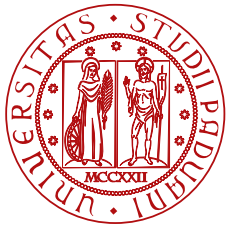


Università degli Studi di Padova

Dipartimento di Ingegneria dell'informazione

Corso di laurea in Ingegneria Informatica


# Clustering a più fasi di un metagenoma


Laureando

**Francesco Tomasella**

**Matricola 2008793**

Relatore

**Prof. Cinzia Pizzi**

**Università degli Studi di Padova**



Anno Accademico 2022/2023

Data di laurea 29/09/2023

**Abstract**

L'importanza crescente della metagenomica ha reso indispensabile lo sviluppo di varie tecniche di analisi dei dati, tra le quali spicca l'analisi tassonomica delle popolazioni microbiche. In questo contesto, sono stati sviluppati differenti approcci di raggruppamento delle specie noti come "binning", che possono essere generalmente suddivisi in due categorie principali: il binning supervisionato, basato su dati di riferimento, e il binning non supervisionato, che opera in modo reference-free.
L'obiettivo fondamentale di questa ricerca è dimostrare come l'impiego sequenziale di due strumenti di binning, AbundanceBin e MetaProb, conduca a una migliorata aggregazione delle comunità microbiche rispetto all'utilizzo singolo di uno di tali strumenti. Questo approccio si basa sullo sfruttamento della complementarità intrinseca tra i vari tool per il binning, con alcuni di questi che dimostrano maggiori performance in presenza di variazioni nell'abbondanza delle specie, mentre altri sono preferibili quando le diverse specie presentano un livello di abbondanza simile all'interno del campione.
Mediante l'utilizzo sequenziale di AbundanceBin e MetaProb in modo non supervisionato si riscontra un miglioramento delle prestazioni rispetto all'uso degli stessi singolarmente, indicando una effettiva correttezza dell'ipotesi.
Nell'ottica di miglioramento dell'analisi metagenomica e della comprensione delle comunità microbiche, sarà necessario proseguire in futuro con studi più estesi su dataset differenti.

# Indice

# 1 Introduzione

Recentemente, la metagenomica ha guadagnato una sempre più grande rilevanza all'interno del panorama della ricerca scientifica. La metagenomica rappresenta lo studio delle comunità microbiche ottenute direttamente da un ambiente, senza la necessità di colture microbiche. Questa metodologia assume una particolare importanza in quanto alcune specie potrebbero non sopravvivere al processo di trasporto fino al laboratorio o potrebbero deperire in condizioni ambientali differenti da quelle d'origine e non ottimali.

Nel contesto più ampio della metagenomica, negli ultimi vent'anni si è assistito alla nascita e allo sviluppo di tecniche per l'analisi tassonomica delle specie presenti nelle comunità microbiche, tramite un processo chiamato binning. Questo approccio consiste nell'aggregare insieme le sequenze appartenenti alla stessa specie.
Il binning ha assunto un ruolo di primaria importanza in quanto la conoscenza e distribuzione delle specie permette di stimarne aspetti come l'abbondanza e di definire come esse interagiscono tra loro o come avviene la loro evoluzione.

La caratterizzazione delle specie riveste una notevole importanza nella scoperta di nuove comunità batteriche, sia negli ambienti esterni che interni. In particolare, le ricerche sul microbiota intestinale hanno guadagnato sempre più rilevanza, permettendo la scoperta continua non solo di nuove specie ma anche delle loro interazioni [7]. L'analisi metagenomica di queste comunità batteriche costituisce un passo fondamentale il cui progresso è ampiamente attribuibile alle nuove ed avanzate tecniche di sequenziamento, tra cui le tecnologie di sequenziamento di nuova generazione NGS (Next Generation Sequencing), e alla progressiva riduzione dei costi associati al sequenziamento stesso.

Lo scopo del presente studio è quello di testare l'impiego a cascata di più di un tool per l'analisi del metagenoma tramite binning per verificare se questo si traduce in un miglioramento delle prestazioni globali nella suddivisione delle sequenze tra le diverse specie. Nell'ambito di questa ricerca, sono stati utilizzati due software: Abun-

danceBin [10] e MetaProb [1]. Il primo è un metodo di binning basato sull'abbondanza delle specie, pertanto è particolarmente indicato per situazioni caratterizzate da elevata variabilità nella distribuzione delle specie. Il secondo, invece, segue un approccio basato sulla composizione del DNA, risultando più appropriato in contesti in cui le specie presenti nel campione condividono un livello di abbondanza molto simile. In questo modo, i due strumenti per il binning si completano, offrendo in linea teorica l'opportunità di raggiungere miglioramenti delle prestazioni nell'analisi tassonomica del metagenoma.

Gli esperimenti sono stati condotti su un dataset simulato scelto appositamente per permettere il test delle prestazioni di entrambi gli strumenti di binning e scelto anche in modo da consentire un confronto con il caso in cui un solo tool di binning venga usato.

All'interno di questa tesi, inizieremo con la spiegazione di alcuni concetti fondamentali per il nostro caso di studio, per poi spiegare il funzionamento dei due tool presi in esame. Dopo aver fatto la valutazione dei risultati ottenuti, verrà poi fatto un confronto con le performance nell'uso dei sotfware per il binning usati singolarmente.

All'interno del presente lavoro di ricerca, inizieremo con l'esposizione dei concetti chiave rilevanti per il nostro specifico ambito di studio. Successivamente, verrà condotta una dettagliata analisi del funzionamento dei due strumenti per il binning presi in esame.
Una volta completata la valutazione dei risultati ottenuti, procederemo infine con una comparazione delle prestazioni del nostro approccio con i software usati singolarmente.

# 2

# Concetti Fondamentali

Prima di cominciare con la descrizione dell'esperimento, è necessario definire alcuni concetti utili a definire il problema e il suo background.

Iniziamo definendo il clustering, ovvero un modo per raggruppare insiemi di oggetti simili basandosi sulle loro somiglianze o dissimilarità [4]. E' un problema che fa parte dell'apprendimento non supervisionato, il cui scopo è trovare pattern o strutture ricorrenti all'interno di dati non etichettati. Oltre alla bioinformatica, il clustering viene utilizzato nel machine learning, nel data mining, nel riconoscimento di pattern e nell'analisi di immagini, per fare qualche esempio.
Il clustering può essere ulteriormente suddiviso in base alla scelta operativa del come creare i cluster. Possiamo infatti avere il clustering gerarchico e il clustering partizionale. Il primo può seguire approcci agglomerativi o top-down oppure divisivi o bottom-up, e in generale trova i cluster con successive iterazioni sui cluster già trovati. Nel secondo caso invece i cluster vengono trovati tutti nello stesso momento, solitamente fornendo già in input il numero desiderato di cluster da formare. Un esempio per quest'ultimo tipo di raggruppamento è senza dubbio il clustering facente uso dell'algoritmo K-means, mentre per il clustering gerarchico è necessario fornire una misura della distanza, un cui esempio è l'utilizzo della distanza euclidea.

La diffusione delle tecinche di *high-throughput next-generation sequencing* (NGS) è stata una rivoluzione per il mondo della ricerca scientifica e ha portato con sè un gran numero di vantaggi, dal costo al tempo necessari per il sequenziamento.
Le varie tecniche di sequenziamento di nuova generazione possono essere divise in base al tipo di read genomiche che restituiscono: alcune come PacBio [6] e Oxford Nanopore [3] permettono di ottenere read, ovvero sequenze di DNA, molto lunghe (anche decine di migliaia di bp), mentre altre come Illumina restituiscono read molto corte (50-300 bp). Il vantaggio delle long read risulta quello di fornire maggiori informazioni riguardo ad alcune strutture e ripetizioni del DNA, a scapito del tasso di errore di sequenziamento che è più alto. Le short read hanno il vantaggio di avere un costo

minore, e un throughput, una copertura e una precisione molto alte.

