UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA INFORMATICA

L'evoluzione dei database di riferimenti e il suo impatto sulla classificazione tassonomica: analisi critica di Kraken 2

Relatore: Prof. Comin Matteo

Laureando: Ferro Riccardo

ANNO ACCADEMICO 2023-2024

Data di laurea 13/03/2025

Abstract

L'aumento delle dimensioni dei database genomici di riferimento rappresenta una sfida cruciale per gli strumenti di classificazione tassonomica, come il software Kraken 2. Esso utilizza k-mer (brevi sottosequenze di DNA) per associare i dati analizzati al loro antenato comune nell'albero tassonomico.
La scelta del database è determinante per l'accuratezza e la sensibilità delle classificazioni, poiché un database più completo aumenta la probabilità di contenere i k-mer necessari. Sono disponibili diversi archivi digitali di riferimenti, tra cui Standard, PlusPF e PlusPFP, ciascuno con vantaggi specifici in base al focus dello studio. Tuttavia, l'incremento delle dimensioni di questi database comporta maggiore imprecisione per il software di classificazione, oltre ad un aumento dell'uso di memoria e del tempo di elaborazione. Kraken 2 implementa ottimizzazioni per mitigare questi problemi, ma i requisiti computazionali rimangono significativi.
Questo studio esplora l'impatto delle dimensioni e della scelta del database sulle prestazioni di Kraken 2, fornendo raccomandazioni per ottimizzare la classificazione tassonomica in diversi contesti di ricerca metagenomica.

Indice

1 Introduzione **1**

2 Tecnologie e Dati: Il Ruolo di Kraken 2 e dei Database nell'Analisi **3**
2.1 Il software: Kraken 2 3
2.2 Il database: RefSeq 6

3 Guida alla lettura del database RefSeq **9**
3.1 Parametri di configurazione 9
3.2 Analisi Dettagliata delle Colonne del Database: Struttura e Funzionalità 10

4 Evoluzione del database RefSeq nel corso degli anni **11**

5 Estrapolazione ed analisi di dati **15**
5.1 Dati acquisiti dalla seconda colonna - Numero minimizers 16
5.1.1 Database Standard 16
5.1.2 Database Plus PFP 18
5.1.3 Analisi in merito al numero di minimizers 20
5.2 Dati acquisiti dalla prima colonna - valori percentuali 21
5.2.1 Database Standard 21
5.2.2 Database Plus PFP 23
5.2.3 Analisi in merito ai valori percentuali 25

6 Conclusioni **27**

Bibliografia **31**

Capitolo 1

Introduzione

La metagenomica è la scienza che si occupa dello studio dei genomi collettivi di comunità microbiche direttamente prelevate dall'ambiente. Le basi per l'interpretazione dei dati metagenomici sono state fornite dalla classificazione tassonomica. Questo processo consiste nell'assegnare un'etichetta (può trattarsi di regno, phylum, classe, ordine, famiglia, genere e specie) a ciascun read in un set di dati metagenomici. I software di classificazione giocano un ruolo cruciale nell'analisi metagenomica, permettendo di assegnare le sequenze di DNA ottenute dal sequenziamento a specifici taxa (gruppi di organismi classificati in base alle loro relazioni evolutive) e facilitando l'identificazione e la caratterizzazione delle comunità microbiche. All'interno della presente tesi di laurea viene proposta un'analisi dell'efficacia della classificazione tassonomica confrontando due dei principali database su cui si appoggia Kraken 2 e le loro rispettive versioni negli anni: 2020, 2022, 2024. Lo studio sottoporrà ad un approfondito esame il software Kraken 2 e valuterà la variazione di precisione nella classificazione metagenomica in relazione alla crescita dei database di riferimenti. L'analisi traccia l'evoluzione delle sequenze univoche delle seguenti specie e dei loro generi:

- Escherichia coli
- Salmonella enterica
- Pseudomonas fluorescens
- Serratia marcescens
- Proteus mirabilis

in relazione ai valori ad esse associati nelle diverse versioni dei database.

I software di classificazione tassonomica metagenomica si basano principalmente su due modalità di funzionamento. Il primo tipo di classificazione viene svolto facendo riferimento ai k-mer. Gli applicativi che fanno uso di processi di questo genere, come Kraken, Kraken 2, Clark, Centrifuge e SKraken, decompongono le sequenze di DNA in brevi sottosequenze di lunghezza fissa, chiamate k-mer. Le quali vengono utilizzate per interrogare un database di riferimenti, che associa ogni k-mer all'antenato comune più basso (LCA) nell'albero tassonomico di tutti i genomi che contengono quella specifica sequenza.
L'alternativa vede i metodi basati sull'allineamento. Gli applicativi che si appoggiano su sistemi di questo tipo (BLAST, MegaBlast e Megan) confrontano le sequenze di DNA metagenomiche con un database di riferimenti di sequenze note. L'allineamento cerca di trovare le migliori corrispondenze tra le concatenazioni, identificando così l'origine tassonomica della sequenza metagenomica. Procedure del genere possono essere computazionalmente costose e lente, soprattutto dato l'aumento delle dimensioni dei database di riferimenti.

Capitolo 2

Tecnologie e Dati: Il Ruolo di Kraken 2 e dei Database nell'Analisi

2.1 Il software: Kraken 2

Il software di classificazione tassonomica che verrà preso come oggetto dell'analisi è Kraken 2, seconda versione della famiglia di programmi Kraken. Questo nuovo modello utilizza una tabella hash compatta e probabilistica per mappare i minimizer (sottostringhe dei k-mer) ai LCA, riducendo in modo significativo l'utilizzo di memoria rispetto alla versione precedente. Inoltre, implementa una tecnica di seed spaziati per migliorare l'accuratezza dei risultati. [1].
Il processo di classificazione del presente applicativo inizia con l'analisi delle sequenze di input mediante l'estrazione di **k-mer (a)**. Kraken 2 utilizza il concetto di ℓ-bp minimizer, un sottoinsieme selezionato in base a un criterio di minimizzazione che riduce la complessità computazionale (nel caso specifico $\ell = 31$), e semplifica il confronto con le sequenze di riferimento.

