ha svolto un ruolo cruciale nella creazione di strumenti software dedicati allo studio del genoma umano. Oggi, ad emergere è l'uso di rappresentazioni grafiche di più genomi che hanno consentito a biologi computazionali di affrontare progetti ambiziosi su una scala di popolazione.

1.1 Overview

Il presente elaborato affornta il tema della pangenomica, un paradigma emergente nella genomica che supera i limiti del modello basato su un unico genoma di riferimento. Il lavoro si articola in diversi capitoli che analizzano progressivamente il passaggio dal concetto di genoma a quello di pangenoma, fino ad arrivare agli strumenti e alle metodologiè computazionali più avanzate per la sua analisi.

Nel Capitolo 1, si introduce il concetto di pangenomica e la sua importanza in vari ambiti scientifici, come la medicina di precisione e l'agricoltura. Viene evidenziato come il pangenoma permetta di cogliere la complessità genetica di una specie meglio di un singolo genoma di riferimento.

Il Capitolo 2 illustra il passaggio dal genoma di riferimento al pangenoma, descrivendo i limiti delle rappresentazioni genomiche lineari e le problematiche legate alla perdita di varianti strutturali significative. Viene spiegata l'evoluzione della genomica computazionale grazie all'uso di modelli più inclusivi, capaci di rappresentare con maggiore fedeltà la diversità genetica intra-specie.

Nel Capitolo 3, si approfondiscono i diversi modelli computazionali di pangenomi, con particolare attenzione ai grafi pangenomici. I grafi rappresentano un approccio avanzato per modellare le relazioni tra sequenze genomiche, offrendo vantaggi come la riduzione della ridondanza e una maggiore precisione nella rappresentazione delle varianti.

Il Capitolo 4 si concentra sugli studi precedenti nel campo della pangenomica e introduce due approcci di riferimento: EUPAN e HUPAN. Vengono messi a confronto i due strumenti, evidenziando come HUPAN sia stato sviluppato per superare le limitazioni di EUPAN, rendendo più efficiente l'analisi di pangenomi umani. Inoltre, viene analizzato il ruolo dei grafi di De Bruijn per la scoperta delle varianti genetiche.

Il Capitolo 5 presenta Pangraph, uno strumento per la visualizzazione e l'analisi interattiva dei grafi pangenomici. Viene descritto il workflow di utilizzo, dall'importazione dei dati in formato GFA fino alla visualizzazione e all'analisi delle varianti genetiche.

Infine, nel Capitolo 6, vengono tratte le conclusioni finali, con una riflessione sull'impatto della pangenomica nella biologia computazionale e sulle prospettive future del settore.

Chapter 2

Dal genoma al pangenoma: un cambiamento di prospettiva

Il termine *pangenoma* è stato definito con il suo significato attuale da Tettelin et al. nel 2005; deriva *pan* dalla parola greca , che significa "intero" o "tutto", mentre il *genoma* è un termine comunemente usato per descrivere il materiale genetico completo di un organismo [1].

Il passaggio dal genoma di riferimento al pangenoma computazionale rappresenta un'evoluzione fondamentale nella genomica, motivata dalle limitazioni intrinseche dei riferimenti lineari e dai progressi tecnologici. Il genoma di riferimento, tradizionalmente utilizzato come modello lineare per rappresentare il DNA umano, non riesce a catturare adeguatamente la vasta diversità genetica presente nella popolazione.

La scoperta di varianti strutturali (SV), che coinvolgono mutazioni di almeno 50 paia di basi, ha evidenziato questa lacuna: molte letture campionate da individui con SV specifiche non riescono ad allinearsi al genoma di riferimento e vengono erroneamente scartate come artefatti [2].

