UNIVERSITÀ DEGLI STUDI DI SALERNO

Progetto Strumenti Formali per la Bioinformatica

"La pangenomica"

Dott.essa Liguori Serena

Matricola:0522501547

Anno accademico 2022/2023

Sommario

Introduzione

La pangenomica computazionale è un campo di ricerca in via di sviluppo che sta cambiando il modo in cui gli informatici affrontano le sfide nell'analisi delle sequenze biologiche. Negli ultimi decenni, i contributi della combinatoria, della teoria dei grafi e delle strutture dati sono stati essenziali nello sviluppo di strumenti software per l'analisi del genoma umano. Questi strumenti hanno permesso ai biologi computazionali di avvicinarsi a progetti ambiziosi su scala di popolazione, come il 1000 Genomes Project. Attualmente, la necessità di tenere conto dell'elevata variabilità dei genomi delle popolazioni nonché della specificità di un singolo genoma in un approccio personalizzato alla medicina sta rapidamente spingendo all'abbandono del paradigma tradizionale dell'utilizzo di un singolo genoma di riferimento. Una rappresentazione grafica di più genomi o un grafo del pangenoma, sta sostituendo il genoma di riferimento lineare.

1 Che cos'è un pangenoma?

Un pangenoma modella l'insieme completo di elementi genomici in una data specie o clade (è un gruppo di organismi monofiletici , cioè composti da un antenato comune e da tutti i suoi discendenti lineari su un albero filogenetico).

La pangenomica è quindi in contrasto con gli approcci genomici basati su riferimenti che mettono in relazione le sequenze con un particolare modello di consenso del genoma (**Figura 1**)

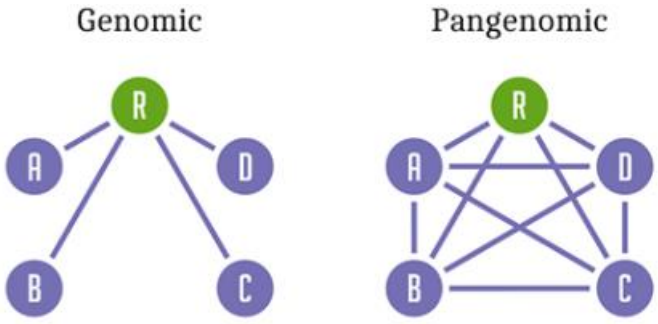

Figura 1: *a sinistra*: nelle analisi genomiche basate sui riferimenti, tutti i genomi (*A … D*) vengono confrontati tra loro tramite la loro relazione con il genoma di riferimento *R;*
a destra: in un contesto pangenomico, tentiamo di modellare le relazioni dirette tra tutti i genomi nella nostra analisi, di cui un particolare riferimento *R* è scelto arbitrariamente.

La pangenomica è stata importante per la microbiologia, dove la diversità genomica l'hanno resa indispensabile, e ha visto sempre più applicazioni ai genomi eucariotici. Le analisi pangenomiche standard si focalizzano sulla presenza o assenza di geni da determinati ceppi e sulla determinazione di un pangenoma centrale (comunemente presente) e accessorio (spesso assente). Tuttavia, hanno avuto l'inclinazione a prestare meno attenzione alla variazione tra queste sequenze e non tentano di fornire un modello preciso che colleghi tra loro molti genomi di base. Al contrario, la maggior parte delle analisi "ad alto rendimento" di grandi genomi dipende dal confronto con un singolo genoma di riferimento. Negli ultimi anni, il sequenziamento ridotto e i costi del de novo assembly hanno supportato la scoperta di livelli significativi di variazione genomica su larga scala in molte specie eucariotiche, inclusi gli esseri umani, arabidopsis, lievito di birra e la mosca della frutta.

Queste osservazioni hanno generato un vasto interesse nell'estendere le operazioni bioinformatiche per utilizzare un modello di riferimento pangenomico. La disponibilità di veri riferimenti pangenomici per l'uomo e altri organismi renderanno sempre più subottimale l'uso di un singolo genoma di riferimento. Ad ogni modo, l'utilizzo efficace dei pangenomi richiede lo sviluppo di nuovi metodi bioinformatici in grado di costruire, interrogare e operare su di essi.
L'interesse nel sostituire i genomi di riferimento lineari con i modelli dei grafi del pangenoma è in gran parte aumentato con la scoperta di limitazioni nell'esecuzione di vari compiti, come la *mappatura della lettura* e *l'identificazione delle varianti*.

2 Limitazioni di un genoma di riferimento lineare

Una variante strutturale (SV) è una mutazione genomica che coinvolge 50 o più paia di basi. L'introduzione di un'accurata tecnologia di sequenziamento a lettura lunga per il rilevamento di SV ha rivelato un numero ancora maggiore di variazioni candidate in un singolo genoma rispetto al genoma di riferimento. La scoperta di così tante varianti ha fatto luce sulla principale limitazione dei riferimenti lineari: le letture campionate da un individuo che porta determinati SV potrebbero non allinearsi al riferimento, in quel caso la lettura è spesso considerata un artefatto e scartata.
Poiché la mappatura delle letture è ancora un passaggio cruciale nella maggior parte delle analisi per l'identificazione di varianti genetiche collegate alla malattia, le applicazioni cliniche devono andare oltre il genoma di riferimento lineare.
Sono stati identificati altri limiti di un riferimento lineare, come le difficoltà nell'introdurre cambiamenti nell'attuale riferimento e il fatto che non catturi sufficientemente la diversità della popolazione.
Inoltre, ci sono alcuni chiari vantaggi nell'usare un riferimento al pangenoma:

- ridurre il bias di riferimento (ridurre la non rielaborazione di letture del campione);
- aumentare l'accuratezza della mappatura durante il sequenziamento di un nuovo individuo;
- aumentare l'accuratezza dell'identificazione di varianti rare;
- migliorare il de novo assembly di un nuovo individuo.