Una volta ottenuti i minimizer, questi vengono processati tramite una serie di trasformazioni che li convertono in una rappresentazione compatta. Per farlo, Kraken 2 applica una **maschera a seme spaziato (b)**, in cui solo determinate posizioni della sequenza (ad esempio s = 7) vengono utilizzate per generare il valore finale, aumentando la sensibilità e riducendo il rumore dovuto a variazioni locali. La sequenza è successivamente elaborata da una **funzione di hash**, che produce un valore unico e compatto. Questo hash viene ulteriormente ridotto per permettere l'accesso rapido a una tabella hash compatta. Nel caso di collisioni, ovvero quando due valori hash conducono alla stessa posizione, viene utilizzato un approccio di linear probing, che ricerca la prossima posizione disponibile.

Infine, ogni record nella tabella hash è rappresentato in un formato binario di 32 bit, suddiviso in due parti principali **(c)**. La prima, composta da 15 bit, immagazzina il codice hash compatto, che identifica in maniera univoca il k-mer. La seconda, che occupa 17 bit, memorizza l'informazione relativa al Lowest Common Ancestor (LCA), ovvero il livello tassonomico più basso comune ai taxa associati a quel k-mer. Questa struttura permette di assegnare rapidamente le sequenze analizzate al livello tassonomico appropriato, sia esso dominio, phylum, genere o specie.