A differenza della genomica, in cui le read sono tutte di una stessa specie, nella metagenomica le read appartengono a genomi di specie diverse che abitano l'ambiene da cui è stato preso il campione.
Per questo motivo, di fondamentale importanza è conoscere il problema del binning metagenomico: l'analisi tassonomica delle comunità microbiche che consiste nel clustering delle read in gruppi specifici per studiarne la diversità [5]. Esistono molteplici criteri per classificare gli strumenti di binning in diverse categorie.
Uno di essi è quello relativo alla modalità di sequenziamento delle read, che porta a 2 possibilità: lo shotgun sequencing, in cui l'intero genoma viene sequenziato, e l'approccio amplicon-based, in cui solo alcune parti di interesse del genoma sono sequenziate.
I tool per il binning possono anche essere separati tra quelli dipendenti dalla tassonomia e quelli che non lo sono. I primi sono considerate tecniche di supervised learning e sono anche definiti reference-based, in quanto permettono di confrontare le read con un database di riferimento, segnando come non assegnate le read non abbastanza simili ai vari dati già presenti all'interno del database. Tra questo tipo di metodi indichiamo ad esempio MEGAN [2] e Kraken [9].
Gli strumenti taxonomy-independent, invece, non confrontano i dati con un database in nessuno step dell'eleborazione, basandosi solamente sulle similarità tra i dati.

Un'altra modalità di classificazione è legata agli approcci utilizzati per il binning. La maggior parte dei metodi si basa sull'analisi della composizione del DNA. In questa categoria, troviamo, ad esempio, MetaProb [1], BiMeta [8] e MetaCluster [11]. Tuttavia, uno dei principali limiti degli approcci composition-based è la decrescita delle prestazioni in presenza di read molto corte.
Ci sono però anche metodi di binning basati sull'abbondanza delle specie, come AbundanceBin [10]. Essi si basano sull'osservazione che la distribuzione dei k-mer in uno stesso genoma è più simile a se stesso rispetto a quella di altri genomi.

# 3 Metodi

In questo capitolo si descriveranno gli approcci metodologici dei due tool analizzati in questa tesi, AbundanceBin e MetaProb, e si descriverà l'approccio sperimentale utilizzato.

## 3.1 AbundanceBin

AbundanceBin è un tool per il binning metagenomico basato sull'abbondanza delle specie [10]. Il suo punto di forza è rappresentato dalla sua capacità di produrre risultati soddisfacenti nel binning anche in situazioni nelle quali le read sono molto corte (intorno alle 75 bp). Il tool lavora in maniera non supervisionata, non richiedendo alcuna informazione riguardante il numero dei bin, in modo simile ai dati che si ottengono in situazioni reali e non simulate, delle quali non conosciamo la composizione dei campioni.

AbundanceBin assume che la distribuzione delle read segua il modello di Lander-Waterman, per cui la coverage delle varie posizioni dei nucleotidi è modellata tramite una distribuzione di Poisson. La procedura di sequenziamento metagenomica può essere vista come un insieme di distribuzioni di Poisson, con ognuna di queste posta a rappresentare una specie differente.
In presenza di $m$ specie diverse, quindi, è possibile individuare $m$ distribuzioni di Poisson. La media di ciascuna di queste distribuzioni, definita da $\lambda$ , rappresenta l'abbondanza delle specie ed è quindi l'elemento che è necessario calcolare per raggiungere una stima corretta relativamente alla loro abbondanza. AbundanceBin si basa quindi sulla risoluzione di un problema di ottimizzazione tramite l'utilizzo di un algoritmo di Expectation-maximization (EM). Una volta raggiunta la convergenza dell'algoritmo EM, è possibile calcolare la probabilità di assegnazione di una read ad un bin, anche se c'è la possibilità che la read rimanga non assegnata. All'algoritmo EM è però necessario fornire in input il numero di bin. Per aggirare questo problema, AbundanceBin adotta

un approccio ricorsivo che si basa sulla divisione del dataset in due bin, proseguendo successivamente con ulteriori suddivisioni dei bin fino a quando non si ottengono bin dall'abbondanza molto diversa, dimensioni del genoma e numero di reads in ogni bin sopra una certa soglia percentuale sul totale di reads del bin genitore.

AbundanceBin dimostra avere delle ottime performance in situazioni nelle quali l'abbondanza delle specie è diversa, seppur non inferiore ad un rapporto 1:2. Nei casi in cui ci sia meno variabilità e le specie abbiano un'abbondanza paragonabile, AbundanceBin non risulta più essere una scelta ottimale e presenta tassi di errore molto alti, poiché con alta probabilità raggrupperà negli stessi bin specie differenti accomunate però da un'abbondanza simile.

## 3.2 MetaProb

MetaProb è un tool il cui obiettivo è affrontare il problema del binning metagenomico [1]. Il processo di binning in MetaProb è diviso in 2 fasi distinte, come visibile in Fig. 3.1:

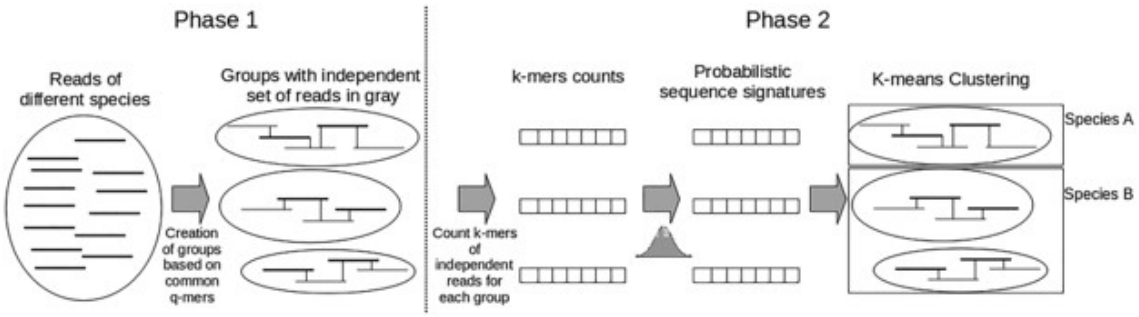


Figura 3.1: Pipeline di lavoro di MetaProb [1]

Nella prima fase le read delle diverse specie presenti nel campione vengono raggruppate in base alla loro capacità di sovrapposizione, misurata attraverso lo strumento dei q-mer, sottostringhe di lunghezza q estratte dalla sequenza nucleotidica utili per la valutazione dell'overlap. Read appartenenti alla stessa specie, infatti, hanno una maggior probabilità di venire raggruppate insieme per via delle loro sovrapposizioni.

Si parte quindi con ogni read che forma un gruppo a sé stante, successivamente avviene l'unione di due gruppi che si verifica di volta in volta solo nel momento in cui ci siano un sufficiente numero di q-mer in comune. Il passo successivo è quello di migliorare ulteriormente il raggruppamento tramite l'utilizzo dei k-mer, sotto-sequenze nucleotidiche di lunghezza k contenute nelle read. Per farlo, MetaProb non esegue una semplice analisi della frequenza dei k-mer ma utilizza una tecnica più sofisticata basata sugli insiemi indipendenti, che garantisce una minore ridondanza in modo da permettere un migliore conteggio dei k-mer.

Nella seconda fase, tramite tecniche di misurazione della distanza, avviene un ulteriore confronto e raggruppamento degli insiemi creati in precedenza. MetaProb, a

differenza di molti altri metodi di binning, per la misura non usa la distanza Euclidea, in quanto vi è la presenza di rumore, ma un nuovo metodo basato sulle signatures delle sequenze. Questo procedimento fa parte degli approcci definiti alignment-free, molto efficienti dal punto di vista computazionale. Il metodo delle signature di sequenze probabilistiche si basa sulla distribuzione del conteggio dei k-mer e sulla rimozione del bias creato dai gruppi sbilanciati.