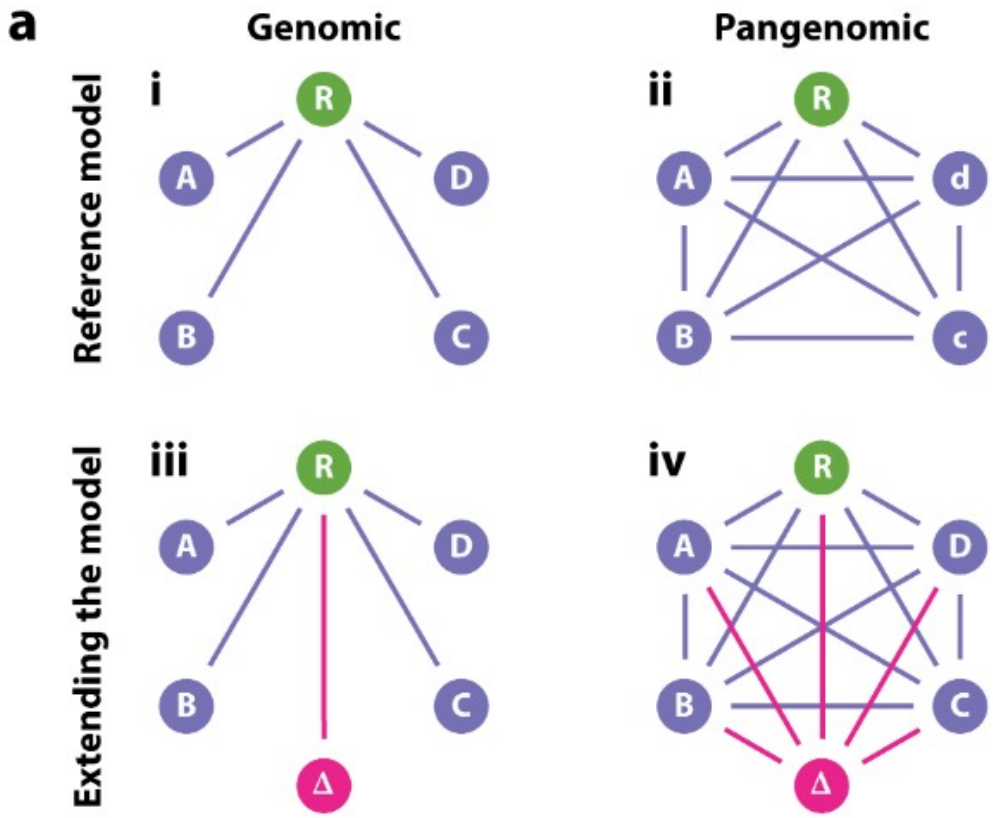

Figure 2.1: Modelli genomici e pangenomici a confronto

2. DAL GENOMA AL PANGENOMA: UN CAMBIAMENTO DI PROSPETTIVA

L'avvento delle tecnologie di sequenziamento a lettura lunga ha permesso di rilevare un numero significativamente maggiore di varianti, dimostrando l'inadeguatezza dei riferimenti lineari nell'integrare una rappresentazione completa della diversità genetica. Altri limiti includono la difficoltà nell'aggiornamento del riferimento esistente e il bias, un errore sistematico o una distorsione che porta a risultati non rappresentativi o imprecisi rispetto alla realtà, nel nostro caso, nell'interpretazione delle sequenze [3]. Il pangenoma, al contrario, offre una soluzione più inclusiva e flessibile. Con una rappresentazione che abbraccia molteplici varianti genetiche, consente di ridurre il bias del riferimento, migliorare l'accuratezza della mappatura durante il sequenziamento di nuovi individui e aumentare la precisione nell'identificazione di varianti rare. Inoltre, migliora l'assemblaggio de novo, facilitando applicazioni cliniche più accurate ed efficaci. A tal punto, la **Figura 2.1** mostra che nelle analisi genomiche basate su riferimento, tutti i genomi (A-D) vengono confrontati tra loro in base alla relazione con il genoma di riferimento (R). (ii) In un contesto pangenomico, si tenta di modellare le relazioni dirette tra tutti i genomi nell'analisi, da cui un particolare riferimento viene scelto arbitrariamente. (iii) Quando si estende l'analisi con un nuovo genoma, , lo si aggiunge al modello genomico confrontandolo con il genoma di riferimento. (iv) Al contrario, l'aggiunta di un nuovo genoma a un'analisi pangenomica lo confronta direttamente con tutti gli altri genomi nel modello. Questo cambiamento segna un progresso cruciale per garantire che la genomica rifletta la reale complessità del genoma umano. Nel 2005 è stato condotto il primo studio di pangenomica, pubblicato sulla rivista PNAS, incentrato sull'analisi genetica di diversi ceppi di Streptococcus agalactiae. Questo batterio è noto per causare frequenti infezioni nei neonati e negli anziani. Come mostrato in Figura 2, lo studio ha evidenziato differenze rilevanti tra i genomi dei ceppi analizzati, dimostrando l'importanza di studiare i genomi di più varianti di una stessa specie per comprenderne meglio l'evoluzione e la patogenicità.

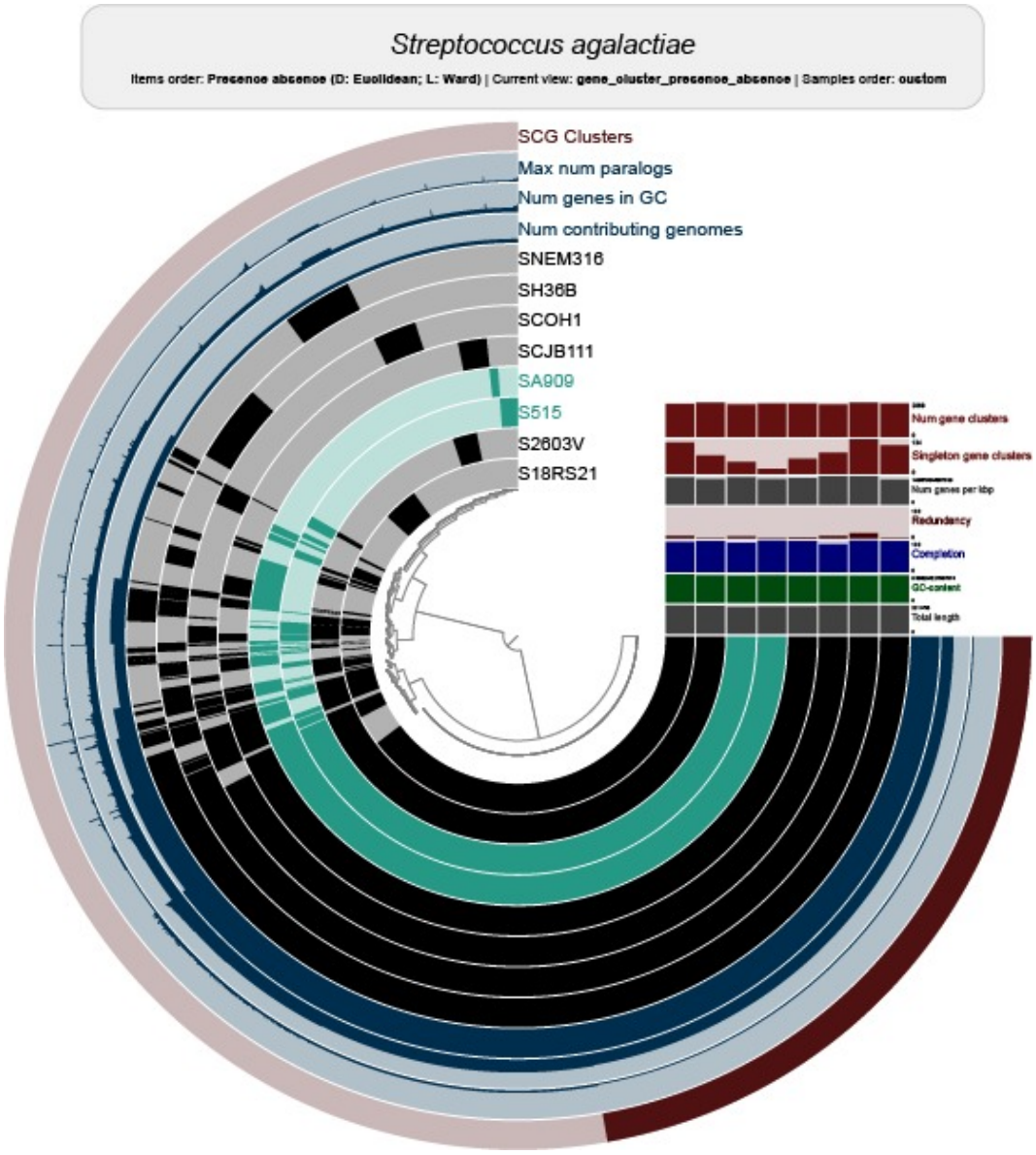

Figure 2.2: Analisi genetica di Streptococcus agalactiae

La pangenomica suddivide i geni di un organismo in due categorie principali, come mostrato in Figura 2.3.

- **Geni Core**: questa categoria comprende i geni presenti in tutti gli individui di una specie. Spesso identificati come geni housekeeping, essi codificano funzioni essenziali per l'organismo, come la formazione della membrana cellulare, le proteine di legame, e le attività regolatorie. Questi geni rappresentano, per così dire, il nucleo genetico della specie, ovvero le funzioni fondamentali che vengono trasmesse di generazione in generazione. Sono elementi genetici difficilmente sostituibili o eliminabili senza provocare gravi conseguenze.

- **Geni variabili**: in questa categoria rientrano i geni che si trovano solo in alcuni ceppi, detti *geni shell* o, in casi estremi, in un unico ceppo, detti *geni cloud*, noti anche come genoma accessorio. Tra questi si trovano, ad esempio, i geni che conferiscono resistenza agli antibiotici o che permettono l'adattamento a specifici ambienti in cui il ceppo si trova a vivere.