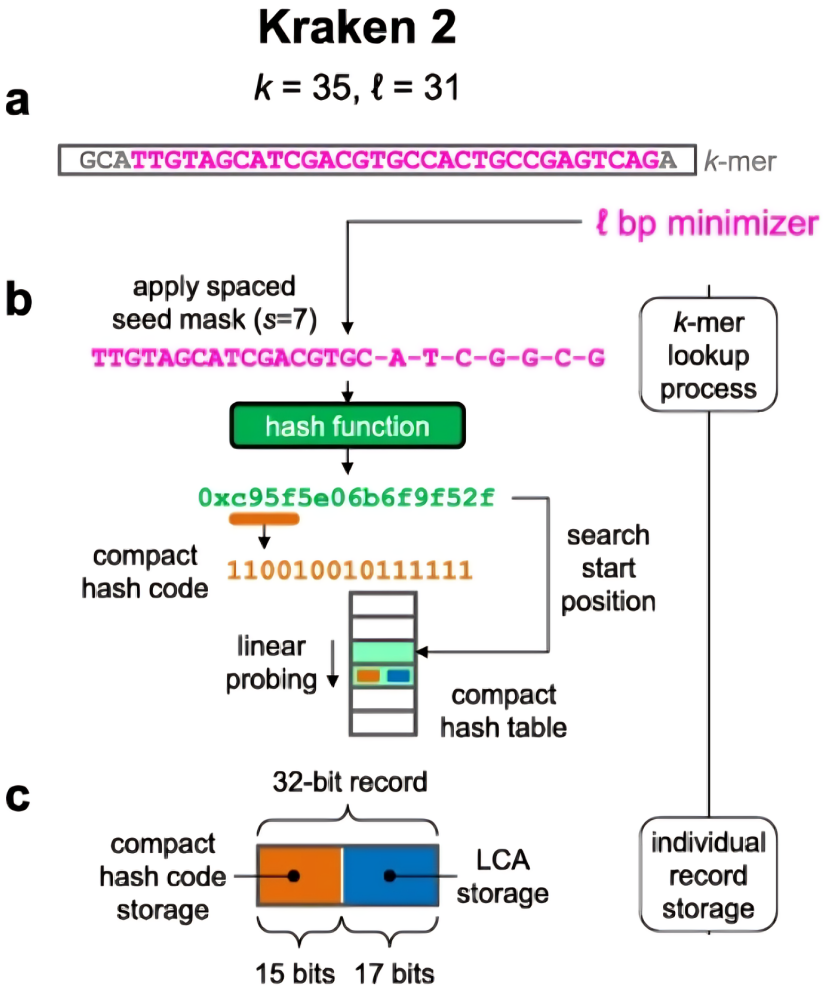

Figura 2.1: Funzionamento Kraken 2

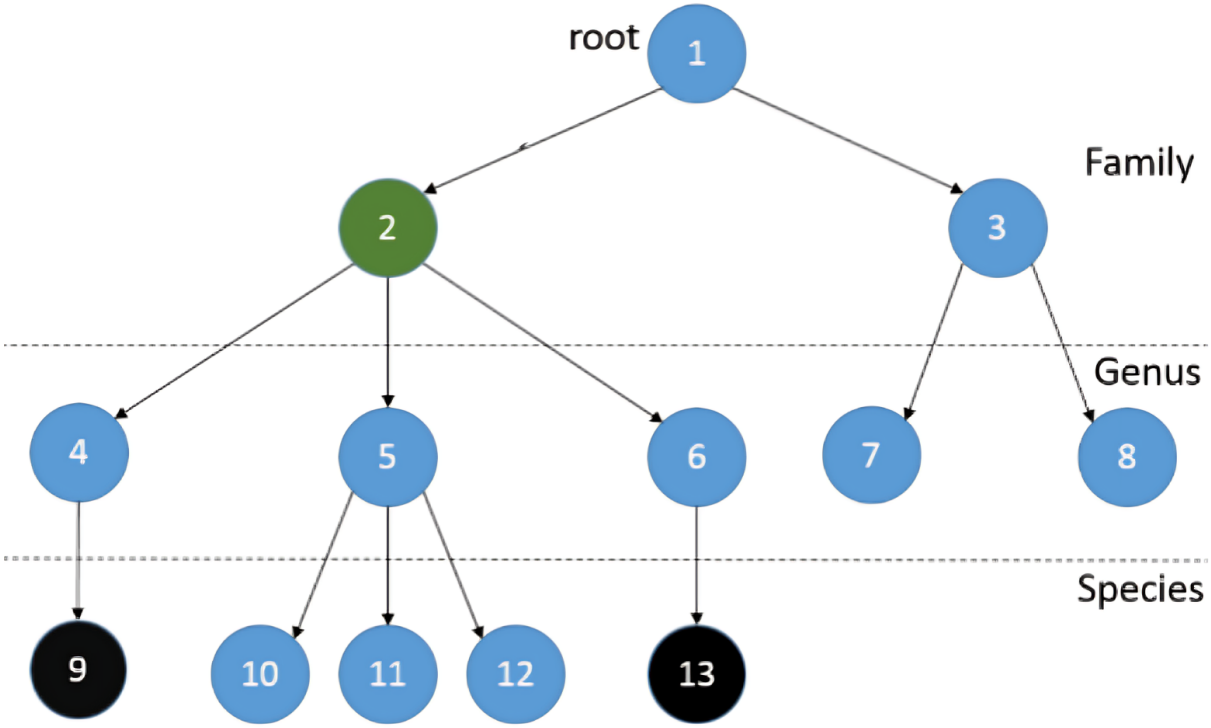

Figura 2.2: Albero tassonomico

2.2 Il database: RefSeq

Il database di riferimenti protagonista di questa trattazione è RefSeq (Reference Sequence) del ”National Center for Biotechnology Information” (NCBI), che è una raccolta annotata e curata di sequenze nucleotidiche (DNA, RNA) e dei loro prodotti proteici. La sua prima pubblicazione è avvenuta nel giugno 2003. Questo sistema di gestione dei dati si concentra nel fornire una sequenza di riferimento unica per ogni molecola. Ultime, ma non meno importanti, sono la cura ed il mantenimento manuale con cui vengono trattate le sequenze del database RefSeq. [2].
Nel corso dell'analisi svolta in questa trattazione verrano esaminate tre versioni di due dei database derivanti da RefSeq, sviluppate specificatamente per il software Kraken2. La scelta è ricaduta su ”Standard” e ”PFP Plus” per tenere traccia dell'evoluzione di quest'ultimi e della classificazione tassonomica a loro associata. Il primo rappresenta il miglior connubio tra accuratezza e velocità nell'insieme dei database a disposizione, avendo una minore esigenza di risorse computazionali. Il secondo include il set più ampio disponibile, potendo così offrire le classificazioni di miglior qualità.

Collection	Contains	Date	Archive size (GB)	Index size (GB)
Standard	Refeq archaea, bacteria, viral, plasmid, human[1], UniVec_Core	9/4/2024	62	80
PlusPFP	Standard plus Refeq protozoa, fungi & plant	9/4/2024	138	188
Standard	archaea, bacteria, viral, plasmid, human, UniVec_Core	9/26/2022	46	60
PlusPFP	Standard plus protozoa, fungi & plant	9/8/2022	99	129
Standard	archaea, bacteria, viral, plasmid, human[1], UniVec_Core	9/19/2020	36.0	47.0
PlusPFP	Standard plus protozoa, fungi & plant	9/19/2020	66.5	90.0

Figura 2.3: Database di riferimenti presi in analisi

Capitolo 3

Guida alla lettura del database RefSeq

Questo capitolo si propone di esplorare la struttura delle basi di dati su cui si appoggia Kraken 2 durante la sua esecuzione. La configurazione tabellare di ”Standard” e ”PFP Plus” è identica. Tra questi due database cambia ”solamente” il contenuto - figura 2.3.

3.1 Parametri di configurazione

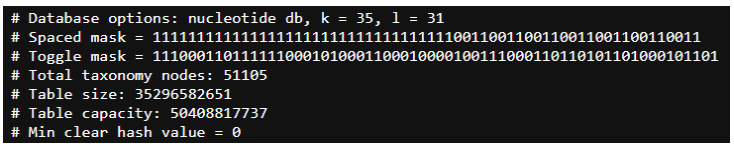

Figura 3.1: Parametri per la configurazione di Kraken 2

Le voci esposte in questa sezione rispecchiano le prime righe del database, fondamentali per configurare correttamente il software e garantire che l’analisi delle sequenze sia il più accurata ed efficiente possibile. Facendo riferimento all’immagine 3.1 la prima riga è **Database options**, questi valori specificano le opzioni del database: il tipo di sequenze (nucleotidi) e la lunghezza dei k-mer e l-mer utilizzati per l’analisi. La riga successiva contiene la **Spaced mask**. Questa sequenza di bit definisce quali posizioni nei k-mer devono essere considerate durante l’analisi. Di seguito è riportata un’altra sequenza di bit, la **Toggle mask**. Che viene utilizzata per alternare tra diverse configurazioni di analisi. Successivamente viene esposto il valore che indica il numero di classificazioni biologiche presenti nel database, ovvero **Total taxonomy nodes**. All’interno delle righe successive si trovano **Table size** e **Table capacity**, indici che mostrano la dimensione attuale della tabella e la sua capacità massima. L’ultimo parametro è **Min clear hash value**, il quale rappresenta il minimo valore di hash che può essere considerato valido. In questo modo è possibile filtrare i k-mer (scartando, ad esempio, quelli con un valore troppo basso).