Finito questo passaggio, si passa al clustering tramite algoritmo k-means, con l'unione dei gruppi in un numero di cluster che può essere fornito o in alternativa stimato. Per la stima del numero di cluster finali, e quindi delle specie, viene usato un metodo ispirato da G-means ma che usa come test statistico il test Kolmogorov-Smirnov a due campioni.

Al sistema di elaborazione della stima nel suo insieme è stato dato il nome di SpeciesNumber: esso prevede che, se un cluster non dovesse passare il test, potrà essere suddiviso, oppure procederà con le sue iterazioni fino al momento nel quale tutti i centri dei k-means passeranno il test.

## 3.3 Approccio sperimentale

Per la verifica dell'ipotesi si è deciso di dividere l'esperimento in due fasi principali. Ognuna di queste corrisponde ad uno dei diversi tool scelti per lo scopo e fa parte della pipeline di lavoro volta a convalidare l'ipotesi, secondo cui l'utilizzo di più strumenti di binning metagenomico in sequenza, di cui uno basato sull'abbondanza e l'altro sulla composizione, porta ad un miglioramento del binning delle specie all'interno dei campioni.
L'idea è di ottenere dei cluster all'interno dei quali sono presenti tutte le read delle specie che hanno una stessa abbondanza con AbundanceBin (Fase 1) e poi separare le specie all'interno di ciascun cluster con MetaProb (Fase 2). Un esempio di pipeline ideale è presentato in Fig. 3.2.

In realtà, si è visto come il passaggio dalla Fase 1 alla Fase 2 non sia di fatto banale richiedendo ulteriori elaborazioni. In Fig. 3.3, si può vedere una schematizzazione più realistica della pipeline che andremo ora a descrivere.

La prima fase consiste nella fornitura dei dati di ingresso ad AbundanceBin, al fine di consentire a quest'ultimo di eseguire un processo di clustering basato sull'abbondanza delle specie. Il software accetta come input un singolo file nel formato FASTA, e offre tra le opzioni la possibilità di scegliere la lunghezza dei k-mer. E' poi possibile richiedere l'esecuzione di una classificazione ricorsiva per la stima del numero di bin presenti all'interno dell'input.

Nel processo di output, il software AbundanceBin produce una serie di file, tra cui quelli relativi ai cluster appena generati. Ciascun cluster è archiviato in un file separato, con l'intento di agevolarne la distinzione, e al suo interno sono incluse le read assegnate

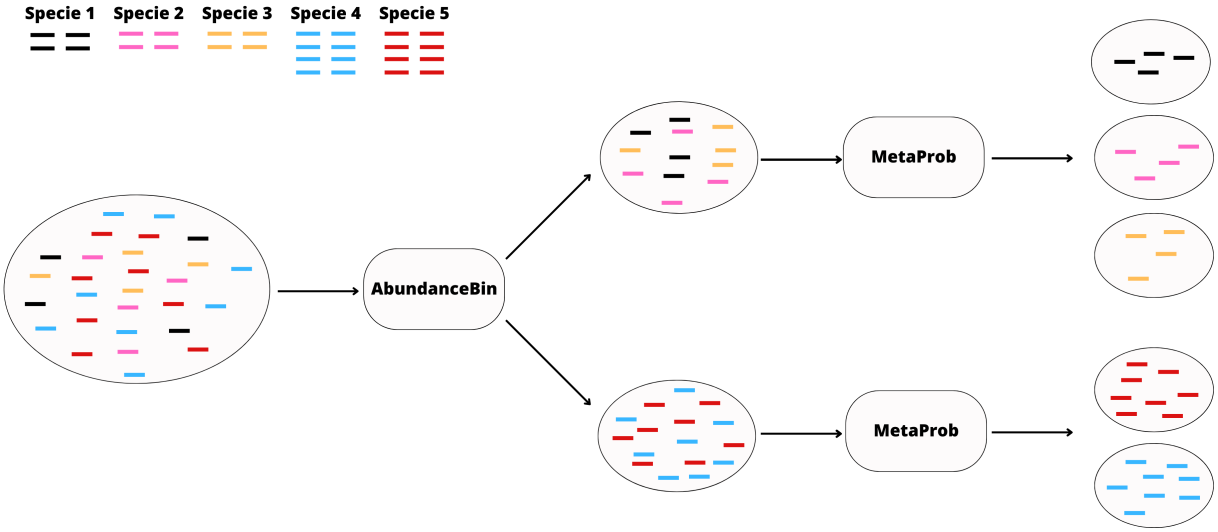


Figura 3.2: Pipeline di lavoro ideale

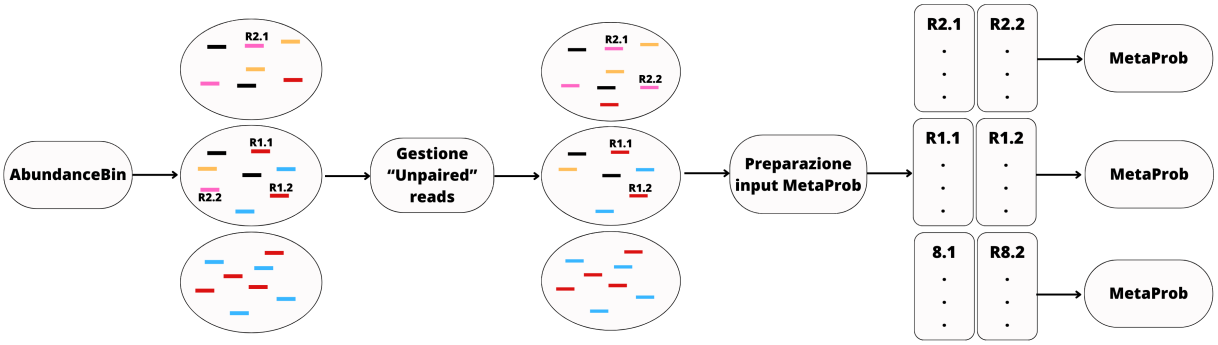


Figura 3.3: Pipeline più realistica, con le operazioni da eseguire tra un tool e l'altro (per motivi di spazio si è omessa la rappresentazione della fase di input ad AbundanceBin, che comunque rimane invariata).

a detto cluster, con la sola descrizione e senza quindi la riga di sequenza nucleotidica presente nel file di input. Al fine di agevolare l'analisi individuale dei singoli cluster e poter proseguire con l'esperimento, è pertanto necessario procedere con la ricostruzione di un file simile all'originale mediante l'accoppiamento di ciascun ID delle read con la sua sequenza nucleotidica corrispondente, trasformando conseguentemente il testo in un file FASTA completo e compatibile con le eventuali nuove computazioni.

Prima di entrare nel vivo della seconda fase, ovvero l'utilizzo del tool MetaProb, è fondamentale eseguire un ulteriore processo di elaborazione e preparazione dei file dei vari cluster precedentemente ottenuti. Ognuno di essi verrà utilizzato individualmente come input per MetaProb. Innanzitutto è necessario definire una strategia per la riassegnazione delle sequenze non accoppiate all'interno dei cluster, poiché AbundanceBin considera ogni read a sé stante e pertanto può assegnare read di fatto "accoppiate" dalla tecnologia di sequenziamento utilizzata in cluster diversi. MetaProb, invece, accetta come input solo file in cui le read sono esplicitamente accoppiate. Le opzioni a disposizione comprendono l'eliminazione diretta di tutte le read non accoppiate o la decisione di mantenerle e procedere con la riassociazione. La determinazione del cluster di de-

stinazione per ciascuna read risulta tuttavia complessa, richiedendo una valutazione caso per caso, basata sui risultati ottenuti e sulla composizione complessiva dei cluster. Nei nostri esperimenti abbiamo utilizzato due possibili approcci: i) riassegnazione delle read solo da cluster con percentuale di read molto elevata; ii) riassegnazione delle read spaiate a partire dal cluster con percentuale di read più elevata e iterazione della riassegnazione finché non restano più read spaiate.

