Università degli Studi di Salerno
Dipartimento di Informatica
Corso di Laurea Triennale in Informatica

Metodi Efficienti per il Trattamento di Sequenze Genomiche: Un'Analisi di Kallisto, Minimap e Miniasm

PROFESSORI
Prof.ssa Rosalba Zizza
Prof.ssa Clelia De Felice
Prof. Rocco Zaccagnino

CANDIDATO
Alessandro Carnevale
Matricola: 0522501994

Tabella dei contenuti

Introduzione

L’analisi delle sequenze genomiche è un campo fondamentale per comprendere la struttura e la funzione del DNA e dell’RNA.

Necessità

Analizzare le sequenze genomiche in modo efficiente

Problema

Gestire la grande quantità di dati generata dalle moderne tecnologie di sequenziamento

Obiettivo

Valutare Kallisto, Minimap e Miniasm per quantificazione, mapping e assemblaggio

RNA-seq

Frammenti di RNA presenti in un trascrittoma vengono sequenziati generando milioni di letture corte che devono essere successivamente analizzate per **determinare il livello di espressione dei trascritti**.

Principali problemi

- **Ambiguità nell'assegnazione delle letture**: lo splicing alternativo genera più isoforme che condividono regioni comuni.
- **Bias dovuti alla lunghezza dei trascritti**: trascritti più lunghi hanno maggiore copertura, richiedendo normalizzazione.

Assemblaggio

Processo di **ricostruzione della sequenza del DNA** partendo da frammenti più piccoli (**reads**) generati dalle tecnologie di sequenziamento.

Principali problemi

- **Ripetizioni genomiche**: possono generare ambiguità nei percorsi di assemblaggio.
- **Errori di sequenziamento**: introducono artefatti che complicano la ricostruzione.

Metodi di assemblaggio

- **Grafo di De Bruijn**: basato su k-mer, efficiente per letture corte.
- **Overlap-Layout-Consensus**: basato sull'identificazione delle sovrapposizioni tra letture.

Analisi delle Sequenze

Metodi Alignment-based

Le letture vengono **allineate esplicitamente** a un genoma o trascrittoma di riferimento.

Vantaggi

- **Alta accuratezza** nella determinazione della posizione delle letture.

Svantaggi

- **Computazionalmente oneroso**: richiede tempo e molta memoria per il confronto delle letture.
- **Sensibile a errori di sequenziamento**: gli errori possono influenzare l'allineamento corretto.

Analisi delle Sequenze

Metodi Alignment-free

Non eseguono un allineamento esplicito, ma **analizzano direttamente le letture** attraverso statistiche basate su **k-mer** o altre rappresentazioni compatte.

Vantaggi

- Più **veloci**, ideali per grandi dataset genomici e trascrittomici.
- Più **robusti agli errori di sequenziamento**, poiché non si basano su una corrispondenza esatta.

Svantaggi

- **Meno precisi** nel rilevare mutazioni e variazioni strutturali.

Tool analizzati

Kallisto

Uno strumento progettato per la **quantificazione dell'espressione genica** a partire da dati RNA-seq, fornendo una soluzione efficiente e veloce.

Principi di funzionamento

- Lo **pseudoallineamento**, alla base di Kallisto, identifica i trascritti di origine di una lettura senza determinarne le coordinate esatte, mantenendo un'elevata efficienza e un'accurata quantificazione.
- Utilizza un **Transcriptome De Bruijn Graph (T-DBG)** per indicizzare il trascrittoma in modo efficiente e velocizzare l'identificazione dei trascritti di origine delle letture.
- Utilizza l'algoritmo **Expectation-Maximization (EM)** per stimare l'abbondanza dei trascritti nei dati RNA-seq e applica il **bootstrapping** per quantificare l'incertezza nelle stime.

Tool analizzati

Minimap

Un software per il **mapping di long reads**, basato su **minimizer**, individuando rapidamente regioni omologhe tra letture e genoma di riferimento.

Flusso di lavoro

1. **Indicizzazione**: le sequenze target vengono convertite in una rappresentazione compatta basata sui minimizer, riducendo il consumo di memoria.
2. **Estrazione dei minimizer**: le letture vengono processate per identificare i minimizer da confrontare con la sequenza di riferimento.
3. **Matching approssimato**: i minimizer delle letture vengono confrontati con quelli della sequenza target per individuare regioni omologhe.
4. **Filtraggio e clustering**: i risultati vengono raffinati per migliorare la precisione del mapping, riducendo errori e ambiguità.

Tool analizzati

Miniasm

Un **assembler de novo** ottimizzato per la velocità, che genera rapidamente una bozza dell'assemblaggio senza una fase di correzione degli errori integrata. Lavora in combinazione con Minimap, che identifica le sovrapposizioni tra le letture.

Flusso di lavoro

1. **Identificazione delle sovrapposizioni**: Minimap rileva sovrapposizioni tra le letture, riducendo la complessità computazionale dell'allineamento all-vs-all.
2. **Costruzione del grafo di sovrapposizione**: Miniasm crea un overlap graph, in cui i nodi rappresentano le letture e gli archi indicano sovrapposizioni significative.
3. **Pulizia del grafo**: eliminazione di connessioni ridondanti, tips e bubbles.
4. **Generazione delle unitigs**: estratte direttamente dalla topologia del grafo, **senza fase di consenso**, mantenendo il tasso di errore delle letture originali.
5. **Esportazione in formato GFA**: il risultato viene salvato in un file GFA, contenente la struttura del grafo e le unitigs assemblate.

Risultati e Conclusioni

Kallisto

- Elevata efficienza computazionale grazie allo pseudoallineamento.
- Accuratezza nelle stime dell'abbondanza dei trascritti.
- Non rileva nuovi trascritti o splicing non annotati.

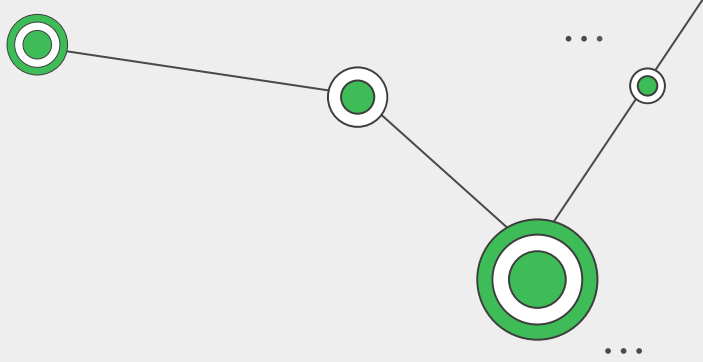
Minimap

- Rapido e a basso consumo di memoria.
- Buona accuratezza nella ricerca di regioni omologhe.
- Difficoltà nella gestione di regioni ripetitive.

Miniasm

- Alta velocità grazie all'eliminazione della fase di consenso.
- Evitando la correzione degli errori di lettura, richiede una successiva fase di polishing.

I metodi **alignment-free** offrono un compromesso tra efficienza computazionale e accuratezza, risultando una risorsa ideale in contesti in cui velocità e scalabilità sono prioritari.



Grazie per l'attenzione