3.2 Analisi Dettagliata delle Colonne del Database: Struttura e Funzionalità

La struttura del database è composta da diverse colonne, ognuna delle quali svolge un ruolo cruciale per garantire una proficua e precisa analisi dei dati. Viene mostrato un esempio di riga per rendere più comprensibile la spiegazione - figura 3.2

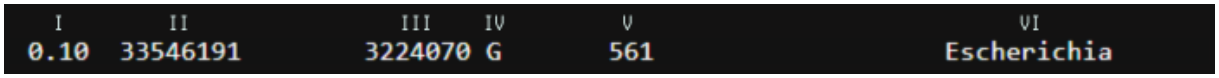

Figura 3.2: Riga del database che rappresenta il genere Escherichia, presa come esempio

La prima colonna identifica la **percentuale di letture** I calcolata in rapporto al numero totale di letture contenute nel dataset analizzato. Indica la proporzione di letture totali assegnate ad un determinato taxon. Per esempio, se un taxon a livello di famiglia ha il 5% (0,05) delle letture, significa che il 5% delle sequenze all'interno del dataset sono state identificate come appartenenti a quella famiglia.
La colonna successiva rappresenta il conteggio totale delle sequenze di DNA assegnate ad un certo taxon. Il valore, chiamato **numero di minimizers** II, fornisce un'indicazione della quantità complessiva dei frammenti di DNA associati a quel gruppo tassonomico. Di seguito, il **numero di minimizers unici** III indica il conteggio delle sequenze di DNA assegnate esclusivamente ad un taxon, senza sovrapposizioni con altri taxa. Evidenziando così la specificità di suddette sequenze per quella particolare unità tassonomica.
Il valore che segue costituisce un'informazione fondamentale per comprendere il livello gerarchico di appartenenza: il **livello tassonomico** IV. Che specifica la posizione del taxon nella gerarchia tassonomica, come regno, phylum, classe, ordine, famiglia, genere o specie.
La penultima colonna del database rappresenta l'**ID tassonomico** V. È l'identificatore numerico unico assegnato a ciascun taxon. Infine, come ultimo valore troviamo il **nome tassonomico** VI. Il quale mostra il nome scientifico del taxon (ad esempio, Escherichia coli) permettendo di risalire al nome a cui è associato l'ID tassonomico, rendendo più leggibili i risultati.

Capitolo 4

Evoluzione del database RefSeq nel corso degli anni

Come discusso in precedenza, il database di riferimenti in esame è nato a metà del 2003. Grazie all'avanzamento delle tecnologie di sequenziamento, che hanno portato a un aumento esponenziale della quantità di dati disponibili, questo è cresciuto in modo significativo sia in termini di dimensioni che di numero di specie incluse. Il database RefSeq è dimensionalmente raddoppiato (gigabasi, Gbp) in media ogni 1,5 anni, con il numero di 31-mer unici in crescita a un ritmo simile. Il numero di specie osservate in RefSeq è raddoppiato all'incirca ogni 3 anni, mentre il numero di generi osservati non è raddoppiato in quasi 6 anni. Questo divario nella crescita tra specie e generi ha portato a un costante aumento del rapporto specie-generi nel tempo, passando da meno di due specie per genere (versione 1) a quasi otto specie per genere (versione 89). Si osserva anche una generale, sebbene fluttuante, diminuzione del rapporto ceppi-specie, passando da otto (versione 1) a circa tre ceppi per specie (versione 89).[3]. Nei grafici qui sotto è possibile notare grandi variazioni intorno all'anno 2008, specialmente per quanto riguarda il grafico **A** figura 4.1. Da luglio 2007 a luglio dell'anno successivo il numero di specie incluse nella versione di RefSeq è aumentato del 20% e il numero totale di record è aumentato del 53%, con una crescita del 45% nella collezione di proteine. Una crescita del genere è dovuta alla cura da parte del personale del NCBI, ma soprattutto ad una enorme espansione del numero di genomi supportati dalle collaborazioni. [4]

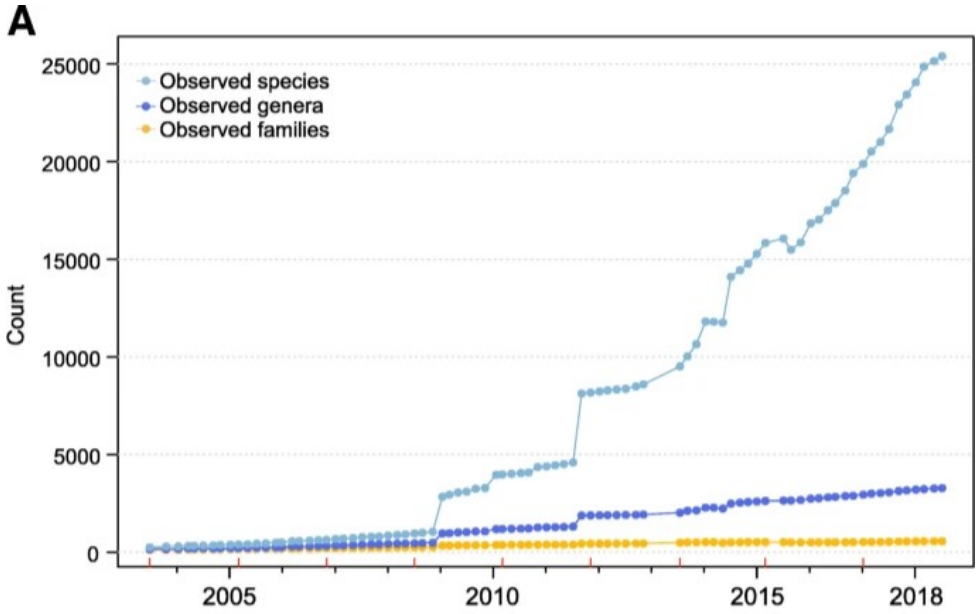

Figura 4.1: Aumento del numero di specie, generi e famiglie osservati[3]