3 Modelli pangenomici

Un *modello pangenomico* è una struttura dati che rappresenta, come detto precedentemente, le sequenze genomiche di una popolazione, una specie, un clade. Il modello funge da entità di coordinamento centrale per descrivere la raccolta di sequenze e genomi nel pangenoma. I modelli pangenomici possono assumere molte forme, ma noi ci concentreremo su quelli dei grafi.

3.1 Modelli grafi

I grafi del pangenoma sono stati proposti come nuovo paradigma per rappresentare i genomi di riferimento. Questa è una rappresentazione naturale poiché i grafi forniscono una struttura dati compatta e concisa per l'esecuzione di diversi compiti, incluse le classiche operazioni di ricerca. Le rappresentazioni basate su grafi del genoma umano possono codificare un gran numero di varianti, come quelli riportati da The 1000 Genomes Project Consortium (2015). L'adozione di grafi del pangenoma nell'esecuzione di attività per l'analisi e il confronto di genomi in presenza di variazioni è solo all'inizio, ma tali approcci di pangenomica hanno dimostrato di superare gli approcci a genoma di riferimento singolo. Non è possibile rappresentare varianti strutturali complesse con l'uso di un singolo genoma di riferimento. Le 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 ad esempio la duplicazione, le inversioni e la traslocazione di segmenti del genoma. Un grafo è una struttura più appropriata per rappresentare rielaborazioni tra più genomi, poiché l'orientamento di bordi, cicli e strutture complesse in un grafo, come le bolle (è un sottografo aciclico diretto determinato da una coppia di vertici, un vertice sorgente s e un vertice terminale t tale che tutti i percorsi da s a t sono vertici disgiunti), rappresentano varianti strutturali in modo che possano essere gestite da algoritmi e strutture dati adatte a indici e grafi di interrogazione.

Data una raccolta di sequenze genomiche, un problema fondamentale nella pangenomica è come costruire un grafo che riassuma i genomi. Ma prima di esaminare tale argomento, dobbiamo considerare varie definizioni:

I ***grafi di variazione*** sono grafi orientati i cui vertici sono etichettati da stringhe non vuote. G=(V, A, W) dove V indica la funzione di etichettatura, A denota l'insieme di archi e W denota un insieme non vuoto di percorsi distinti.

In tale definizione i percorsi corrispondono a sequenze che vogliamo mantenere nella nostra rappresentazione. È possibile che si voglia rappresentare le varianti che sono compatibili con un insieme di variazioni di sequenza. Questo porta alla definizione di grafi di sequenza. I grafi di sequenza rappresentano l'insieme di andamenti di un grafo di variazione ma poiché questi andamenti non sono esplicitamente etichettati, cioè distinti, sono rappresentate anche le varianti non nell'insieme di input che sono indotte dagli archi del grafo di variazione (**Figura 2**).

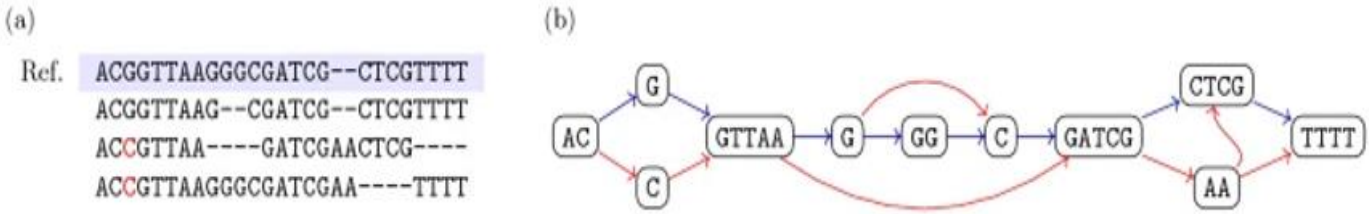

Figura 2: Esempio di una variante rappresentata nel grafo ma non nei genomi di input

a) Un allineamento di sequenze multiple di un genoma di riferimento lineare e altri tre genomi che contengono variazioni rispetto al riferimento;

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

- La definizione di grafo di variazione che abbiamo fornito è semplice e può essere adattata a diversi contesti. Nel caso in cui vogliamo rappresentare un insieme di genomi, il grafo di variazione è chiamato **grafo del genoma**: utilizzati per rappresentare le intere relazioni del genoma. I percorsi attraverso questi grafi rappresentano le ricombinazioni dei genomi inclusi nel modello.