Una volta ottenuti cluster completamente composti da read accoppiate, è possibile procedere alla suddivisione del file in due file distinti, conformemente alle specifiche richieste da MetaProb per l'input di paired-end reads. Un file conterrà tutte le read di un'estremità, mentre l'altro ospiterà le read corrispondenti all'altra estremità.

In aggiunta ai file di input, per l'esecuzione di MetaProb le impostazioni relative al software sono state mantenute ai valori predefiniti, i quali sono stati indicati dal software stesso come quelli ottimali in termini di prestazioni per le short paired-end reads. Per quanto concerne il numero di specie previste, tale informazione non è stata specificata, in accordo con l'obiettivo sperimentale. Di conseguenza, è stato attivato l'algoritmo K-means per stimare dinamicamente il numero di specie. La dimensione dei q-mer, parametro necessario alla creazione delle adiacenze dei grafi, è stata fissata a 5 come di default, e la feature selezionata per rappresentare le informazioni contenute nei vari gruppi è stata la feature 1, quella predefinita.

Al termine dell'esecuzione di MetaProb, è stata effettuata la raccolta dei dati riguardanti il numero di cluster generati e la distribuzione delle specie all'interno di essi, allo scopo di valutare eventuali miglioramenti nelle prestazioni conseguenti all'applicazione di un secondo strumento di binning metagenomico.

# 4 Risultati

In questo capitolo descriveremo il setting sperimentale, in particolare il dataset utilizzato, il formato dei dati, le misure di qualità del clustering utilizzate e discuteremo i risultati sperimentali ottenuti.

## 4.1 Dataset

Per l'analisi è stato usato uno dei dataset utilizzati in [1], in particolare S7 che si addice alla verifica dell'ipotesi che vogliamo testare, ovvero se l'utilizzo di più fasi di clustering porta ad un miglioramento dei risultati rispetto all'uso di un solo strumento di clustering.

Come è possibile vedere dalla Tabella 4.1, il dataset S7 è composto da short paired-end reads, suddivise tra 5 specie con abbondanze 1:1:1:4:4 e con distanza filogenetica a livello di ordine e genere a cui appartengono. Il dataset è stato realizzato dagli autori di [8] ed è stato simulato tramite MetaSim, un tool per la generazione di read metagenomiche, usando l'Illumina error profile con un tasso di errore dell'1%.

Nella Tabella 4.2 è possibile osservare le 5 diverse specie presenti all'interno del dataset e la coverage corrispondente.

| **Dataset** | **Specie** | **Distanza Filogenetica** | **Abundance Ratio** | **Totale Reads** |
|---|---|---|---|---|
| S7 | 5 | Ordine e Genere | 1:1:1:4:4 | 3 307 100 |

Tabella 4.1: Composizione del dataset S7

| Specie | Nome | Coverage |
|---|---|---|
| Specie 1 | Actinobacillus pleuropneumoniae serovar 5b str. L20 | 10 |
| Specie 2 | Aliivibrio salmonicida LFI1238 | 10 |
| Specie 3 | Haemophilus somnus 129PT | 10 |
| Specie 4 | Pasteurella multocida 36950 | 40 |
| Specie 5 | Vibrio cholerae M66-2 | 40 |

Tabella 4.2: Specie presenti all'interno di S7

La scelta di S7 come dataset per i test è basata sulla distribuzione delle abbondanze delle sequenze al suo interno: la presenza di specie con abbondanze uguali a gruppi, permette di effettuare un'analisi vicina allo scopo dell'esperimento e di verificare quindi che utilizzando congiuntamente metodi di binning basati sull'abbondanza e sulla composizione si ha un miglioramento dei risultati rispetto all'utilizzo di uno solo di essi.

## 4.2 Il formato FASTA

Il formato dei file contenenti le reads del dataset S7 è il formato FASTA, un formato di file testuale usato in ambito bioinformatico per rappresentare sequenze nucleotidiche o di amminoacidi. Ciascun elemento della sequenza, sia esso un nucleotide o un amminoacido, è rappresentato da una lettera.

Il file è strutturato in modo che ogni sequenza sia composta da una prima riga di descrizione, che inizia tramite il simbolo > , seguito da un identificatore di sequenza (SeqID). Possono poi essere aggiunte altre informazioni riguardanti la sequenza, come il nome dell'organismo di provenienza o eventuali annotazioni. Dopo la riga di descrizione inizia la sequenza vera e propria, che può essere presentata su una o più righe in base alla sua lunghezza ed è rappresentata utilizzando i codici definiti da IUPAC rispettivamente per gli acidi nucleici e per gli amminoacidi. All'interno della Tabella 4.3, sono elencati i codici supportati dal formato per gli acidi nucleici, nel nostro caso i più rilevanti.

| | |
|---|---|
| **A** –> adenosine | **S** –> G C (Strong) |
| **C** –> cytidine | **W** –> A T (weak) |
| **G** –> guanine | **B** –> G T C |
| **T** –> thymidine | **D** –> G A T |
| **U** –> uridine | **H** –> A C T |
| **R** –> G A (purine) | **V** –> G C A |
| **Y** –> T C (pyrimidine) | **N** –> A G C T (any) |
| **K** –> G T (keto) | **-** –> gap of indeterminate length |
| **M** –> A C (amino) | |

Tabella 4.3: Codici IUPAC per gli acidi nucleici supportati dal formato FASTA

La semplicità del formato ne ha permesso la diffusione, in quanto la sua natura testuale comporta una maggiore propensione all'analisi e alla manipolazione, aspetti fondamentali in un contesto composto da una grande quantità di dati da analizzare.

## 4.3 Metriche di valutazione

Per la valutazione individuiamo tre metriche principali: la precisione, l'accuratezza e l'F-score.
Per i calcoli di queste metriche di valutazione è importante specificare che per l'assegnazione delle etichette ai cluster si è deciso di scegliere la specie con la maggioranza di read, e di mantenere anche più cluster con la stessa etichetta, per prevenire i casi in cui ci dovessero essere un numero di cluster maggiore del numero di specie.

La precisione è una misura che indica quanti tra i casi forniti identificati siano effettivamente corretti. Più concretamente, nel nostro caso la precisione rappresenta l'abilità nel collocare read appartenenti ad una specie nel cluster corretto, ed è così calcolata:

$$Precision = \frac{TP}{TP+FP}$$

L'accuratezza o recupero come misura rappresenta il rapporto di istanze positive correttamente individuate dal software. Rapportata al contesto di questo lavoro, l'accuratezza può essere raffigurata come la capacità di inserire nello stesso cluster tutte le read di una stessa specie.

$$Recall = \frac{TP}{TP+FN}$$

La F-measure o F-score, è la media armonica di precisione e accuratezza, ed è un dato utile per tenere sotto controllo congiuntamente precisione ed accuratezza

$$F - measure = 2 * \frac{Precision*Recall}{Precision+Recall}$$

## 4.4 Analisi dei risultati

Procediamo ora con l'analisi dei risultati, soffermandoci separatamente sulle due fasi dell'esperimento al fine di acquisire una comprensione dettagliata del processo di elaborazione dei dati adoperato da AbundanceBin e MetaProb. Tale analisi ci permetterà successivamente di eseguire un confronto tra l'approccio sequenziale a più fasi e le performance ottenute tramite l'utilizzo dei singoli strumenti per il binning metagenomico.