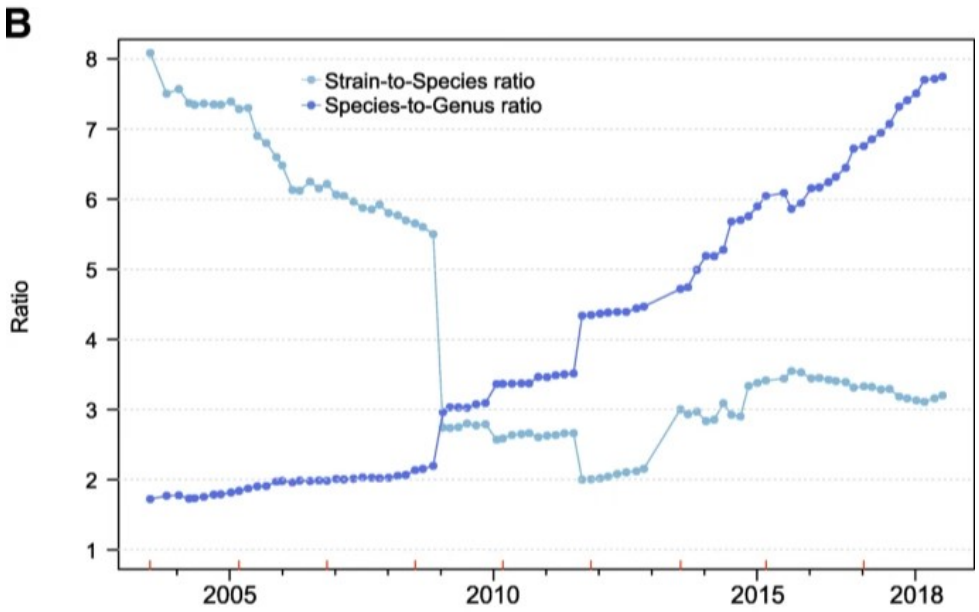

Figura 4.2: Rapporto ceppi-specie e rapporto specie-generi[3]

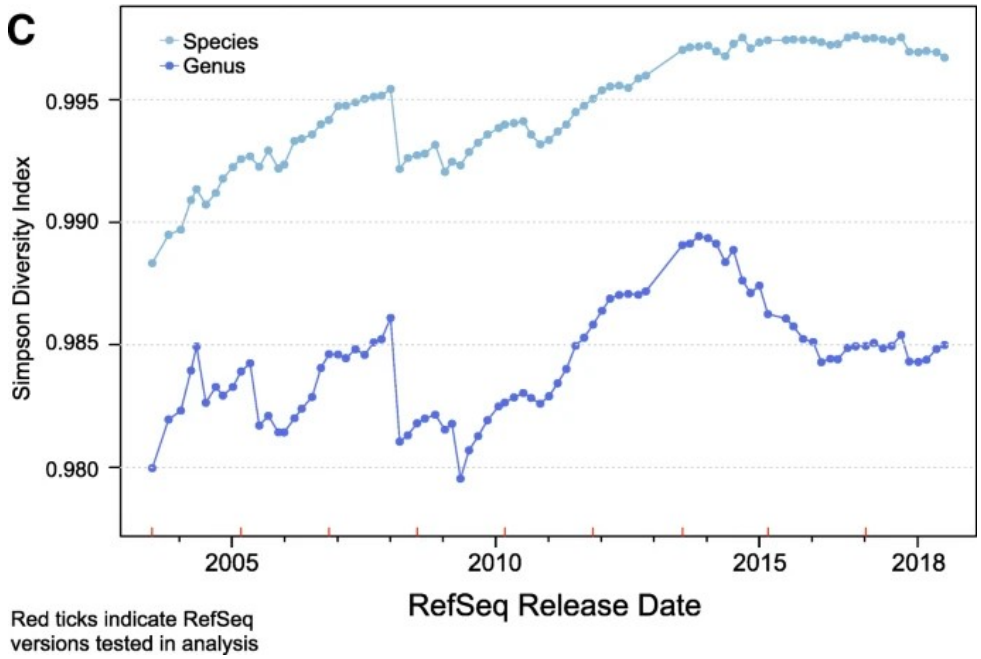

Figura 4.3: Indice di diversità di Simpson per specie e per generi [3]

La diversità dei generi è diminuita nella maggior parte delle versioni dal novembre 2013.

- **A** - Il numero di specie osservate ha superato il numero di generi osservati, man mano che RefSeq è cresciuto.
- **B** - Il rapporto tra ceppi e specie tende a diminuire, mentre il rapporto tra specie e generi tende ad aumentare con la crescita di RefSeq.
- **C** - Indice di diversità di Simpson delle specie in RefSeq batterico tende ad aumentare ad ogni rilascio, mentre la diversità di Simpson dei generi tendeva ad aumentare fino a novembre 2013, quando ha iniziato a diminuire.

Capitolo 5

Estrapolazione ed analisi di dati

All'interno del presente capitolo sono mostrate le tabelle che raggruppano i dati necessari allo svolgimento dell'analisi dei database oggetto di questa trattazione. In aggiunta, vengono inseriti anche i grafici associati ad essi, così da rendere la lettura dei dati più semplice ed intuitiva. Le suddette informazioni sono state reperite analizzando a mano ogni singolo database in esame - basi di dati consultabili nel dettaglio in figura 2.3:

- Standard 4 settembre 2024
- Standard 26 settembre 2022
- Standard 19 settembre 2020
- Plus PFP 4 settembre 2024
- Plus PFP 8 settembre 2022
- Plus PFP 19 settembre 2020

Questi database, discendendo da RefSeq, hanno avuto una crescita simile a quest'ultimo. Il grafico sottostante permette di visualizzare chiaramente questo andamento.