I **grafi di sequenza** sono stati utilizzati per la prima volta per rappresentare più allineamenti di sequenza. È un grafo orientato i cui vertici sono etichettati da stringhe non vuote. G=(V,A)dove V indica l'etichetta funzione e A denota l'insieme degli archi. È possibile notare che un grafo di sequenza è un grafo di variazione G=(V,A,W) con lo stesso insieme di vertici, con W costituito da tutti i possibili cammini nel grafo. Per questo motivo, le proprietà dei grafi di variazione valgono anche per i grafi di sequenza.
Il *grafo di sequenza* serve a comprimere molte sequenze di input ridondanti in una struttura dati più piccola che è ancora rappresentativa dell'insieme completo. I grafi di sequenza possono avere i loro nodi o bordi etichettati con sequenze di DNA, ma per semplicità ci concentreremo sul caso etichettato come nodo. In un grafo di sequenza etichettato come nodo, i bordi indicano quando si verificano le concatenazioni dei nodi che collegano nelle sequenze modellate dal grafo.

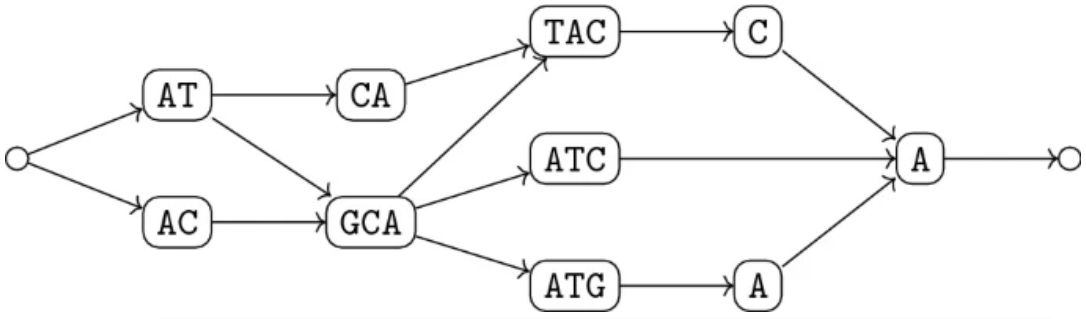

Figura 3: Esempio di grafo di sequenza con due vertici fittizi: una sorgente e un pozzo.

4 Costruire un pangenoma

- Un pangenoma può essere rappresentato come una **raccolta di sequenze**. Diversi approcci supportano la costruzione, l'annotazione e l'interrogazione di queste raccolte di sequenze pangenomiche.
  PANSEQ [1] trova nuove regioni, determina il nucleo e il genoma accessorio.
  PGAP [2] estende l'approccio di ***PANSEQ*** con moduli per l'analisi evolutiva e funzionale ed è implementato come singolo eseguibile autonomo. Il lavoro recente si è concentrato sul ridimensionamento di queste tecniche a genomi sempre più grandi.
  HUPAN [3] estende il modello di raccolta di sequenze ai genomi eucariotici umani e di grandi dimensioni, prendendo genomi assemblati come input e trovando sequenze non di riferimento al loro interno rispetto a un genoma di riferimento.
  HUPAN utilizza la strategia "map-to-pan". Ha una serie di miglioramenti:
  (1) l'assemblaggio de novo di ogni singolo genoma viene eseguito con SGA, un programma che richiede poca memoria;
  (2) viene creata una strategia di estrazione di sequenze non di riferimento più rapida;
  (3) si ritiene che sia le sequenze completamente non allineate che le sequenze parzialmente non allineate generino le regioni genomiche non di riferimento;
  (4) viene proposto un rigoroso processo di screening per distinguere le sequenze non umane dalle sequenze non di riferimento.
  Figura 4 mostra il diagramma di costruzione del pan-genoma in HUPAN.

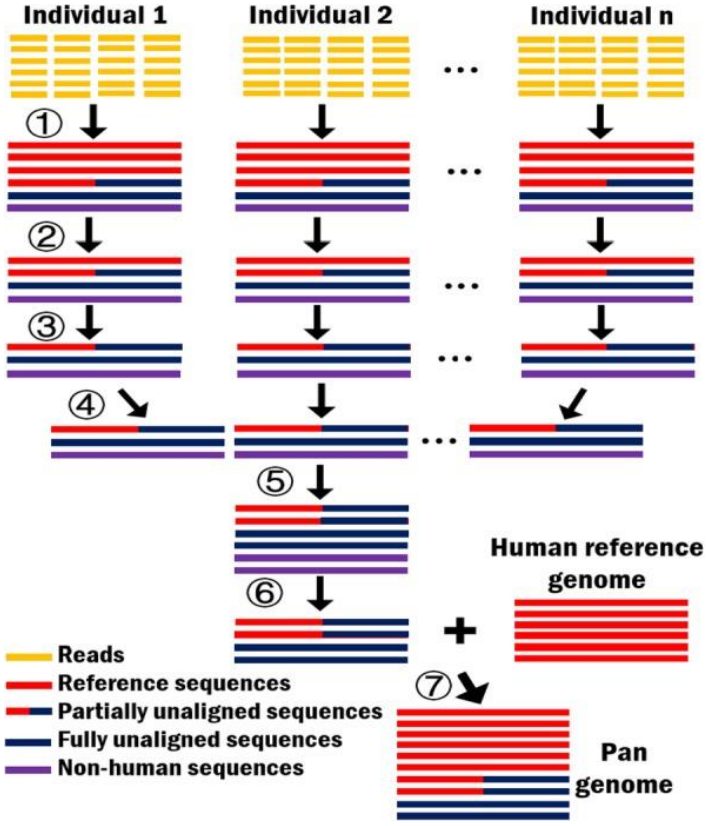

Figura 4:
① assemblaggio de novo tutte le letture in contig,
② rimozione di contig simili al genoma di riferimento umano,
③ estrazione di sequenze non allineate (comprese sequenze completamente non allineate e sequenze parzialmente non allineate),
④ unione di sequenze non allineate da più individui,
⑤ rimozione sequenze ridondanti,
⑥ rimuovendo potenziali contaminazioni e
⑦ costruendo un pan-genoma combinando il genoma di riferimento umano e nuove sequenze