### 4.4.1 Fase 1

Dalla Tabella 4.4 possiamo notare come AbundanceBin suddivida i dati del dataset in ingresso in tre diversi cluster. All'interno di questi ultimi è possibile individuare la presenza di read spaiate, ovvero coppie di paired-end reads dove una di esse è stata assegnata ad un cluster differente rispetto all'altra.

| Cluster | Read Totali | Read accoppiate | Read Spaiate | % read spaiate |
|---|---|---|---|---|
| Cluster 1 | 2 405 443 | 2 152 948 | 252 495 | 10.5 % |
| Cluster 2 | 786 613 | 453 664 | 332 949 | 42.33 % |
| Cluster 3 | 115 044 | 20 506 | 94 538 | 82.18 % |

Tabella 4.4: Composizione dei cluster generati da AbundanceBin

Considerando la distribuzione delle abbondanze delle diverse specie all'interno del dataset S7, l'aspettativa successivamente all'applicazione di AbundanceBin era quella di ottenere una suddivisione in due cluster: uno per le due specie con abbondanza 4 e un altro per le tre specie con abbondanza 1. Tuttavia, l'analisi dei risultati presentati nella Tabella 4.5 rivela che il numero di cluster generati è pari a tre e non due come previsto.

I tre cluster risultanti evidenziano un efficace raggruppamento delle specie con abbondanza più elevata, ossia le specie 4 e 5 (come illustrato nella Tabella 4.5), che sono prevalentemente collocate all'interno del Cluster 1. Le tre specie con abbondanza inferiore sono invece distribuite tra il Cluster 2 e il Cluster 3, quest'ultimo caratterizzato principalmente dalla presenza di read spaiate.

Nonostante il numero effettivo di cluster differisca dalle previsioni iniziali, guardando più in dettaglio la composizione dei cluster possiamo affermare che la suddivisione delle specie all'interno dei cluster risulta congruente con l'abbondanza attesa, dimostrando l'efficacia di AbundanceBin nell'identificare e raggruppare le specie in base all'abbondanza nel dataset S7.

| Cluster | Specie 1 | Specie 2 | Specie 3 | Specie 4 | Specie 5 |
|---|---|---|---|---|---|
| Cluster 1 | 10 000 | 18 723 | 8 097 | 1 065 139 | 1 303 484 |
| Cluster 2 | 225 827 | 108 303 | 203 762 | 109 648 | 139 073 |
| Cluster 3 | 48 029 | 23 340 | 40 507 | 1 389 | 1 779 |

Tabella 4.5: Numero di read per specie nei cluster generati da AbundanceBin

Le Tabelle 4.4 e 4.5 non forniscono informazioni sufficienti, in quanto non mettono in luce il numero di read appartenenti a ciascuna specie all'interno di ogni cluster che sono effettivamente accoppiate o spaiate.
Questa distinzione è di fondamentale importanza, poiché le read spaiate sono comunemente ritenute errori generati dal processo di analisi eseguito da AbundanceBin.

Per questo motivo è stata introdotta la Tabella 4.6 per presentare una rappresentazione più dettagliata e accurata dei dati. Questa Tabella evidenzia, per ciascuna specie

all'interno di ogni cluster, quante delle read associate siano effettivamente accoppiate o spaiate. La colonna "% Read", inoltre, permette una visione a colpo d'occhio della struttura del cluster e della percentuale con cui le specie appaiono al suo interno. Questo approccio mira a garantire una valutazione più precisa dei risultati ottenuti, consentendo una migliore analisi delle informazioni emerse.

| Specie | Read | % Read | Di cui paired | Di cui unpaired |
|---|---|---|---|---|
| Specie 1 | 10 000 | 0.42% | 30.04 % | 69.96 % |
| Specie 2 | 18 723 | 0.78% | 27.98 % | 72.02 % |
| Specie 3 | 8 097 | 0.34% | 30.04 % | 69.96 % |
| Specie 4 | 1 065 139 | 44.30% | 90.62 % | 9.38 % |
| Specie 5 | 1 303 484 | 54.16% | 90.30 % | 9.70 % |

(a) Cluster 1

| Specie | Read | % Read | Di cui paired | Di cui unpaired |
|---|---|---|---|---|
| Specie 1 | 225 827 | 28.70% | 80.51 % | 19.49 % |
| Specie 2 | 108 303 | 13.77% | 74.09 % | 25.91 % |
| Specie 3 | 203 762 | 25.90% | 81.82 % | 18.18 % |
| Specie 4 | 109 648 | 13.94% | 9.90 % | 90.10 % |
| Specie 5 | 139 073 | 17.69% | 10.10 % | 89.90 % |

(b) Cluster 2

| Specie | Read | % Read | Di cui paired | Di cui unpaired |
|---|---|---|---|---|
| Specie 1 | 48 029 | 41.75% | 18.46 % | 81.54 % |
| Specie 2 | 23 340 | 20.29% | 17.93 % | 82.07 % |
| Specie 3 | 40 507 | 35.21% | 18.40 % | 81.60 % |
| Specie 4 | 1 389 | 1.21% | 0 % | 100 % |
| Specie 5 | 1 779 | 1.55% | 0 % | 100 % |

(c) Cluster 3

Tabella 4.6: Numero di read per specie nei cluster generati da AbundanceBin

Dai dati presentati traspare come il metodo AbundanceBin dimostri prestazioni notevoli, includendo oltre il 90% delle read delle specie con abbondanza maggiore all'interno del Cluster 1. Il rimanente 10% di queste read, tuttavia, è principalmente composto da read spaiate che vengono distribuite tra i Cluster 2 e 3.
Per quanto riguarda le specie con abbondanza 1, la loro suddivisione è prevalentemente osservabile tra i Cluster 2 e 3, con alcune read in quantità minore erroneamente assegnate al Cluster 1.

Per ciascuna di queste tre specie, è degno di nota che la stragrande maggioranza delle read risulti essere assegnata al Cluster 2 come coppia di read paired-end, mentre

l'andamento opposto è osservato all'interno del Cluster numero 3.
In questo caso, la prevalenza delle read per ciascuna specie è costituita da read spaiate.

A questo punto, appurata la composizione dei cluster ottenuti dall'elaborazione del dataset da parte di AbundanceBin, è possibile determinare i valori di precisione, richiamo (recall) e F-measure.

Nella Tabella 4.7 sono riportati i risultati calcolati di queste misurazioni, che risultano sostanzialmente concordi con quelli presentati per AbundanceBin in [1], consentendo così di procedere con l'esperimento vista la possibilità di eseguire una comparazione diretta.

La precisione si presenta con un valore particolarmente modesto, poiché il metodo di clustering non ha l'obiettivo di eseguire un raggruppamento in base alle singole specie, ma piuttosto in cluster basati sulle diverse abbondanze. Di conseguenza, non ci aspettiamo di ottenere cluster contenenti singole specie.
Questo stesso principio contribuisce a un valore di recall notevolmente elevato, il quale risulta coerente con i risultati, dal momento che la maggior parte delle read di ciascuna specie è comunque inclusa all'interno di un unico cluster, nonostante la presenza di read spaiate o mal collocate.

| | |
|---|---|
| Precision | 0.477 |
| Recall | 0.879 |
| F-measure | 0.618 |

Tabella 4.7: Valori delle metriche di valutazione di AbundanceBin

### 4.4.2 Fase 2

In seguito alla fase iniziale, è stato necessario apportare modifiche alla struttura dei cluster al fine di renderli conformi al formato di input richiesto da MetaProb.
Alla luce dei numerosi casi di read spaiate che si verificavano all'interno del Cluster 3, è stata presa la decisione di implementare una procedura di riassegnazione e trasferimento di ciascuna read insieme alla sua controparte all'interno dei Cluster 1 e 2. Questa operazione mirava a ridurre il cluster di partenza alle sole read accoppiate.

In un secondo momento, sono state esplorate due alternative: nella prima opzione, sono state eliminate tutte le ulteriori read spaiate presenti all'interno dei Cluster 1 e 2. La seconda opzione, invece, ha coinvolto l'ulteriore assegnazione delle read rimaste spaiate nel cluster con il più alto numero in percentuale di read spaiate, ovvero il Cluster 2, a quello con una percentuale minore di read spaiate, nel nostro caso il Cluster 1, unico rimasto.