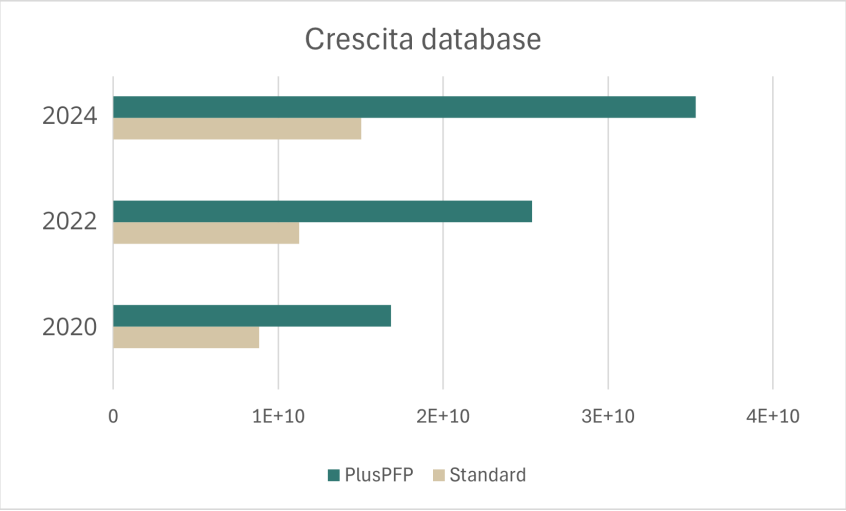

Figura 5.1: Crescita numero minimizers totale Standard e PlusPFP

5.1 Dati acquisiti dalla seconda colonna - Numero minimizers

5.1.1 Database Standard

Standard Generi			
Genoma	**2020**	**2022**	**2024**
Escherichia	23104354	28107421	33579637
Salmonella	17834838	19707280	16560050
Pseudomonas fluorescens group	57646130	56212721	52854442
Serratia	42572204	47744463	51035491
Proteus	11303417	13426967	15161511

Figura 5.2: Database Standard - Numero minimizers dei generi in esame

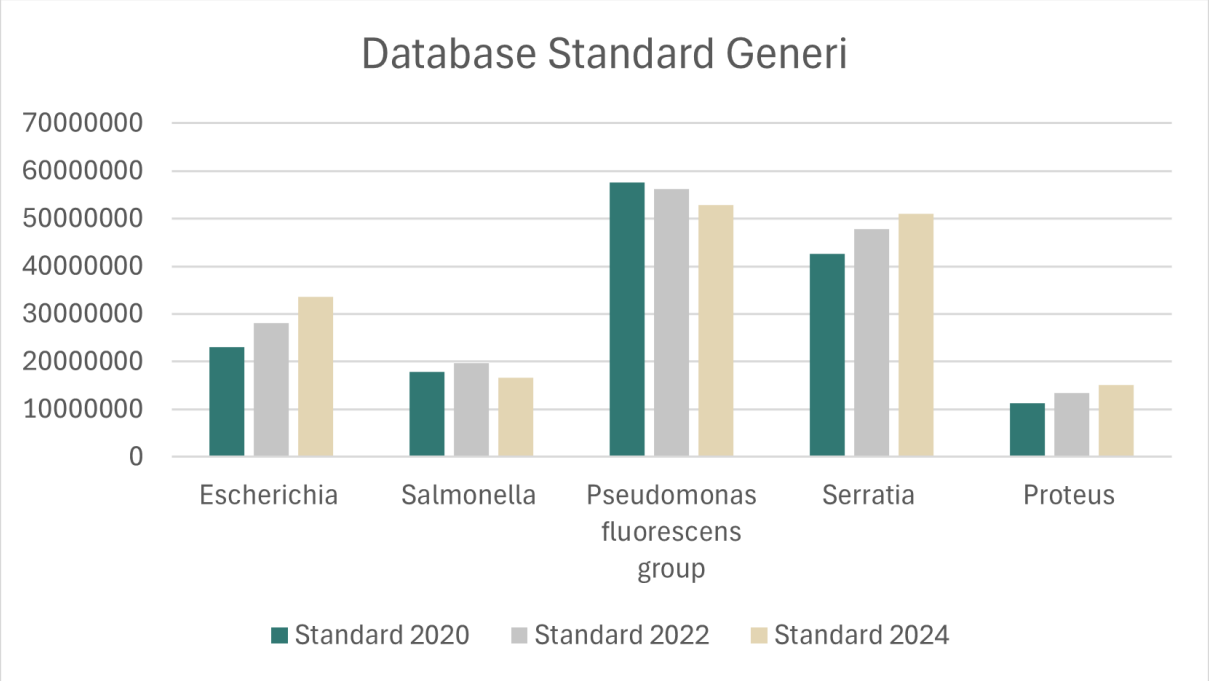

Figura 5.3: Grafico associato

Standard Specie			
Genoma	**2020**	**2022**	**2024**
Escherichia coli	15351495	19646568	23196177
Salmonella enterica	14291134	15347113	12026911
Pseudomonas fluorescens	22351073	21066749	16341701
Serratia marcescens	7401286	8024286	7997582
Proteus mirabilis	3457807	3511147	3764977