- Piuttosto che raccogliere sequenze uniche che rappresentano una raccolta di genomi, possiamo considerare **piccole varianti** tra la raccolta e un genoma di riferimento. Tale modello implica direttamente un grafo aciclico diretto, ordinato lungo il genoma di riferimento, con bolle nei siti di variazione. Questo approccio di costruzione del pangenoma è utilizzato in diversi mappatori di lettura del grafo del genoma. Decidere quale variazione dovrebbe essere aggiunta a un grafo non è banale e ha incoraggiato studi sull'utilità del grafo e algoritmi per determinare quale variazione è utile.
- Agli assemblatori basati su grafi di **De Bruijn** può essere assegnata una qualità pangenomica attraverso l'aggiunta di "colori" ai loro nodi. Ogni colore fornisce una mappatura tra uno specifico campione biologico e un sottoinsieme del grafo. **CORTEX [4]** ha dimostrato per la prima volta che i grafi di De Bruijn colorati potevano eseguire analisi su scala di popolazione con un'efficiente implementazione di grafi.

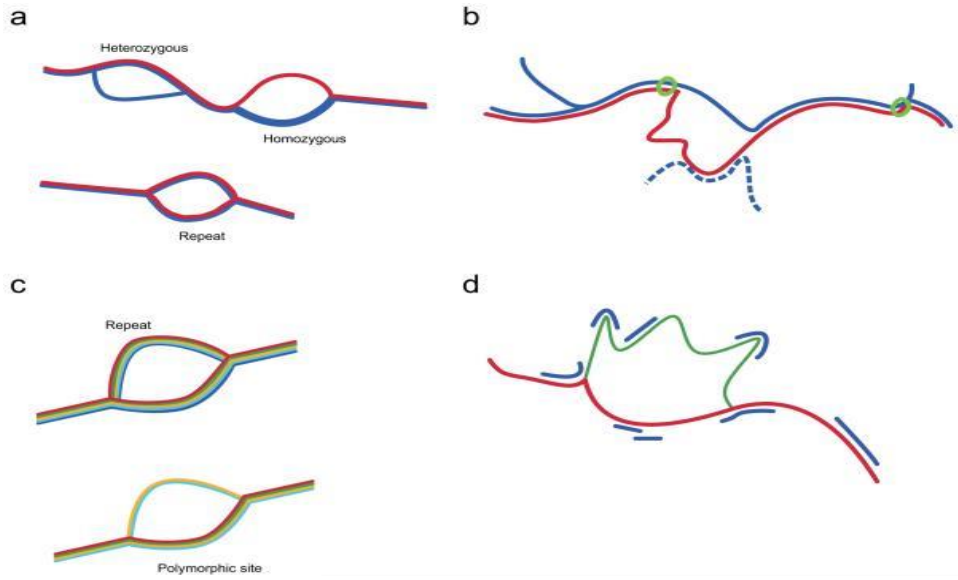

Figura 5: Rappresentazione schematica di quattro metodi di analisi della variazione utilizzando grafi di de Bruijn colorati;

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

b) Anche quando l'allele di riferimento (rosso) non forma una bolla pulita, possiamo identificare siti varianti omozigoti tracciando la divergenza del percorso di riferimento da quello del campione. Una volta trovato un breakpoint, prendiamo il contig più lungo del campione (cioè il percorso fino all'incrocio successivo) 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).

c) 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.

d) La probabilità di un dato genotipo può essere calcolata dalla copertura (blu) di ciascun allele (verde, rosso), tenendo conto dei contributi di 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.

Recenti miglioramenti alla costruzione di grafi di De Bruijn colorati, come ***BIFROST [5]***, consentono la costruzione di grafi di De Bruijn colorati da insiemi di sequenze molto grandi e supportano ulteriormente aggiornamenti efficienti di questi modelli pangenomici. Numerosi metodi utilizzano la costruzione di De Bruijn compattata per elaborare grafi del pangenoma.

4.1 Problema

Un problema fondamentale nella pangenomica computazionale è costruire un grafo di variazione. Questo problema si presenta in due versioni, a seconda che l'input sia un insieme di sequenze o un allineamento multiplo delle sequenze.

Problema 1(costruzione del grafo dell'allineamento): Sia G= $\{g_1, \ldots, g_m\}$ un insieme di *m* genomi allineati, tutti di lunghezza *n*. La costruzione del grafo dal problema dell'allineamento chiede di trovare un grafo di variazione che sia compatibile con *G*.

Problema 2 (costruzione del grafo dei genomi): *Sia G=* $\{g_1, \ldots, g_m\}$ *un insieme di m genomi. Il problema della costruzione del grafo dei genomi chiede di trovare un grafo d variazione G che esprime tutti i genomi in G.*

Questo problema è più generale del Problema 1, poiché non esiste una divisione in blocchi da rispettare per tutti i genomi (**Figura 7**).