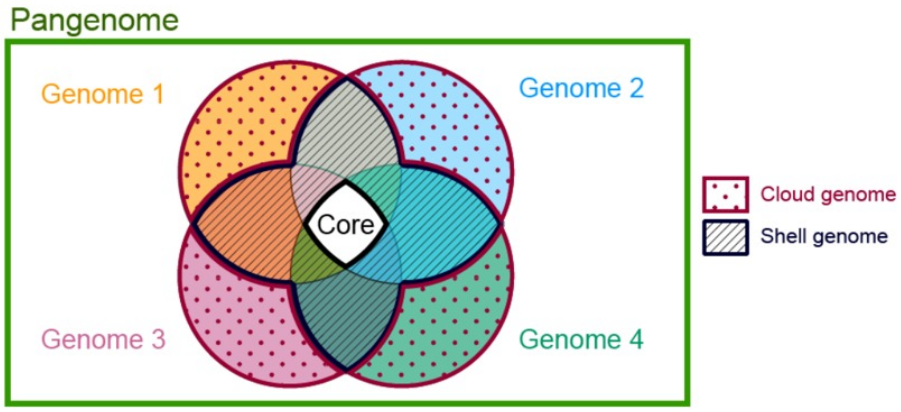

Figure 2.3: Classificazione dei geni di un pangenoma

La **Figura 2.3** in alto mostra la classificazione dei geni del pangenoma in base alla loro presenza in tutti i ceppi di una specie (geni core), in alcuni (geni shell) o in uno solo (geni cloud).

Chapter 3

Modello pangenomico

Il pangenoma può essere rappresentato attraverso diversi modelli computazionali, i quali mirano a catturare la complessità e la diversità del materiale genetico in maniera efficace. Essi possono assumere molte forme, fra i modelli più utilizzati si annoverano: le sequenze multiple concatenate, i grafi di de Bruijn, e i grafi pangenomici più complessi, che integrano informazioni su varianti genetiche, frequenze alleliche e relazioni strutturali, offrendo una struttura dati compatta e versatile. Un esempio significativo è rappresentato dalle varianti riportate dal consorzio The 1000 Genomes Project Consortium (2015). Tali varianti strutturali possono modificare un genoma in un genoma simile ma funzionalmente diverso e sono il risultato di rielaborazioni di segmenti di sequenza nel genoma, come duplicazione, inversioni e traslocazione di segmenti.

- **Sequenze Multiple Concatenate**
  In questo modello, i genomi di vari individui vengono concatenati per creare una sequenza unica più lunga, nella quale le varianti sono annotate rispetto a un riferimento.
  - Vantaggi: È semplice da implementare e non richiede infrastrutture computazionali sofisticate.
  - Svantaggi: Non scala bene con l'aumento del numero di genomi; inoltre, l'annotazione delle varianti rispetto al riferimento è soggetta

a bias, risultando in una rappresentazione riduttiva.

- **Grafi di de Bruijn**

  Questi grafi sono comunemente usati nel sequenziamento genomico per rappresentare il DNA attraverso k-mers (sottostringhe di lunghezza k). I nodi rappresentano i k-mers, mentre gli archi rappresentano la loro adiacenza.

  - Uso nei pangenomi: Possono essere estesi per includere varianti genetiche, rendendoli utili per assemblare genomi di riferimento alternativi.
  - Sfide computazionali: La gestione della complessità aumenta rapidamente con la crescita della dimensione del dataset e della diversità genetica.

- **Grafi del Pangenoma (Graph-Based Models)** Questo è il modello più avanzato e flessibile. Un grafo pangenomico è una struttura dati che rappresenta i genomi come un grafo diretto e aciclico (Directed Acyclic Graph, DAG):

  - I nodi rappresentano segmenti di DNA (sequenze).
  - Gli archi indicano le possibili connessioni tra segmenti, derivanti da varianti genetiche o diverse organizzazioni del genoma.

3.1 Grafi

I grafi rappresentano una delle strutture più adatte per descrivere la complessità delle relazioni tra più genomi. Grazie alla loro capacità di modellare varianti strutturali attraverso elementi come bordi orientati, cicli e sottostrutture particolari (ad esempio, le "bolle"), consentono una rappresentazione dettagliata e versatile delle differenze genomiche.

Le bolle, in particolare, sono sottografi aciclici diretti definiti da una

coppia di vertici: un nodo sorgente S e un nodo terminale T. Tutti i percorsi da S a T sono privi di sovrapposizioni nei nodi intermedi, una caratteristica che permette di identificare e gestire varianti come inserzioni o delezioni tramite algoritmi ottimizzati e strutture dati avanzate.

Un aspetto cruciale della pangenomica è costruire grafi che riassumano efficacemente una collezione di sequenze genomiche. Prima di approfondire le tecniche per generare tali grafi, è utile introdurre alcune definizioni chiave. Un grafo di variazione è una struttura orientata in cui i vertici sono etichettati con stringhe non vuote. Formalmente, un grafo di variazione può essere descritto come: G=(V,A,W); G = (V, A, W); G=(V,A,W):

- VVV: insieme dei nodi etichettati;
- AAA: insieme degli archi diretti che connettono i nodi;
- WWW: insieme di percorsi distinti che rappresentano sequenze genomiche di interesse, inseriti nella rappresentazione.

Tuttavia, in alcune applicazioni, l'insieme delle varianti non è noto a priori. In questi casi, il grafo deve rappresentare non solo le sequenze di input ma anche quelle compatibili con esse, introducendo così il concetto di grafo di sequenza.

Un grafo di sequenza G = (V, A) dove V indica l'etichetta funzione e A l'insieme degli archi, rappresenta un'estensione del grafo di variazione, in cui i percorsi possibili includono tutte le combinazioni indotte dagli archi. Di conseguenza, i grafi di sequenza catturano anche varianti non esplicitamente presenti nei genomi di input. Questa caratteristica consente di rappresentare in modo compatto le sequenze ridondanti in una struttura più snella, senza perdere la complessità del dataset originale.

3.2 Vantaggi nell'uso di grafi

I grafi pangenomi rappresentano uno strumento avanzato per la ricerca genomica e la biologia computazionale, offrendo numerosi vantaggi rispetto ai modelli lineari tradizionali. Di seguito ne vengono elencati i principali.

- **Rappresentazione della Variabilità**

  I grafi consentono di modellare varianti strutturali complesse, come duplicazioni, inversioni e delezioni, utilizzando nodi e archi. Questa rappresentazione permette di gestire le varianti in modo efficiente, includendone di rare e complesse, come traslocazioni o fusioni geniche.

- **Riduzione della Ridondanza** Attraverso l'uso di grafi di sequenza, molteplici sequenze genomiche possono essere compresse in una struttura dati compatta ma rappresentativa dell'intero dataset. Tale struttura è espandibile, consentendo l'aggiunta di nuove sequenze genomiche senza dover ristrutturare completamente il modello.

  - Eliminazione della duplicazione delle informazioni: : Le regioni genomiche condivise tra più individui o specie vengono rappresentate una sola volta, riducendo lo spazio richiesto per la memorizzazione.
  - Efficienza computazionale: Grazie alla compressione dei dati, i grafi pangenomici riducono il costo computazionale associato all'analisi di dataset genomici di grandi dimensioni.

- **Mappatura delle Sequenze**

  Un classico esempio di applicazione è il miglioramento della mappatura delle letture:

  - In un modello lineare, l'allineamento di sequenze multiple spesso genera disallineamenti e indel, causando una perdita di informazioni utili.