Figura 5.4: Database Standard - Numero minimizers delle specie in esame

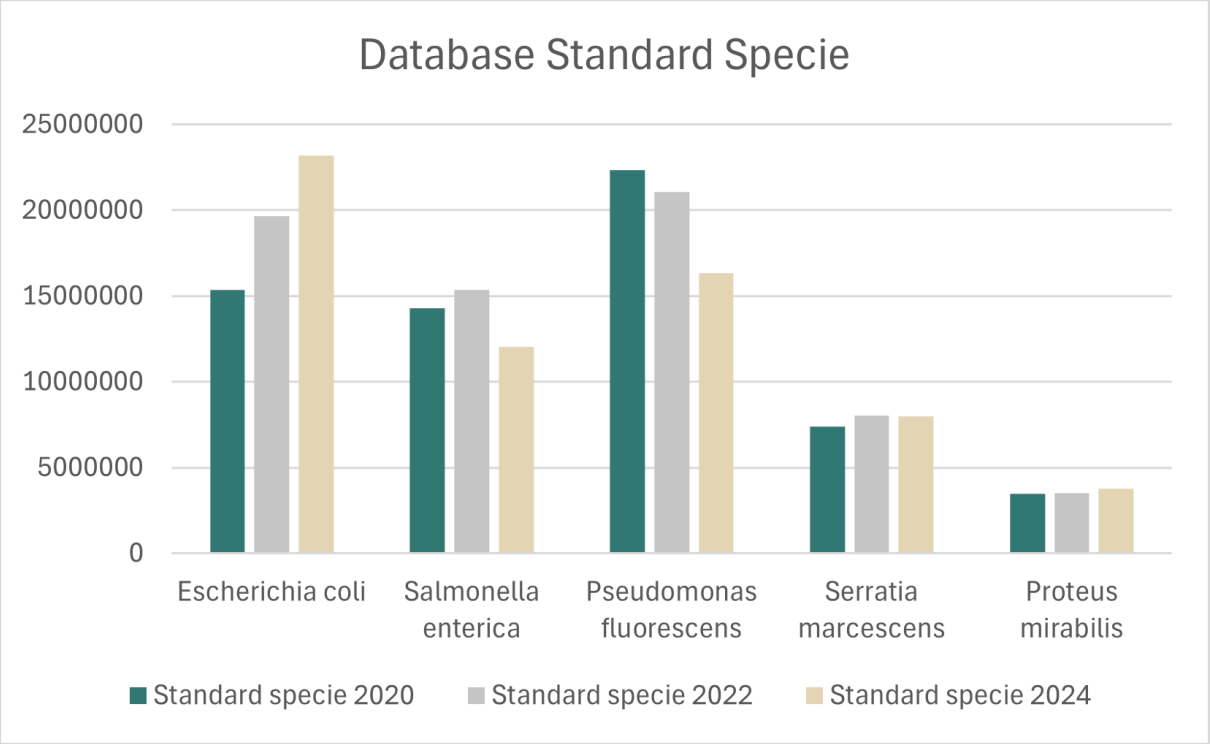

Figura 5.5: Grafico associato

5.1.2 Database Plus PFP

PlusPFP Generi			
Genoma	**2020**	**2022**	**2024**
Escherichia	23093889	28252168	33546191
Salmonella	17828750	19811607	16546893
Pseudomonas fluorescens group	57637201	56256757	52837412
Serratia	42561825	47797183	51008099
Proteus	11291357	13434333	15122169

Figura 5.6: Database PlusPFP - Numero minimizers dei generi in esame

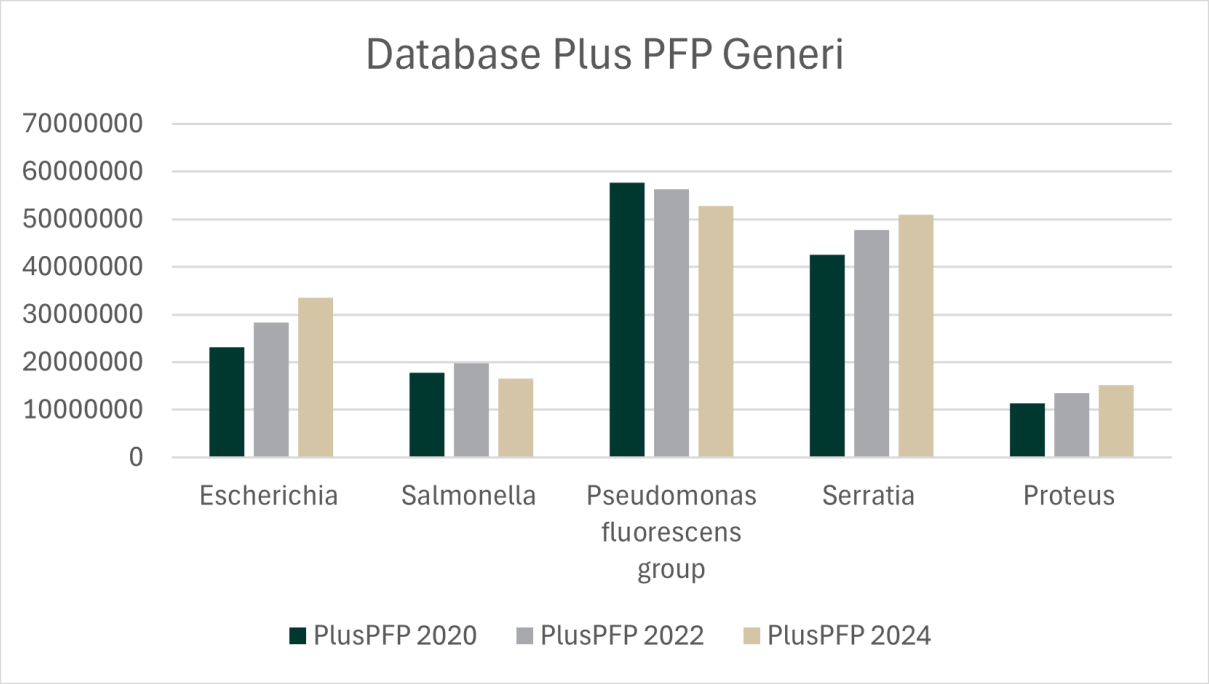

Figura 5.7: Grafico associato

PlusPFP Specie			
Genoma	**2020**	**2022**	**2024**
Escherichia coli	15343722	19764999	23172019
Salmonella enterica	14285953	15442846	12016671
Pseudomonas fluorescens	22348653	21079129	16336653
Serratia marcescens	7399532	8055834	7992573
Proteus mirabilis	3453363	3524498	3749991