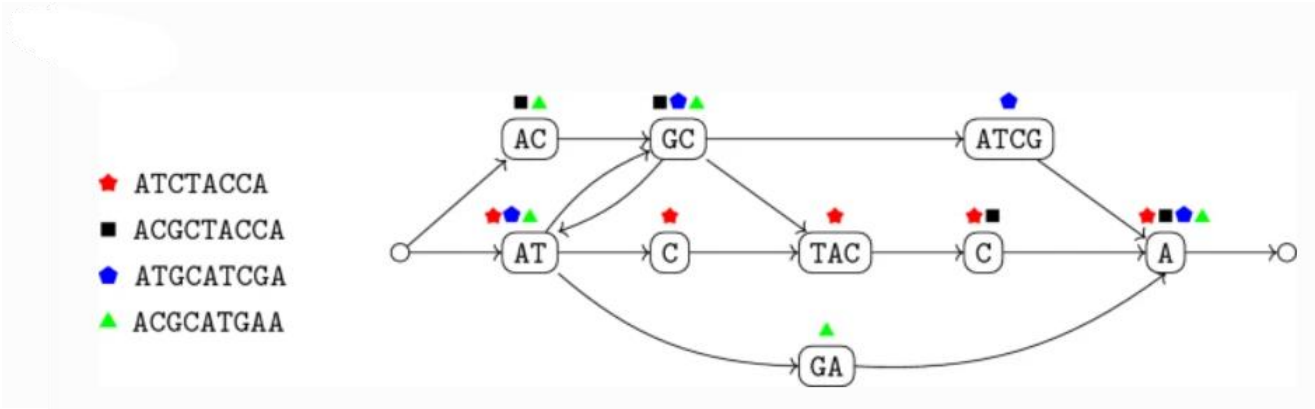

Figura 7: Esempio di un grafo di variazione costruito da quattro sequenze, ciascuna rappresentata da un simbolo di colore diverso. Coloriamo solo i vertici per semplificare la figura.
In questo caso ogni sequenza è allineata rispetto al grafo delle variazioni (la prima sequenza è anche il grafo iniziale.

Possiamo considerare un grafo di variazione come una struttura dati astratta per la quale sono state proposte alcune implementazioni concrete **[6].** Tali implementazioni presentano diversi compromessi. Ad esempio, non tutti consentono facilmente aggiornamenti nel grafo delle variazioni, ovvero utilizzano strutture dati dinamiche. Inoltre, utilizzano diverse strategie di compressione e memorizzano anche i filamenti, per consentire a un vertice di rappresentare due stringhe a complemento inverso. Descriviamo un modello leggermente semplificato, in cui due stringhe a complemento inverso sono rappresentate con due vertici che sono collegati insieme, ad esempio condividendo un identificatore per la coppia.

La ***prima implementazione***, **VG [7],** utilizza una tabella hash per rappresentare gli archi, ma ciò richiede troppa memoria.
Una ***seconda implementazione***, **XG [8]**, è invece statica, ovvero i vertici e gli archi non possono essere aggiornati. Utilizza vettori di bit per codificare i vertici e gli elenchi di adiacenza, risultando in una struttura veloce ed efficiente in termini di memoria.
La ***terza implementazione***, **ODGI [9]**, rappresenta archi e percorsi tramite codifica delta, in cui viene memorizzata solo la differenza tra gli identificatori di due vertici consecutivi. Si osservi che quando il grafo è simile a un singolo percorso (il che è vero in quasi tutti i casi pratici), questa codifica abbina una grande prestazione di runtime con un piccolo utilizzo della memoria. Notiamo che esistono altri approcci che forniscono un sistema di coordinate basato sull'insieme di percorsi, ad esempio **ODGI 2021[10]** (**ODGI è pubblicato come software libero sotto la licenza open source del MIT**).
Pur essendo un netto miglioramento rispetto ai metodi precedenti, l'ODGI ha due limitazioni:

- la coordinata di un vertice appartenente a due diversi percorsi non è intuitiva;
- un vertice che non appartiene a nessuno dei percorsi in W non ha coordinate.

Il superamento di queste due limitazioni è una sfida teorica.
Un problema più pratico è come memorizzare un grafo del pangenoma in un file. Il formato più utilizzato per questo scopo è **GFA**, inizialmente proposto per rappresentare i grafi di assemblaggio **[11].** È un formato testuale per rappresentare grafi etichettati. Il principale limite di GFA deriva dal suo scopo originario. Poiché un grafo di assemblaggio non ha una connessione diretta con il genoma di riferimento lineare, non è garantito che un file GFA fornisca un sistema di coordinate valido per l'intero grafo. Per superare questo problema, è stata applicata un'estensione, chiamata **rGFA [12]**, dove viene selezionato un cammino di riferimento e determina un sistema di coordinate per il cammino. L'rGFA considera solo cammini corrispondenti a varianti semplici del cammino di riferimento, cioè non sono ammessi cicli nel grafo.

5 Indicizzazione dei grafi del pangenoma

I grafi grandi come i grafi del genoma devono essere indicizzati per ottenere un'efficienza adeguata per operazioni di base come il pattern matching o il read mapping. Le strutture dati dell'indice per i grafi del pangenoma supportano un accesso casuale efficiente agli elementi e alle caratteristiche del grafo. Occorre prestare attenzione per garantire che queste strutture di indici non richiedano spese generali significative relative al contenuto informativo del grafo. Strutture dati succinte e un'attenta codifica di questi dati sono necessarie per adattare in modo affidabile grafi di grandi dimensioni nella memoria principale dei sistemi informatici di base. Particolari modelli di indici sono al centro della visualizzazione basata su grafi, della mappatura delle letture e dei sistemi di chiamata delle varianti.