La necessità di adottare questa doppia strategia è stata motivata dall'output generato da AbundanceBin, il quale, a differenza di MetaProb, presenta alcune difficoltà nella

gestione delle paired-end reads, risultando in read spaiate le cui controparti sono assegnate a cluster diversi.
Al fine di definire una strategia efficace per mitigare queste problematiche, è stato deciso di sperimentare entrambe le soluzioni e confrontarle, allo scopo di ottenere una comprensione più precisa degli errori commessi da AbundanceBin e trovare la soluzione migliore al problema.

In entrambi i casi, si è optato per il riaccoppiamento delle read spaiate contenute all'interno del Cluster 3, in quanto all'interno di quest'ultimo oltre l'80% delle read risultava essere spaiato, indicando chiaramente un errore nell'assegnazione delle stesse.

Nel primo scenario, la decisione è stata di procedere prima con il riaccoppiamento delle read spaiate all'interno del Cluster 3, associandole alle rispettive controparti presenti nei Cluster 1 e 2. Successivamente, si è proceduto all'eliminazione delle read che rimanevano spaiate tra i Cluster 1 e 2.
Questa scelta è stata motivata dalla volontà di evitare il rischio di riassegnare erroneamente le read ad un cluster inappropriato, considerando che si presumeva che la maggioranza delle read spaiate fosse il risultato di errori commessi da AbundanceBin. Un ulteriore motivo di tale decisione risiedeva nel fatto che, anche riportando le read in un altro cluster, non si poteva garantire con certezza assoluta la corretta riallocazione delle read, specialmente considerando che in una situazione realistica in cui il numero e l'abbondanza delle specie non sono noti, non è possibile l'assegnazione precisa delle read in base alle etichette.

| Cluster | Specie | Coppie Read |
|---|---|---|
| Cluster 1.A | Specie 1 | 682 |
| | Specie 2 | 4 533 |
| | Specie 3 | 401 |
| | **Specie 4** | **447 044** |
| | Specie 5 | 16 281 |
| Cluster 1.B | Specie 1 | 1 871 |
| | Specie 2 | 369 |
| | Specie 3 | 1 648 |
| | Specie 4 | 34 194 |
| | **Specie 5** | **103 031** |
| Cluster 1.C | Specie 1 | 21 |
| | Specie 2 | 6 |
| | Specie 3 | 3 |
| | Specie 4 | 2 627 |
| | **Specie 5** | **470 805** |

Tabella 4.8: Elaborazione Cluster 1 con MetaProb

Dopo aver eliminato le read spaiate, è stato possibile preparare il file di input per MetaProb, cluster per cluster, per poi eseguirlo. I risultati di questo processo di computazione sono dettagliati nelle Tabelle 4.8 e 4.9, che evidenziano come MetaProb abbia

| Cluster | Specie | Coppie Read |
|---|---|---|
| Cluster 3.A | **Specie 1** | **3 281** |
| | Specie 2 | 303 |
| | Specie 3 | 564 |
| | Specie 4 | 0 |
| | Specie 5 | 0 |
| Cluster 3.B | Specie 1 | 708 |
| | Specie 2 | 519 |
| | **Specie 3** | **1 502** |
| | Specie 4 | 0 |
| | Specie 5 | 0 |
| Cluster 3.C | Specie 1 | 445 |
| | Specie 2 | 1 271 |
| | **Specie 3** | **1 660** |
| | Specie 4 | 0 |
| | Specie 5 | 0 |

Tabella 4.9: Elaborazione Cluster 3 con MetaProb

successivamente suddiviso sia il Cluster 1 sia il Cluster 3 in ulteriori 3 cluster ciascuno.

Il Cluster 3, una volta eseguito lo spostamento delle read spaiate, conteneva esclusivamente read provenienti dalle tre specie con abbondanza 1. Di conseguenza, ci si attendeva la creazione di tre cluster distinti e una suddivisione equa delle read tra queste specie: specie 1, specie 2 e specie 3. Tuttavia, i risultati presentano una situazione leggermente diversa.

Il primo cluster generato contiene approssimativamente l'80% di read appartenenti alla specie 1, il Cluster 3.B ospita oltre il 50% di read della specie 3, mentre il Cluster 3.C, che dovrebbe rappresentare la specie 2, è composto per circa il 49% da read della specie 3 e circa il 38% dalla specie 2.
Questo fenomeno potrebbe essere attribuito al fatto che il Cluster 3 generato da AbundanceBin, dopo essere stato purificato dalle read spaiate presenti inizialmente, conteneva solamente 10.253 coppie di read paired-end. Tale cifra rappresenta meno dell'1% del numero totale di read presenti inizialmente nel dataset, confermando il fatto che si tratta di un cluster "spurio" generato da AbundanceBin
La specie 2, inoltre, risulta quella con numero di read totali minori rispetto a tutte le altre all'interno del dataset.

Il Cluster 1 generato da AbundanceBin, come precedentemente evidenziato nella Tabella 4.6, era il destinatario delle read appartenenti alle specie 4 e 5, cioè le due specie con un'abbondanza pari a 4. In una situazione ideale, se questo cluster fosse stato il destinatario esclusivamente delle specie con abbondanza maggiore, ci saremmo aspettati la creazione di due ulteriori cluster da parte di MetaProb, uno per la specie 4

e uno per la specie 5.
Tuttavia, tenendo conto che AbundanceBin può introdurre errori e read erroneamente assegnate all'interno del cluster, è necessario riconsiderare le nostre aspettative.
Più realisticamente, considerando i dati in nostro possesso prevediamo di ottenere 3 cluster, uno per ciascuna delle due specie con abbondanza maggiore e uno contenente tutte le read delle rimanenti specie, le quali hanno abbondanze simili tra loro ma diverse dalle due specie dominanti.

I risultati ottenuti confermano che MetaProb è in grado di individuare effettivamente 3 cluster distinti, come previsto: il Cluster 1.A appena creato contiene read delle quali oltre il 95% appartenengono alla specie 4, mentre il Cluster 1.C ospita read composte al 99% da sequenze appartenenti alla specie 5.

Il Cluster 1.B richiede un'analisi più approfondita. Questo cluster contiene per il 73% read appartenenti alla specie 5 e per il 23% read appartenenti alla specie 4.
Nonostante rappresenti solo il 10% delle read totali del cluster fornito in input, va notato che il numero di read delle specie con un'abbondanza pari a 1 è notevolmente inferiore, dato che queste read sono principalmente il risultato di errori di categorizzazione introdotti da AbundanceBin.
Inoltre, è importante sottolineare che MetaProb è naturalmente più adatto a lavorare con cluster che contengono specie con abbondanze simili. Questo è evidente nella corretta classificazione delle specie 4 e 5. Tuttavia, MetaProb incontra delle difficoltà quando la variazione nell'abbondanza delle specie è più marcata, come nel caso del Cluster 2. Questo aspetto motiva l'approccio adottato in questo esperimento, che mirava a migliorare le prestazioni di MetaProb rimuovendo uno dei suoi punti deboli, limite simile a quello riscontrato in altri software basati sulla composizione del DNA.

L'analisi del Cluster 2 di AbundanceBin ed elaborato da MetaProb richiede un'attenta disamina. Dopo il riaccoppiamento delle read del Cluster 3 con le loro controparti e l'eliminazione delle rimanenti read spaiate, MetaProb ha suddiviso il cluster iniziale in 10 cluster distinti. Questo numero appare notevolmente elevato, e per comprenderlo appieno, è essenziale esaminare la composizione originaria del Cluster 2 prima dell'analisi condotta dal secondo tool di binning.

Come evidenziato nella Tabella 4.6, il Cluster 2 era costituito per il 95% da read appartenenti alle specie 1, 2 e 3, mentre il restante 5% era attribuito alle specie 4 e 5. L'aspettativa era quindi di ottenere un numero di cluster ragionevolmente compreso tra 3 e 6, tenendo conto che all'interno del cluster erano presenti tutte le specie e che il tool poteva tendere a sovrastimare il numero di cluster in cui suddividere il file di input.