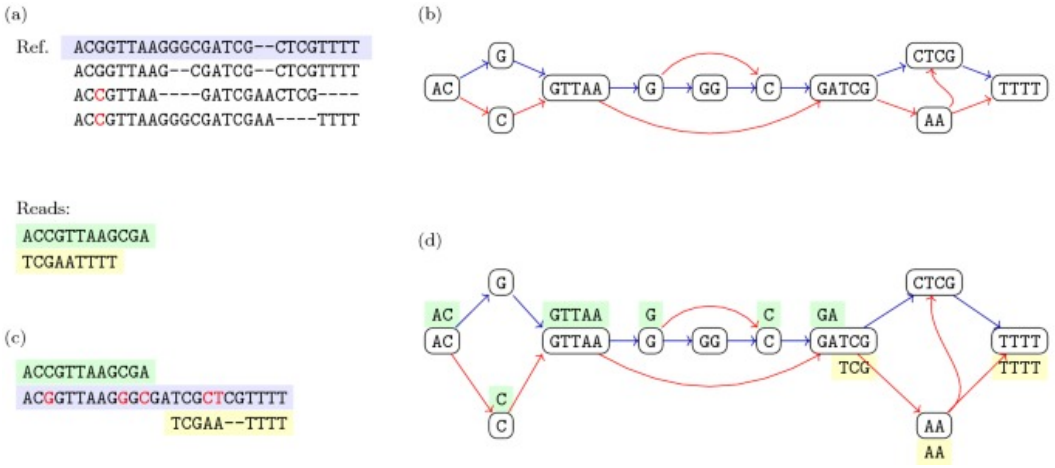

Figure 3.1: Letture di mappatura su un genoma di riferimento

– In un grafo di variazione, invece, le letture possono essere mappate senza mismatch, sfruttando la rappresentazione delle varianti strutturali.

  In **Figura 3.1** è mostrato un esempio pratico di come un grafico pangenomico migliori la qualità delle letture di mappatura su un genoma di riferimento.

– Un allineamento di sequenze multiple di un genoma di riferimento lineare e altri tre genomi che contengono variazioni rispetto al riferimento.

– Un grafico di variazione costruito dalla matrice dell'allineamento multiplo dei genomi; in rosso i bordi che rappresentano le variazioni nel grafico e formano le tipiche "bolle" nel grafico. Si noti che il grafico può contenere un percorso che non rappresenta alcun genoma di input (ad esempio, ACCGTTAAGGGCGATCGAACTCGTTTT).

– Mappatura di due letture (ACCGTTAAGCGA e ACCGTTAAGCGA) sul genoma di riferimento lineare. Si noti che gli allineamenti inducono disallineamenti e indel.

– Mappatura delle stesse letture sul grafico di variazione. Si noti che,

in questo caso, la mappatura è possibile senza alcun disallineamento [4].

- **Applicazioni Cliniche Evolutive**

  - In ambito clinico, i grafi consentono una rappresentazione più accurata delle mutazioni e delle varianti rare, migliorando la diagnosi e la personalizzazione delle terapie geniche, consentendo lo sviluppo di terapie mirate e personalizzate.
  - Facilitano lo studio dei processi evolutivi, come la ricombinazione e l'introgressione genetica tra popolazioni o specie, analizzandone le differenze su larga scala, ricostruendo anche la storia evolutiva di diverse popolazioni.

Chapter 4

Related work

Se al 2005 risale il primo studio di pangenomica che vede come protagonista il batterio *Streptococcus agalactiae*, inaugurando una nuova era nella comprensione della variabilità genetica all'interno delle specie microbiotiche, è nel 2010 che la pangenomica si estende al genoma umano.

Durante tale studio, vennero analizzati solo due genomi rappresentativi di Asia e Africa, e rilevate circa 5Mb di nuove sequenze assenti nel genoma di riferimento per ogni individuo. [5] In uno studio successivo [6], la rianalisi delle nuove sequenze da 5 Mb di un individuo cinese ha mostrato che sequenze da 3,7 Mb potevano essere allineate al genoma di riferimento umano GRCh38. In un articolo più recente, Sherman et al. hanno riportato un pan-genoma africano [7], conteneva circa 300 Mb di sequenze uniche mancanti nel genoma di riferimento umano, in particolare, la maggior parte di queste nuove sequenze erano specifiche per individuo e solo 81 Mb di sequenze erano mostrate in due o più individui.La crescita esplosiva dei dati di sequenziamento dell'intero genoma umano porta sfide significative e grandi opportunità per studiare il pan-genoma di una popolazione specifica [8] [9]. Tuttavia, costruire le sequenze del pan-genoma da centinaia di genomi individuali è una sfida enorme.

A tal punto è stato segnalato uno strumento EUPAN basato su una strategia "map-to-pan" e applicato a più di 3000 genomi di riso. Tuttavia,

a causa delle grandi dimensioni del genoma umano, EUPAN non può essere applicato per l'analisi del pan-genoma umano a causa dell'enorme requisito di dimensione della memoria della fase di assemblaggio de novo (sono necessari più di 500 Gb di memoria per assemblare un genoma umano da dati di sequenziamento 30 volte).
Nella prossima sezione vedremo le basi per la costruzione del pangenoma HUPAN.

4.1 EUPAN e HUPAN a confronto

A differenza del pangenoma EUPAN, in questa sezione viene presentato uno strumento di analisi del pangenoma umano, HUPAN, illustrato nel documento [9]. Anch'esso utilizza la strategia "map to-pan" per determinare i PAV genici per ogni individuo.

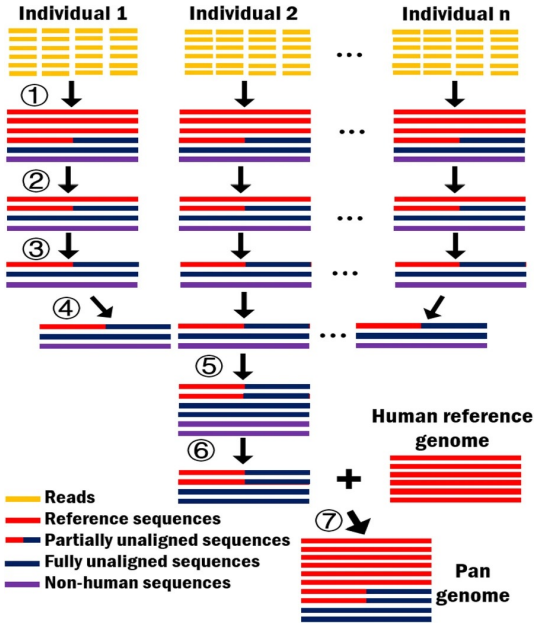

Figure 4.1: Diagramma di costruzione del pan-genoma in HUPAN.