5.1 Indicizzazione delle sequenze utilizzando un grafo

L'indice FM [13] è un indice di testo, basato sulla trasformata di Burrows-Wheeler (BWT) **[14],** che viene spesso utilizzato con sequenze di DNA. Una variante dell'indice FM, l'RLCSA **[15],** run-length, codifica la BWT, consentendogli di archiviare e indicizzare una raccolta di sequenze simili in modo efficiente in termini di spazio. Se conosciamo un buon allineamento globale delle sequenze, possiamo utilizzare tale informazione per rendere l'indice sia più piccolo che più veloce. Questo approccio è stato ulteriormente sviluppato nell'indice di allineamento FM.

5.2 Indicizzazione dei grafi aciclici

Una classe di metodi di indicizzazione dei grafi supporta solo i grafi aciclici, spesso rappresentati come grafi aciclici diretti (***DAG***). Questo vincolo può esistere sia perché l'aciclicità del grafo fornisce garanzie che semplificano il problema, sia perché caratteristiche accessorie dell'implementazione software del metodo ne precludono l'uso su grafi ciclici.

GENOME Mapper [16], il primo allineatore di lettura basato SU grafi, era limitato a tali grafi. Anche la sua indicizzazione era relativamente semplice. Utilizza un semplice indice *k* -mer basato su hash, con $k \leq 13$ per limitare l'utilizzo della memoria.

GCSA [17] è stato il primo tentativo di generalizzare la BWT per i grafi. Applica una serie di trasformazioni del grafo che preservano lo spazio della sequenza del grafo creando un ordinamento non ambiguo per i nodi. Quando la complessità del grafo è bassa, queste trasformazioni sono ragionevolmente veloci e non aumentano significativamente la dimensione del grafo. Tuttavia, a una certa soglia di densità delle varianti, il grafo trasformato diventa rapidamente troppo grande per essere gestito.

6 Mappatura del grafo del genoma

Negli ultimi anni sono stati sviluppati metodi efficienti per mappare le letture su grandi grafi del pangenoma, molti si basano su recenti ricerche sull'allineamento e sui progressi nell'indicizzazione, di cui abbiamo già parlato nei capitoli precedenti.

Sebbene questi strumenti di mappatura abbiano come target tutti i grafi di sequenza, vi sono differenze significative nei tipi di grafi che gestiscono. Diversi strumenti si applicano solo ai grafi di variazione aciclici formati aggiungendo varianti a un riferimento lineare.

La maggior parte di questi strumenti enfatizza la mappatura dei dati *di sequenziamento di nuova generazione (NGS)* a lettura breve. A nostra conoscenza, **GRAPH ALIGNER** e **V-MAP** sono gli unici strumenti di mappatura grafica progettati per dati di sequenziamento a lettura lunga **[18].**

Sebbene V-MAP supporti anche le letture NGS, la strategia di seeding di GRAPH ALIGNER lo limita alle letture lunghe.

Per l'indicizzazione, la maggior parte degli strumenti di mappatura dei grafi ha optato per qualche variazione di una tabella *k* -mer. GRAPH ALIGNER, GenomeMapper, Seven Bridges' mapper e V-MAP utilizzano tutti questa strategia.

La maggior parte dei mappatori di grafi utilizza algoritmi di allineamento basati su grafi. Le eccezioni sono GENOMEMapper, che si allinea a tutti i percorsi che PARTONO da un seme, e **HISAT2**. L'algoritmo di allineamento HISAT2 si basa su un insieme complesso di euristiche che dipendono fortemente dal suo indice di corrispondenza esatta. Questo lo rende eccezionalmente veloce, sebbene possa anche danneggiare la qualità dell'allineamento attorno agli indel.

A causa del recente sviluppo di questi metodi, sono stati condotti pochi studi comparativi indipendenti sulle loro prestazioni e accuratezza. GRAPH ALIGNER è l'unico mappatore che incorpora le ricerche più recenti sugli algoritmi di allineamento dei grafi. Utilizza un algoritmo di allineamento a bande per ottenere una velocità impressionante allineando letture lunghe ai grafi del genoma **[18].**

7 Applicazioni dei modelli pangenomici

Sebbene le tecniche pangenomiche grafiche possano essere applicate in tutta la biologia, il lavoro più recente si è concentrato su una manciata di applicazioni in cui possono fornire vantaggi sostanziali. La riduzione del bias di riferimento e le rappresentazioni coerenti degli alleli producono miglioramenti significativi nel rilevamento delle variazioni strutturali e possono ridurre i costi di esecuzione della genotipizzazione.

7.1 Identificazione delle varianti e la genotipizzazione

In genere, l'identificazione delle varianti e la genotipizzazione indicano diversi aspetti di un processo di inferenza del genoma guidato da riferimenti. La genotipizzazione consiste nel determinare se una variante osservata in precedenza è presente in un nuovo campione, mentre l'identificazione delle varianti comporta il rilevamento di variazioni ancora non osservate. Quando il nostro sistema di riferimento è un genoma lineare, questi due passaggi sono spesso uniti. Un singolo processo rileverà la variazione candidata e dedurrà un genotipo campione in ciascun locus variabile. Utilizzando l'identificazione delle varianti multi-campione, la fasatura del genotipo e l'imputazione del genotipo, possiamo condividere le informazioni tra i campioni per migliorare l'accuratezza della nostra ricostruzione dei genomi da letture brevi. Tuttavia, questo approccio di chiamata congiunta è costoso e non applicabile quando abbiamo solo pochi nuovi genomi da ricostruire. Inoltre, non aiuta la nostra interpretazione primaria di nuovi dati di sequenziamento durante la mappatura di lettura.
L'allineamento a qualsiasi allele in un grafo di riferimento del pangenoma è efficiente quanto l'allineamento all'allele di riferimento in una sequenza di riferimento lineare.