L'analisi della composizione dei 10 cluster creati da MetaProb rivela come le poche migliaia di read appartenenti alle specie 4 e 5 siano sparse tra i vari cluster, con la specie 4 che non risulta mai essere predominante in nessuno dei 10 gruppi.

| Cluster | Specie | Coppie Read | Cluster | Specie | Coppie Read |
|---|---|---|---|---|---|
| Cluster 2.A | Specie 1 | 4 698 | Cluster 2.F | Specie 1 | 554 |
| | **Specie 2** | **20 952** | | **Specie 2** | **30 002** |
| | Specie 3 | 12 111 | | Specie 3 | 0 |
| | Specie 4 | 3 449 | | Specie 4 | 1 |
| | Specie 5 | 1 853 | | Specie 5 | 0 |
| Cluster 2.B | Specie 1 | 6 508 | Cluster 2.G | Specie 1 | 133 |
| | Specie 2 | 64 | | Specie 2 | 585 |
| | **Specie 3** | **6 628** | | Specie 3 | 5 |
| | Specie 4 | 757 | | Specie 4 | 441 |
| | Specie 5 | 72 | | **Specie 5** | **2 087** |
| Cluster 2.C | Specie 1 | 2 381 | Cluster 2.H | Specie 1 | 1 608 |
| | Specie 2 | 258 | | **Specie 2** | **3 111** |
| | **Specie 3** | **91 188** | | Specie 3 | 299 |
| | Specie 4 | 53 | | Specie 4 | 345 |
| | Specie 5 | 1 | | Specie 5 | 3 039 |
| Cluster 2.D | **Specie 1** | **3 178** | Cluster 2.I | **Specie 1** | **98 564** |
| | Specie 2 | 1 246 | | Specie 2 | 0 |
| | Specie 3 | 2 235 | | Specie 3 | 1 079 |
| | Specie 4 | 347 | | Specie 4 | 0 |
| | Specie 5 | 86 | | Specie 5 | 0 |
| Cluster 2.E | Specie 1 | 268 | Cluster 2.J | **Specie 1** | **11 104** |
| | Specie 2 | 766 | | Specie 2 | 2 |
| | **Specie 3** | **1 257** | | Specie 3 | 774 |
| | Specie 4 | 114 | | Specie 4 | 60 |
| | Specie 5 | 24 | | Specie 5 | 41 |

Tabella 4.10: Suddivisione di MetaProb del cluster 2

Di questi 10 cluster, 4 contengono un numero di coppie di read superiore a 30.000, con il Cluster 2.I che raggiunge quasi 100.000 coppie di read. Gli altri 6 cluster ospitano un numero inferiore di read, con quantità che variano dalle 2.400 nel Cluster 2.E alle 14.000 coppie nel Cluster 2.B.
Le specie 1 e 3 sono concentrate principalmente nei Cluster 2.I e 2.C, che sono i più numerosi, e in entrambi i casi, rappresentano circa il 98% della composizione totale di questi cluster. Ciò suggerisce che MetaProb sia stato in grado di riconoscere correttamente la presenza di queste specie e di categorizzarle in modo accurato. Per entrambe le specie, infatti, il 75% delle read ad esse associate nell'input è concentrato all'interno di questi due cluster.
Per quanto riguarda la specie 2, circa il 90% delle sue read è suddiviso tra i Cluster 2.A e 2.F, nei quali essa costituisce la specie predominante.
Questo ci mostra come, in realtà, i restanti cluster contengano un numero limitato di read, e che, di conseguenza, le read appartenenti alla stessa specie tendono comunque ad essere posizionate per la maggior parte nello stesso cluster.

Esaminiamo ora i risultati derivanti dall'impiego del secondo approccio per rimuo-

vere le read spaiate dai cluster di input. Come precedentemente menzionato, in questo caso le read sono state prima di tutto riaccoppiate dal Cluster 3 ai rispettivi Cluster 1 e 2. Successivamente, si è proseguito con il riaccoppiamento delle read del cluster con la percentuale più elevata di read spaiate, ovvero il Cluster 2, con le read spaiate rimanenti nel Cluster 1. Di conseguenza, i Cluster 2 e 3 presentano con entrambi gli approcci lo stesso contenuto, in quanto le read spaiate precedentemente rimosse dal Cluster 2 sono trasferite comunque tutte all'interno del Cluster 1, l'unico rimasto. Poiché i cluster in ingresso a MetaProb sono identici, anche quelli in uscita lo saranno. Pertanto, ci concentreremo ora sull'analisi dei risultati ottenuti per il Cluster 1, tenendo conto delle considerazioni effettuate in precedenza per gli altri due.

| Cluster | Specie | Coppie Read |
|---|---|---|
| Cluster 1.A | Specie 1 | 4 370 |
| | Specie 2 | 15 285 |
| | Specie 3 | 3 211 |
| | **Specie 4** | **541 144** |
| | Specie 5 | 22 428 |
| Cluster 1.B | Specie 1 | 4 077 |
| | Specie 2 | 812 |
| | Specie 3 | 3 649 |
| | Specie 4 | 37 410 |
| | **Specie 5** | **110 605** |
| Cluster 1.C | Specie 1 | 51 |
| | Specie 2 | 7 |
| | Specie 3 | 21 |
| | Specie 4 | 3 967 |
| | **Specie 5** | **581 932** |

Tabella 4.11: Cluster 1 dopo approccio 2, riaccoppiamento reads

Il Cluster 1 contiene ora circa 250.000 read in più rispetto al primo caso. È pertanto cruciale comprendere se il movimento delle read ha influenzato positivamente o negativamente le prestazioni dell'algoritmo di suddivisione dei cluster eseguito da MetaProb.

Come nel metodo precedente, ci aspettiamo la presenza di 2 o 3 cluster, uno per ciascuna delle specie con un'abbondanza di 4 e forse uno per le poche read relative alle specie 1, 2 e 3.

Il risultato ottenuto è di 3 cluster, in linea con l'elaborazione precedente e distribuiti come indicato nella Tabella 4.11. Come è possibile osservare, le specie predominanti individuate in ciascun cluster sono le medesime dell'elaborazione precedente. Di conseguenza, le considerazioni effettuate in precedenza relativamente alle percentuali di distribuzione delle read delle specie all'interno dei cluster rimangono valide.

| Approccio | Precision | Recall | F-measure |
|---|---|---|---|
| Approccio 1 | **0.920** | 0.688 | 0.787 |
| Approccio 2 | 0.912 | **0.812** | **0.859** |

Tabella 4.12: Valutazione dei risultati finali

I risultati differiscono da quelli precedenti, come mostrato nella Tabella 4.12, con un sensibile aumento della recall e di conseguenza un non trascurabile incremento di F-measure. La precisione risulta invece leggermente inferiore.

La similitudine osservata tra alcuni dei risultati può essere attribuita principalmente al fatto che l'unico cluster tra i tre in ingresso a MetaProb che è stato soggetto a variazioni è il Cluster numero 1. La nostra scelta di spostare le read è stata motivata dall'ipotesi che i cluster con una elevata percentuale di read spaiate contenessero anche un significativo numero di read erroneamente assegnate. L'implementazione di entrambe le strategie rappresentava una ricerca attenta di confronto e di una soluzione ottimale, poiché molte altre strategie erano inapplicabili a causa del nostro obiettivo di adottare approcci riproducibili in contesti reali.