Figura 4.1

1. assemblaggio de novo di tutte le letture in contig;
2. rimozione di contig simili al genoma di riferimento umano;
3. estrazione di sequenze non allineate (incluse sequenze completamente non allineate e sequenze parzialmente non allineate);
4. fusione di sequenze non allineate da più individui;
5. rimozione di sequenze ridondanti;
6. rimozione di potenziali contaminazioni;
7. costruzione del pan-genoma che combina il genoma di riferimento umano e nuove sequenze;

	UPA	EUPANO
# contig grezzi (> 500 bp)	610.537	610.537
lunghezza dei contig grezzi (bp)	2.709.735.693	2.709.735.693
# contigs dopo il filtraggio	24.150	–
lunghezza dei contig dopo il filtraggio (bp)	76.168.613	–
# assemblaggi non corretti	1037	1050
Lunghezza dei contig non assemblati (bp)	5.483.408	5.657.999
# Contig completamente non allineati	5371	5394
Lunghezza dei contig completamente non allineati (bp)	5.000.779	5.014.971
# Contig parzialmente non allineati	1187	1197
Lunghezza dei contig parzialmente non allineati (bp)	5.435.999	5.628.509
Tempo CPU (ore)	42	275
Memoria massima (Gb)	92	250

Figure 4.2: Confronto tra HUPAN ed EUPAN

L'approccio di mappatura delle sequenze genetiche presenta alcune limitazioni. Prima di tutto, se il campione contiene sequenze diverse o assenti rispetto al riferimento usato, le letture potrebbero non essere mappate correttamente. Inoltre, le sequenze di riferimento, soprattutto per gli

eucarioti, sono incomplete in alcune aree, come quelle vicino ai telomeri e ai centromeri, e le letture da queste regioni potrebbero causare errori. Inoltre, potrebbe non esserci una sequenza di riferimento disponibile, come nel caso del sequenziamento ecologico. Anche i metodi di chiamata delle varianti si concentrano generalmente su un solo tipo di variante, ma in presenza di varianti diverse nello stesso campione, ciò potrebbe portare a errori. Sebbene esistano metodi per rilevare grandi variazioni strutturali, non possono determinare con precisione posizione, dimensione o sequenza. Infine, la mappatura tende a ignorare le informazioni sulle variazioni genetiche all'interno della specie.

Un'alternativa a questi problemi è l'assemblaggio de novo, che non dipende da un riferimento e può rilevare diversi tipi di varianti. Tuttavia, anche questo approccio ha delle limitazioni, poiché tende a trattare la sequenza come se fosse uniforme, ignorando le variazioni. Esistono algoritmi che cercano di affrontare questo problema, ma non sono ancora soluzioni universali, specialmente per campioni con alta variabilità genetica. Per affrontare queste limitazioni, vengono proposti nuovi algoritmi di assemblaggio de novo che si concentrano sulla rilevazione e caratterizzazione della variabilità genetica in uno o più individui. Questi algoritmi migliorano i classici grafici di de Bruijn, colorando i nodi e gli spigoli del grafico in base ai campioni in cui appaiono. Questo approccio integra informazioni provenienti da più campioni, comprese sequenze di riferimento e varianti note, e consente di rilevare variazioni anche senza un riferimento, migliorando l'accuratezza e genotipizzando varianti già conosciute. Un esempio di questo approccio è Cortex, che ha contribuito a progetti come il 1000 Genomes Project.

I grafici di De Bruijn sono strumenti usati per rappresentare le sovrapposizioni tra sequenze di DNA e sono alla base di molti algoritmi di assemblaggio del genoma, come AllPaths-LG, SOAPdenovo, Abyss e Velvet.

Il grafico di De Bruijn colorato è un'evoluzione del classico grafico di De Bruijn, pensata per gestire dati provenienti da più campioni contemporaneamente. In pratica, ogni nodo del grafico viene "colorato" per indicare da quale campione proviene la sequenza, permettendo così di visualizzare le differenze e le somiglianze tra diversi genomi. Questi campioni possono includere sequenze di riferimento, varianti già note o anche dati raccolti da esperimenti diversi.

Uno degli utilizzi principali di questo grafico è scoprire varianti genetiche e fare genotipizzazione, ovvero identificare le differenze tra il DNA di diversi individui. Un modo per farlo è attraverso l'analisi delle bolle nel grafico: quando esiste una variazione tra due sequenze, si crea una struttura simile a una biforcazione nel grafico, che può essere riconosciuta e analizzata.

Tuttavia, individuare queste bolle non è sempre facile. A volte si possono generare falsi positivi, cioè varianti che in realtà non esistono davvero. Per ridurre questi errori, si può usare un genoma di riferimento aploide, che aiuta a distinguere le reali differenze da quelle dovute a ripetizioni o errori. Questo è particolarmente utile per identificare varianti omozigoti, ovvero quelle presenti in entrambe le copie del genoma di un individuo.

L'efficacia di questo metodo dipende da diversi fattori, come la complessità del grafico, la copertura della sequenza, la lunghezza del k-mer (cioè la dimensione dei frammenti di DNA usati per costruire il grafico) e il tasso di errore nella lettura della sequenza.

Per migliorare l'analisi, è stato sviluppato un algoritmo specifico chiamato bubble-calling (BC), progettato per identificare queste bolle in modo più accurato e aiutare così a individuare le varianti genetiche in un individuo.

- Scoperta di varianti in un singolo individuo diploide consanguineo (blu) con una sequenza di riferimento (rosso). I veri polimorfismi generano bolle che divergono dal riferimento, mentre le strutture ripetute portano a bolle osservate anche nel riferimento.

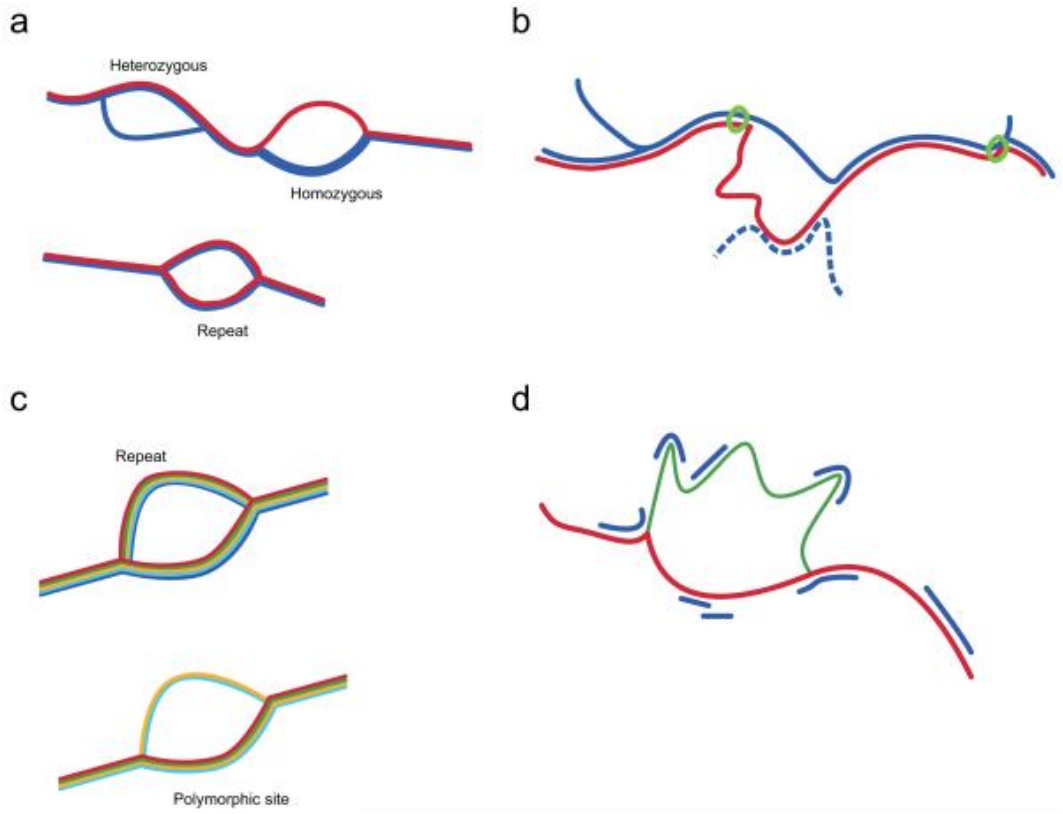

Figure 4.3: Rappresentazione schematica di quattro metodi di analisi delle variazioni mediante grafici de Bruijn colorati