8 Aplotipo e genotipizzazione in pangenomica e pantrascrittomica

La trasformata di Graph Burrows-Wheeler è stata recentemente utilizzata da Sirén et al. (2020) per costruire in modo efficiente un indice dell'intero genoma di 5.008 aplotipi di 1KGP. È importante notare che il GBWT presentato da Sirén et al. (2020) è diverso dal grafo BWT posizionale originale proposto da Novak et al. (2017) e porta a una rappresentazione più pratica ed efficiente degli indici sensibili all'aplotipo, ovvero indici dei grafi del pangenoma in cui i percorsi rappresentano gli aplotipi distinti negli individui.
Sibbesen et al. (2021) hanno utilizzato il GBWT per rappresentare un grafo del pangenoma per gli aplotipi annotato con le informazioni aggiuntive di un grafo di splicing. Quindi la quantificazione delle trascrizioni dai dati RNA-seq viene ottenuta tenendo conto delle informazioni sull'aplotipo e quindi implementando un allineatore RNA-seq al grafo del pangenoma. L'idea principale di Sibbesen et al. (2021) consiste nel rappresentare gli esoni di un grafo di giunzione direttamente in un grafo del pangenoma mappando gli esoni alle sequenze aplotipiche del grafo del pangenoma. In questo modo, propongono uno strumento per la mappatura dei dati RNA-seq in grado di tenere conto delle variazioni aplotipiche nell'analisi dei trascritti.

9 Ricostruzione dell'aplotipo virale

Un'altra applicazione della pangenomica computazionale si pone nell'assemblaggio del genoma virale. Durante l'infezione, i virus replicano il proprio genoma miliardi di volte utilizzando un meccanismo di replicazione soggetto a errori, quindi molti dei genomi risultanti sono copie inesatte. Questi sono anche indicati come aplotipi virali, che insieme formano un pangenoma virale.

Una delle principali sfide nella ricostruzione dell'aplotipo virale è la grande quantità di letture e l'alto grado di somiglianza tra quelle letture. Ciò richiede algoritmi di costruzione di grafi altamente efficienti. Un'altra sfida consiste nell'acquisire la variazione all'interno di un campione filtrando attentamente eventuali errori di sequenziamento. Queste sfide vengono affrontate utilizzando diversi tipi di grafi e traggono grande vantaggio dai progressi nelle rappresentazioni del pangenoma. La **Figura 8** presenta un'istanza di un set di dati di sequenza virale per illustrare le strutture di dati.

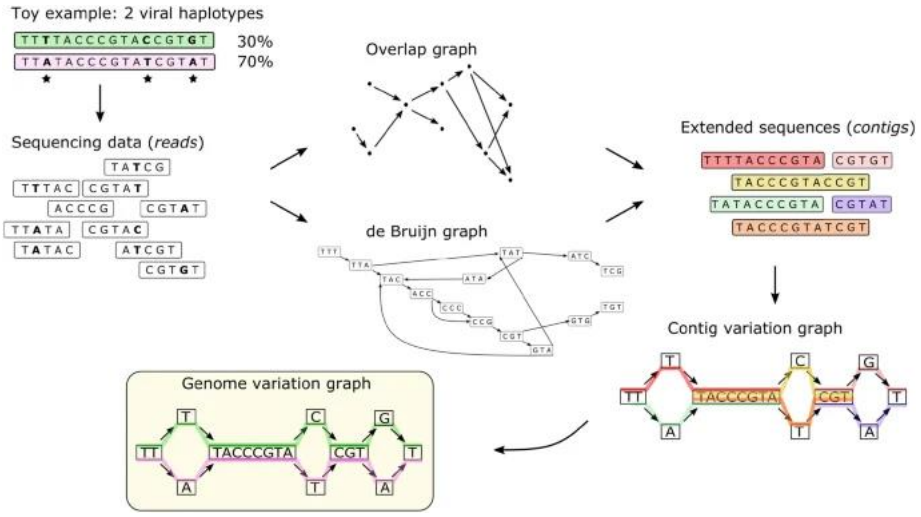

Figura 8: Un esempio per illustrare il processo di assemblaggio dell'aplotipo virale.
In questo esempio, il compito è ottenere il grafo della variazione del genoma (un pangenoma virale) ricostruendo gli aplotipi virali dai dati di sequenziamento, con aplotipi presenti in diverse abbondanze (30% vs. 70%). Le stelle sotto le sequenze originali indicano le tre posizioni in cui i due aplotipi differiscono.

Le tre strutture di dati coinvolte nel processo di assemblaggio sono:

(1) un grafo di sovrapposizione, in cui i vertici rappresentano letture di sequenziamento e gli archi indicano sovrapposizioni suffisso-prefisso;

(2) un grafo di de Bruijn, dove i vertici rappresentano k-meri e gli archi indicano sovrapposizioni di lunghezza k−1;

(3) un grafo di variazione, dapprima costruito dalle sequenze estese (contigs) ottenute attraverso l'assemblaggio del genoma, che può essere trasformato in un grafo di variazione del genoma che rappresenta gli aplotipi a lunghezza intera.