Il nostro approccio è più realistico e differisce dalla soluzione proposta in [10], dove viene dimostrato come l'uso combinato di AbundanceBin e MetaCluster, un tool di binning basato sulla composizione del DNA simile a MetaProb, migliori le prestazioni. Tuttavia, in tale test, il numero corretto di bin è fornito come input al secondo strumento, il che rende tale approccio inapplicabile in situazioni non controllate e con specie sconosciute, come spesso si verifica nella pratica. Anche se MetaProb consente di specificare il numero di bin totali in input, abbiamo scelto di seguire la strada più realistica e pragmatica. Verificando l'ipotesi attraverso il nostro approccio, potremmo teoricamente migliorare il binning e, di conseguenza, ottenere una migliore stima e suddivisione delle specie all'interno dei campioni raccolti, il che è più adatto alle situazioni reali e maggiormente auspicabile.

## 4.5 Confronto con metodi di binning ad una fase

| | MetaProb | AbundanceBin | AbBin + MP 1 | AbBin + MP 2 |
|---|---|---|---|---|
| Precision | 0.818 | 0.477 | **0.920** | 0.912 |
| Recall | 0.745 | **0.879** | 0.688 | 0.812 |
| F-measure | 0.780 | 0.618 | 0.787 | **0.859** |

Tabella 4.13: Confronto dei risultati

Nella Tabella 4.13 viene presentato un confronto tra i dati di precisione, recall e F-measure raccolti per lo stesso dataset utilizzato in questo esperimento, sia quando è stato utilizzato un singolo strumento per il binning metagenomico, sia nel caso in cui

ne siano stati utilizzati due in sequenza, come nel nostro caso. Per i valori di MetaProb usato singolarmente, sono stati usati quelli presentati in [1], calcolati usando lo stesso dataset S7 e senza fornire il numero di bin in ingresso.

Come è possibile notare i risultati hanno mostrato miglioramenti come previsto, con entrambi gli approcci. I casi caratterizzati dall'utilizzo di due strumenti per il binning hanno presentato una precisione superiore rispetto a quelli con l'uso di singoli tool, e il secondo approccio si è contraddistinto per la presenza di un valore di recall alto, portando ad avere la F-measure globalmente migliore.

E' comunque necessario considerare che nonostante i valori calcolati siano alti, alla fine dell'esperimento il numero di cluster ottenuti è 16, molto maggiore rispetto alle 5 specie reali. Ciononostante, i risultati permettono di fare delle considerazioni in merito al rapporto tra il numero elevato di cluster ottenuti e i valori di precisione e recall. Le specie risultano infatti comunque distribuite per la stra grande maggioranza all'interno di pochi cluster specifici, portando ad avere una stima magari non attendibile del numero di specie presenti, ma ad una accettabile categorizzazione e suddivisione delle specie.

E' inoltre importante ricordare che a causa degli output contenenti read spaiate forniti da AbundanceBin, vi è l'introduzione di errori e, soprattutto, di incertezza sulla strategia da adottare per ricondurre le read ad un formato accoppiato all'interno dello stesso cluster, rendendole compatibili per l'utilizzo successivo in MetaProb.

La scelta della strategia può avere un impatto significativo, influenzando l'input del secondo strumento, che potrebbe trovarsi con una quantità di read errate introdotte manualmente durante la fase di riaccoppiamento delle read.
Il rischio è anche quello di avere una quantità di errori in ingresso che si propagano ed amplificano in uscita, rendendo la selezione della strategia di gestione delle read spaiate un aspetto cruciale. Inoltre, esiste la possibilità che una scelta metodologica per il trattamento delle read spaiate possa funzionare con un insieme di dati in ingresso e non essere altrettanto efficace con un altro, rendendo più complessa la ricerca di una soluzione.

# 5

# Conclusioni e lavori futuri

Questo studio ha analizzato l'utilizzo in sequenza di due tool per il binning metagenomico su un dataset composto da short paired-end reads, con lo scopo di verificare se l'utilizzo congiunto di due software, uno basato sull'abbondanza e uno basato sulla composizione del DNA, potesse portare ad un miglioramento delle prestazioni rispetto all'uso dei software singoli.

I risultati di questo studio hanno rivelato valori elevati per tutte le metriche di valutazione dei risultati, suggerendo pertanto la validità dell'ipotesi iniziale in particolare con l'approccio di riassegnazione iterativa delle read spaiate. Questi risultati appaiono altamente promettenti e possono offrire importanti indicazioni per lo sviluppo di nuovi e più prestanti strumenti per il binning metagenomico non supervisionato.
Inoltre, corroborano l'intuizione proposta dagli autori di AbundanceBin [10], portando l'idea di complementarietà di due tool per il binning in un contesto realistico, ovvero non fornendo mai il numero di bin in ingresso.
Durante la fase di ricerca sono inoltre emersi degli aspetti importanti relativi alle condizioni di ottenimento dei risultati, tra cui l'importanza della scelta della strategia per la riassegnazione delle read spaiate ottenute dai cluster dopo l'esecuzione del software AbundanceBin.

Per verificare ulteriormente l'ipotesi avanzata in questa tesi, in futuro si auspica che l'applicazione dell'approccio scelto in questa tesi possa essere estesa ad un maggior numero di dataset, in modo da confermarne l'attendibilità.
In aggiunta, l'esplorazione di strumenti per il binning metagenomico alternativi e più recenti potrebbe costituire un ulteriore passo avanti nella ricerca. L'adozione di tali strumenti potenzialmente più avanzati potrebbe contribuire a un ulteriore miglioramento delle prestazioni e rappresentare un ulteriore sviluppo significativo nel campo.

# Bibliografia


[1] Samuele Girotto, Cinzia Pizzi e Matteo Comin. «MetaProb: accurate metagenomic reads binning based on probabilistic sequence signatures». In: *Bioinformatics* 32.17 (ago. 2016), pp. i567–i575. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw466. URL: https://doi.org/10.1093/bioinformatics/btw466.

[2] Daniel H Huson et al. «MEGAN analysis of metagenomic data». In: *Genome research* 17.3 (2007), pp. 377–386.

[3] Miten Jain et al. «The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community». In: *Genome biology* 17 (2016), pp. 1–11.

[4] T. Soni Madhulatha. *An Overview on Clustering Methods*. 2012. arXiv: 1205.1117 [cs.DS].

[5] Sharmila S. Mande, Monzoorul Haque Mohammed e Tarini Shankar Ghosh. «Classification of metagenomic sequences: methods and challenges». In: *Briefings in Bioinformatics* 13.6 (set. 2012), pp. 669–681. ISSN: 1467-5463. DOI: 10.1093/bib/bbs054. eprint: https://academic.oup.com/bib/article-pdf/13/6/669/482874/bbs054.pdf. URL: https://doi.org/10.1093/bib/bbs054.

[6] Anthony Rhoads e Kin Fai Au. «PacBio sequencing and its applications». In: *Genomics, proteomics & bioinformatics* 13.5 (2015), pp. 278–289.

[7] Andrew B Shreiner, John Y Kao e Vincent B Young. «The gut microbiome in health and in disease». In: *Current opinion in gastroenterology* 31.1 (2015), p. 69.

[8] Le Van Vinh et al. «A two-phase binning algorithm using l-mer frequency on groups of non-overlapping reads». In: *Algorithms for Molecular Biology* 10 (2015), pp. 1–12.

[9] Wood e Salzberg. «Kraken: ultrafast metagenomic sequence classification using exact alignments.» In: *Genome Biology 2014 15:R46* (2014). DOI: doi:10.1186/gb-2014-15-3-r46.

[10] Yu-Wei Wu e Yuzhen Ye. «A Novel Abundance-Based Algorithm for Binning Metagenomic Sequences Using l-tuples». In: *Journal of Computational Biology* 18.3 (2011). PMID: 21385052, pp. 523–534. DOI: 10.1089/cmb.2010.0245. eprint: https://doi.org/10.1089/cmb.2010.0245. URL: https://doi.org/10.1089/cmb.2010.0245.

[11] Bin Yang et al. «MetaCluster: unsupervised binning of environmental genomic fragments and taxonomic annotation». In: *Proceedings of the first ACM international conference on bioinformatics and computational biology*. 2010, pp. 170–179.