- Anche quando l'allele di riferimento (rosso) non forma una bolla pulita, possiamo identificare i siti delle varianti omozigoti tracciando la divergenza del percorso di riferimento da quello del campione. Quando troviamo un punto di interruzione, prendiamo il contig più lungo nel campione (vale a dire il percorso fino alla giunzione successiva) e chiediamo se il percorso di riferimento ritorna prima di questo punto (cerchio verde = sequenza di ancoraggio). L'algoritmo (divergenza del percorso) non è influenzato dalla sequenza ripetuta all'interno dell'allele di riferimento presente altrove nel genoma del campione (punteggiato in blu).

- Quando vengono combinati molti campioni (ciascuno di un colore diverso), è possibile distinguere le bolle indotte dalla ripetizione (in cui entrambi i lati della bolla sono presenti in tutti i campioni) dai veri siti

varianti.

- La probabilità di un dato genotipo può essere calcolata dalla copertura (blu) di ciascun allele (verde, rosso), tenendo conto dei contributi da altre parti del genoma. In questo esempio, il campione è eterozigote, quindi ha una copertura di entrambi gli alleli, sebbene non sufficiente per consentire l'assemblaggio completo.

Numerosi metodi utilizzano la costruzione di De Bruijn compattata per elaborare grafi del pangenoma.

4.2 Costruzione di un pangenoma

Il passo principale nella pangenomica computazionale è costruire un "grafico di variazione", che rappresenta le differenze tra i genomi. Questo può essere fatto in due modi: partendo da un insieme di sequenze (che corrispondono ai percorsi nel grafico) oppure da un allineamento multiplo di sequenze. Il secondo approccio è più semplice, ma dipende molto dalla qualità dell'allineamento. L'obiettivo è creare un grafico che rappresenti uno o più genomi. Questo processo avviene in due fasi: prima si costruisce un grafico di sequenza, poi si estraggono i percorsi che rappresentano i genomi. Va notato che può esistere più di un grafico che rappresenta lo stesso set di genomi e alcuni di questi grafici potrebbero non sembrare un allineamento tradizionale, come nel caso in cui contengano un ciclo. Un allineamento consiste nel modificare una sequenza inserendo spazi vuoti (caratteri "-") per rendere le sequenze comparabili. Due stringhe sono considerate uguali se, rimuovendo gli spazi vuoti, sono identiche.

Costruzione del grafo dall'allineamento: Sia $G = (g1, ..., gm)$ un insieme di m genomi allineati, tutti di lunghezza n. Bisogna trovare un grafo di variazione G che sia compatibile con G.

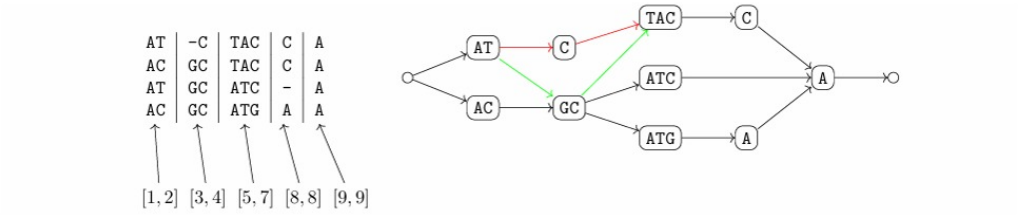

Figure 4.4: Esempio di allineamento e grafico di variazione

In Figura 4.4, vediamo un esempio di allineamento (sinistra) di quattro genomi ed un grafico di variazione corrispondente (destra). L'insieme I di intervalli disgiunti è nella parte inferiore sinistra delle figure ed ogni intervallo è collegato con l'insieme corrispondente di colonne dell'allineamento. Il grafico ha due vertici fittizi: una sorgente ed un pozzo, in modo che ciascuno sia ottimale per quasi tutte le istanze della prima formulazione. Il problema della costruzione di un grafico di variazione a partire da un allineamento di genomi chiede di trovare un grafico compatibile con i genomi allineati. Tuttavia, non esiste una funzione obiettivo chiara per scegliere tra i vari grafici possibili, rendendo il problema poco definito. Inoltre, alcune funzioni obiettivo semplici non portano ai grafici desiderati. Una proprietà importante di un grafico di variazione è che le sequenze condivise tra i genomi dovrebbero appartenere allo stesso vertice del grafico, ma un grafico che minimizza la funzione obiettivo potrebbe avere vertici separati per le sequenze condivise. Per risolvere questo problema, si potrebbero usare funzioni obiettivo che minimizzano il numero di vertici o la lunghezza totale delle etichette dei vertici. Tuttavia, anche queste soluzioni possono non essere ideali, poiché non sempre distinguono tra grafici compatti (dove i vertici sono etichettati con stringhe) e non compatti (dove ogni vertice ha un'etichetta singola). Inoltre, un allineamento multiplo non può sempre rappresentare alcune variazioni strutturali, come inversioni o trasposizioni, e in questi casi non è possibile partire da un allineamento affidabile per costruire il grafico di variazione. Nel contesto delle implementazioni dei grafi pangenomici, sono state sviluppate diverse soluzioni per rappresentare e gestire in modo efficiente i dati. Queste

implementazioni differiscono principalmente per l'approccio utilizzato nella codifica degli archi e dei vertici, nonché per la gestione della memoria e delle prestazioni. Di seguito vengono descritti tre dei principali approcci adottati. La prima implementazione, VG [10] utilizza una tabella hash per rappresentare gli archi, ma richiede troppa memoria. Una seconda implementazione, XG [11], è statica, il che significa che i vertici e gli archi non possono essere aggiornati. Utilizza bitvector per codificare i vertici e le liste di adiacenza, risultando in una struttura veloce ed efficiente in termini di memoria. La terza implementazione, ODGI [12], rappresenta gli archi e i cammini tramite delta encoding, dove viene memorizzata solo la differenza tra gli identificatori di due vertici consecutivi. Questo encoding, quando il grafo è simile a un singolo cammino (cosa che accade nella maggior parte dei casi pratici), offre un buon compromesso tra prestazioni di esecuzione e utilizzo di memoria. Un problema più pratico è come memorizzare un grafo pangenomico in un file. Il formato più utilizzato per questo scopo è GFA, inizialmente proposto per rappresentare i grafi di assemblaggio. Si tratta di un formato testuale per rappresentare grafi etichettati. La principale limitazione di GFA deriva dal suo scopo originale: poiché un grafo di assemblaggio non ha una connessione diretta con il genoma di riferimento lineare, un file GFA non garantisce un sistema di coordinate valido per l'intero grafo. Per superare questo problema, è stata proposta un'estensione chiamata rGFA (Li et al. 2020), in cui viene selezionato un cammino di riferimento che stabilisce un sistema di coordinate per il cammino stesso. Successivamente, ogni vertice del grafo viene associato a un vertice del cammino di riferimento per ottenere un sistema di coordinate per l'intero grafo. In altre parole, rGFA considera solo cammini che corrispondono a varianti semplici del cammino di riferimento, ovvero non sono consentiti cicli nel grafo. Tale approccio comporta delle limitazioni, ma superarle è una sfida teorica.