Si noti che questo esempio è una rappresentazione semplicistica della realtà: gli errori di sequenza non vengono mostrati; quindi, tutte le sovrapposizioni tra le letture sono esatte.

10 Conclusione

I grafi del pangenoma stanno per diventare un modello onnipresente nella genomica grazie alla loro capacità di rappresentare qualsiasi variante genetica senza essere influenzati da bias di riferimento. Tuttavia, nonostante questo grande potenziale, la loro diffusione è ostacolata dalla mancanza di strumenti in grado di gestire e analizzare i grafi del pangenoma in modo semplice ed efficiente.

Nel prossimo futuro, prevediamo di ottenere prontamente a basso costo assiemi completi, risolti per aplotipo. Le fasi di costruzione, indicizzazione e allineamento del modello *richiedono in genere più tempo per i grafi del pangenoma rispetto ai genomi di riferimento lineari*.

A causa di questi problemi, alcuni sostengono che è probabile che i modelli genomici lineari rimarranno importanti anche in futuro. Molti dei lavori che abbiamo considerato prevedono un futuro in cui i sistemi di riferimento sono grafi, ma solo una manciata (principalmente quelli basati su grafi di variazione) produce allineamenti o chiamate genotipiche nel contesto di un grafo pangenomico.

11 Bibliografia

[1] Laing C, Buchanan C, Taboada EN, Zhang Y, Kropinski A, et al. 2010. Analisi della sequenza pan-genomica utilizzando Panseq: uno strumento online per l'analisi rapida delle regioni genomiche centrali e accessorie. *BMC Bioinformatica* 11 :461.

[2] Zhao Y, Wu J, Yang J, Sun S, Xiao J, Yu J. 2011. PGAP: pipeline di analisi pan-genomi . *Bioinformatica* 28 :416–418

[3] Duan Z, Qiao Y, Lu J, Lu H, Zhang W, et al. 2019. HUPAN: una conduttura di analisi del pan-genoma per i genomi umani . *Biologia del genoma* 20 :149.

[4] Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. 2012. Assemblaggio de novo e genotipizzazione di varianti utilizzando grafi de Bruijn colorati. *Nature Genetics* 44 :226–232

[5] Holley G, Melsted P. 2019. Bifrost - Costruzione altamente parallela e indicizzazione di grafi de Bruijn colorati e compatti. *bioRxiv*

[6] Vaddadi K, Srinivasan R, Sivadasan N. 2019. *Read mapping on genome variation graphs* :17

[7] Eizenga JM, Novak AM, Sibbesen JA et al (2020) Pangenome graphs. Annu Rev Genomics Hum Genet 21(1):139–162. https://doi.org/10.1146/annurev-genom-120219-080406

[8] Garrison E, Sirén J, Novak A et al (2018) Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol 36:875–879. https://doi.org/10.1038/nbt.4227

[9] Garrison E, et al (2019) seqwish: A variation graph inducer. https://github.com/ekg/seqwish

[10] Guarracino A, Heumos S, Nahnsen S, et al (2021) ODGI: understanding pangenome graphs. bioRxiv:2021.11.10.467921. https://doi.org/10.1101/2021.11.10.467921

[11] Li H, Chin J, Durbin R, et al (2017) GFA: Graphical Fragment Assembly (GFA) Format Specification. http://gfa-spec.github.io/GFA-spec/

[12] Li H, Feng X, Chu C (2020) The design and construction of reference pangenome graphs with minigraph. Genome Biol. https://doi.org/10.1186/s13059-020-02168-z

[13] Ferragina P, Manzini G. 2005. Indicizzazione del testo compresso . *Diario dell'ACM* 52 :552–581

[14] Burrows M, Wheeler DJ. 1994. Un algoritmo di compressione dei dati senza perdita di ordinamento dei blocchi.

[15] Mäkinen V, Navarro G, Sirén J, Välimäki N. 2010. Archiviazione e recupero di raccolte di sequenze altamente ripetitive

[16] Schneeberger K, Hagmann J, Ossowski S, Warthmann N, Gesing S, et al. 2009. Simultaneous alignment of short reads against multiple genomes. *Genome Biology* 10: R98

[17] Sirén J, Välimäki N, Mäkinen V. 2014. Indexing graphs for path queries with applications in genome research. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 11:375–388

[18] Rautiainen M, Marschall T. 2019. GraphAligner: Rapid and versatile sequence-to-graph alignment. *bioRxiv :810812*

Survey: Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, Chang X, Seaman JD, Rounthwaite R, Ebler J, Rautiainen M, Garg S, Paten B, Marschall T, Sirén J, Garrison E.

Pangenome Graphs. Annu Rev Genomics Hum Genet. 2020 Aug 31;21:139-162. doi: 10.1146/annurev-genom-120219- 080406. Epub 2020 May 26. PMID: 32453966; PMCID: PMC8006571.

Survey: Baaijens, J.A., Bonizzoni, P., Boucher, C. et al. Computational graph pangenomics: a tutorial on data structures and their applications. Nat Comput 21, 81–108 (2022). https://doi.org/10.1007/s11047-022-09882-6