4.3 Indicizzazione di un grafo

Affinche possa esserci un accesso rapido e diretto alle sequenze e alle informazioni genetiche contenute in un grafo, è necessario indicizzarlo. Ciò consente di accedere direttamente alla risorsa senza dover esplorare tutto il grafo una volta che viene eseguita una ricerca.

L'indicizzazione di un grafo pangenomico può avvenire in vari modi, ma generalmente implica la creazione di strutture che associano ogni sequenza o variante del grafo a una posizione specifica, rendendo il recupero delle informazioni più rapido e mirato. Questi sono alcuni approcci chiave:

- **Rappresentazione compatta e strutture di dati**: Per ridurre il consumo di memoria e migliorare le prestazioni, si usano tecniche come l'encoding delta (ad esempio in odgi), che codifica le differenze tra i vertici consecutivi, o l'uso di bitvector e tabelle hash. Queste tecniche consentono di rappresentare in modo efficiente un grafo pangenomico e di indicizzare rapidamente le varianti e i cammini.

- **Coordinate spaziali**: L'indicizzazione può essere basata su un sistema di coordinate, come quello proposto in rGFA. In questo caso, viene selezionato un cammino di riferimento che stabilisce un sistema di coordinate per il grafo, associando ogni vertice a una posizione specifica. Le varianti o i cammini alternativi vengono quindi indicizzati rispetto a questa posizione di riferimento, facilitando il recupero.

- **Strutture di indicizzazione per il recupero rapido**: Vengono impiegate strutture di dati come gli alberi di ricerca o le tabelle di hash per associare rapidamente le sequenze genomiche alle loro posizioni nel grafo. Queste strutture consentono di localizzare velocemente varianti specifiche o porzioni di sequenza all'interno di un grafo pangenomico complesso.

- **Supporto per la ricerca di varianti**: L'indicizzazione deve supportare

la ricerca di varianti geniche, che possono essere presenti in molteplici individui o in forme diverse. A tal fine, l'indicizzazione può essere progettata per memorizzare e localizzare rapidamente varianti strutturali, come inserzioni, delezioni o SNP (polimorfismi a singolo nucleotide).

Chapter 5

Implementazione

5.1 Utilizzo di Pangraph per la Visualizzazione dei Grafi Pangenomici

Pangraph è un software progettato per la visualizzazione, l'analisi e l'esportazione di grafi pangenomici, sviluppato per supportare l'esplorazione della variabilità genetica a livello di popolazione. Questo strumento si distingue per la sua interfaccia user-friendly e le funzionalità avanzate, che consentono di interagire con rappresentazioni grafiche di dati genomici complessi. L'obiettivo principale è quello di facilitare lo studio e l'interpretazione delle varianti genetiche attraverso un approccio grafico.

Pangraph si colloca all'interno del contesto della pangenomica computazionale, affrontando le sfide descritte nei capitoli precedenti, come la gestione delle bolle genetiche e la rappresentazione di varianti rare. [13]

Una delle sue funzionalità principali è la possibilità di caricare grafi in un formato **GFA** (Graphical Fragment Assembly), ovvero un formato utilizzato per rappresentare **grafi di assemblaggio genomico**, consentendo di rappresentare **grafi di Brujin**, **String Graph** e altre strutture.

Tra le altre funzioni troviamo visualizzare questi file tramite una rappresentazione interattiva che consente agli utenti di:

- Zoomare e spostarsi nel grafo;
- Cliccare su un nodo per visualzzare informazioni dettagliate, come la seguenza genomica;
- Evidenziare le bolle presenti;

Una delle caretteristiche chiave di Pangraph è la possibilità di identificare ed evidenziare graficamente le varianti genetiche (bolle). Questo consente agli utenti di:

- Distinguere visivamente come SNP, inserzioni e delezioni;
- Differenziare le varianti con un codifica a colori, migliorando l'itnerpretazione dei dati genomici.

Gli utenti possono esportare le informazioni relative ai nodi del grafo in formato CSV, rendendo possibile l'integrazione con altri strumenti di analisi e la condizione dei risultati. I dati esportati includono: ID del nodo, sequenze associate ed eventuali annotazioni relative alle bolle.

5.2 Workflow di Utilizzo

Avvio del Software:

Per utilizzare Pangraph, l'utente deve scaricare il progetto dalla repository Github [13]. Una volta installate le dipendenze richieste, il programma viene eseguito tramite il comando:

```
python -m src.app
```

L'applicazione è accessibile tramite browser all'indirizzo `http://127.0.0.1:8050`.

Caricamento dei Dati:

L'utente può caricare un file GFA specificando il percorso del file nell'interfaccia. Un esempio di file semplice può includere:

```
S       1       ACCCTA
S       2       ACCGTA
L       1       +       2       +       1M
```

Dopo il caricamento, il grafo viene immediatamente visualizzato.

Interazione con il Grafo:

Cliccando su un nodo del grafo, l'utente può accedere a:

- L'ID del nodo.
- La sequenza associata.
- Informazioni aggiuntive sul ruolo del nodo (ad esempio, se parte di una bolla).

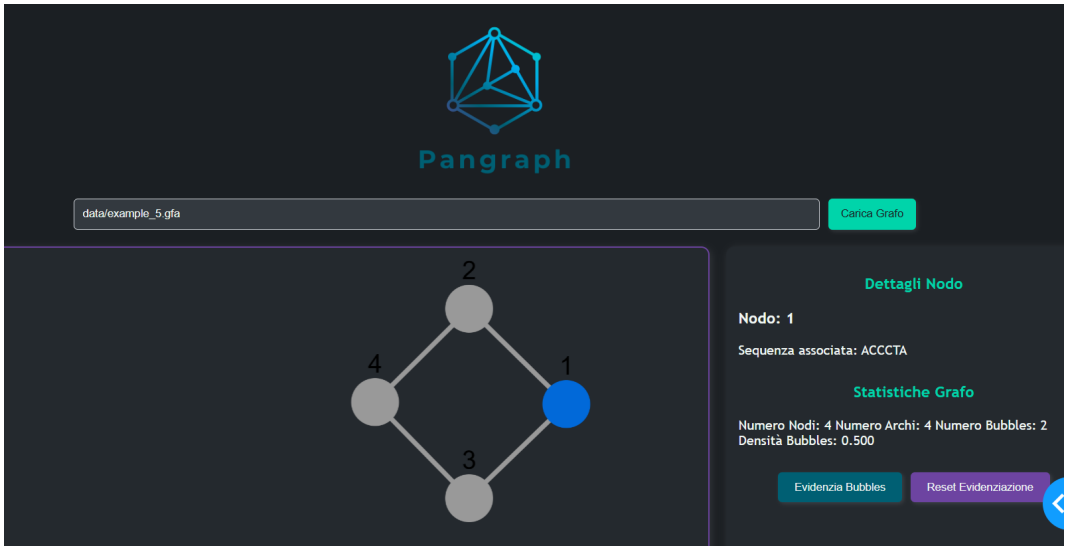

Figure 5.1: Pangraph - Rappresentazione Grafo

Evidenziazione, Reset ed Esportazione:

Con il pulsante *Evidenzia Bubbles*, Pangraph colora i nodi coinvolti in bolle genetiche, distinguendoli dal resto del grafo. Il pulsante *Reset Evidenziazione* consente di tornare alla visualizzazione originale. Infine, con il pulsante *Esporta in CSV*, l'utente può salvare i dati dei nodi in un file CSV per ulteriori analisi.

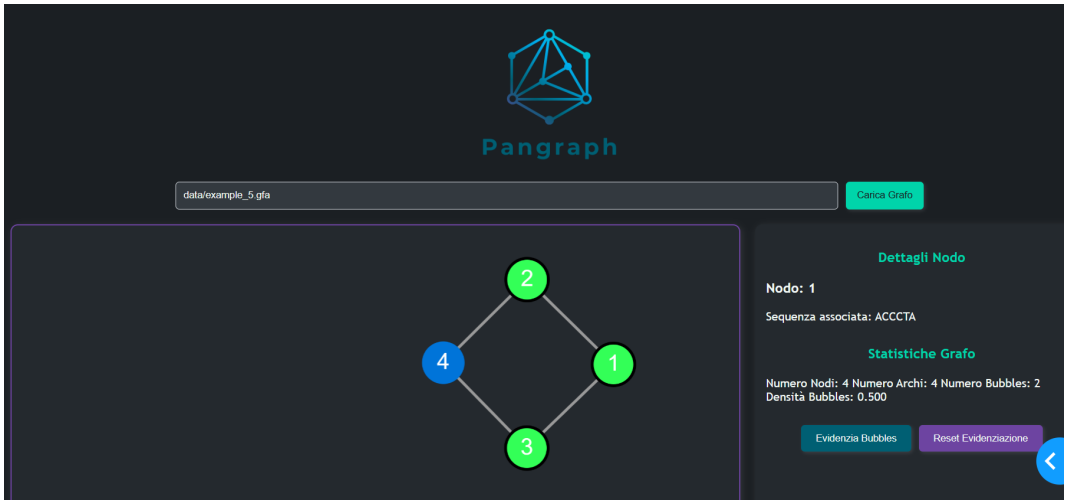

Figure 5.2: Pangraph - Evidenziazione Bolle

Applicazioni di Pangraph

Grazie alla capacità di identificare varianti genetiche rare, Pangraph può essere impiegato nel contesto clinico per:

- Diagnosticare malattie genetiche;
- Personalizzare terapie basate sul genotipo di un paziente;

Nel contesto Biotecnologico e Agriculturale può essere di supporto per:

- Identificare geni utili per il miglioramento delle colture;
- Analizzare resistenze genetiche in diverse popolazioni vegetali;
- Analisi comparative tra genomi di una stessa specie;
- Studio della diversità genetica a livelo di popolazione;

Pangraph rappresenta quindi un ponte tra la complessità della pangenomica computazionale e la necessità di strumenti intuitivi e accessibili. Grazie alle sue funzionalità interattive e alla capacità di gestire grafi complessi, si pone come un alleato fondamentale per ricercatori e professionisti in ambiti che spaziano dalla bioinformatica alla genomica applicata.

Chapter 6

Conclusione

La pangenomica rappresenta un cambiamento di paradigma nell'analisi genomica, permettendo di suprerare i limiti del modello. In particolare, i grafi stanno emergendo come uno strumento fondamentale, grazie alla loro capacità di rappresentare tutte le varianti genetiche senza essere limitati dai bias legati ai riferimenti. Tuttavia, nonostante le grandi potenzialità, la loro adozione è ancora limitata dalla carenza di strumenti efficaci per gestire e nalaizzare questi grafi in maniera semplice e veloce.

Tuttavia, l'integrazione delle variazioni nel sistema di riferimento comporta delle difficoltà. In particolare le fasi di costruzione, indicizzazione e allineamento richiedono generalmente più tempo rispetto ai modelli lineari tradizionali. A causa di queste problematiche, viene sostenuto che i modelli lineari continueranno a giocare un ruolo importante nel futuro della pangenomica.

Pangraph rappresenta un ponte tra la complessità della pangenomica computanzionale e la necessità di strumenti intuitivi e accessibili. Grazie alle sue funzionalitò interattive e alla capacità di gestire grafi complessi, si pone come un alleato per ricercatori e professionisti in ambiti che spaziano dalla bioinformatica alla genomica applicata.

References

[1] *Pan-genome*. https://en.wikipedia.org/wiki/Pan-genome.

[2] Hang Li, Xiaowen Feng, and Chong Chu. *The design and construction of reference pangenome graphs*. https://arxiv.org/pdf/2003.06079. 2020.

[3] Jordan M. Eizenga et al. "Pangenome Graphs". In: *Annual Review of Genomics and Human Genetics* (2020).

[4] Jasmijn A. Baaijens et al. "Computational graph pangenomics: a tutorial on data structures and their applications". In: *Natural Computing* (2022).

[5] R. Li et al. "Building the sequence map of the human pan-genome". In: *Nature Biotechnology* 28 (2010), pp. 57–63.

[6] Faber-Hammond JJ and Brown KH. "Anchored pseudo-de novo assembly of human genomes identifies extensive sequence variation from unmapped sequence reads". In: *Human Genetics* 135 (2016), pp. 727–740.

[7] Sherman RM et al. "Assembly of a pan-genome from deep sequencing of 910 humans of African descent". In: *Nature Genetics* 51 (2019), p. 30.

[8] Maretty L et al. "Sequencing and de novo assembly of 150 genomes from Denmark as a population reference". In: *Nature* 548 (2017), p. 87.

[9] Zhongqu Duan et al. "HUPAN: a pan-genome analysis pipeline for human genomes". In: *Genome Biology* (2019).

[10] Erik Garrison et al. “Variation graph toolkit improves read mapping by representing genetic variation in the reference”. In: *Nature Biotechnology* ().

[11] Guarracino A, Heumos S, and Nahnsen S. *ODGI: understanding pangenome graphs*. https://doi.org/10.1101/2021.11.10.467921.

[12] Erik Garrison. “Graphical pangenomics”. PhD thesis. University of Cambridge, 2019.

[13] Jacopo De Dominicis and Martina Giugliano. *Pangraph: Visualizzatore di Grafi Pangenomici*. https://github.com/Jacopodd/Pangraph. Disponibile su GitHub. 2025. URL: https://github.com/Jacopodd/Pangraph.

List of Figures