Figura 5.8: Database PlusPFP - Numero minimizers delle specie in esame

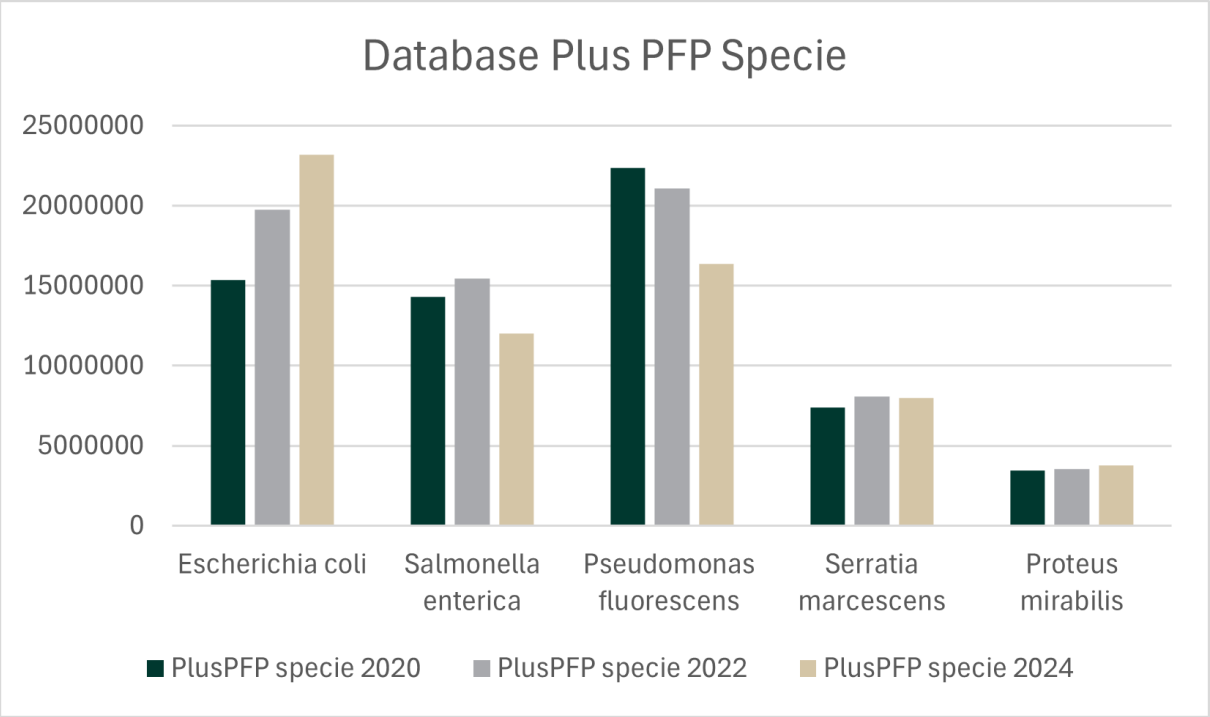

Figura 5.9: Grafico associato

5.1.3 Analisi in merito al numero di minimizers

I grafici relativi a questo valore indicano in modo pressoché uniforme che il suo andamento generale è molto altalenante sia a livello di genere che di specie. Ad esempio, è possibile vedere in figura 5.5 che l'evoluzione del numero di minimizers nel tempo varia molto tra le specie. Uno dei motivi per cui la misura in esame aumenta è l'aggiunta di nuove specie all'interno del database di riferimenti. Questa crescita porta a due trend principali.

Il primo è rappresentato da **Pseudomonas Fluorescens**, in cui l'espansione del database sembra ridurre i minimizers associati. Questo accade perché i dati inseriti non vanno ad accrescere il suo sottoalbero. Di conseguenza, alcuni minimizers vengono associati a più specie rispetto alla versione precedente della base di dati. Ciò sposta 'in alto' la classificazione all'interno dell'albero tassonomico e riduce la **sensibilità**(capacità di identificare le specie presenti nel campione).

L'andamento opposto è rappresentato da **Escherichia Coli**. In questo caso i nuovi inserimenti espandono il sottoalbero della specie. Portando ad un aumento del numero di minimizers associati. Ciò comporta un maggiore utilizzo di memoria e tempi di esecuzione più lunghi, ma può migliorare la sensibilità nella classificazione, talvolta riducendo la **precisione**(proporzione di classificazioni corrette sul totale delle classificazioni effettuate) [1].

5.2 Dati acquisiti dalla prima colonna - valori percentuali

5.2.1 Database Standard

Standard Generi			
Genoma	**2020**	**2022**	**2024**
Escherichia	0,26	0,25	0,22
Salmonella	0,2	0,18	0,11
Pseudomonas fluorescens group	0,65	0,5	0,35
Serratia	0,48	0,42	0,34
Proteus	0,13	0,12	0,1

Figura 5.10: Database Standard - Valori percentuali dei generi in esame

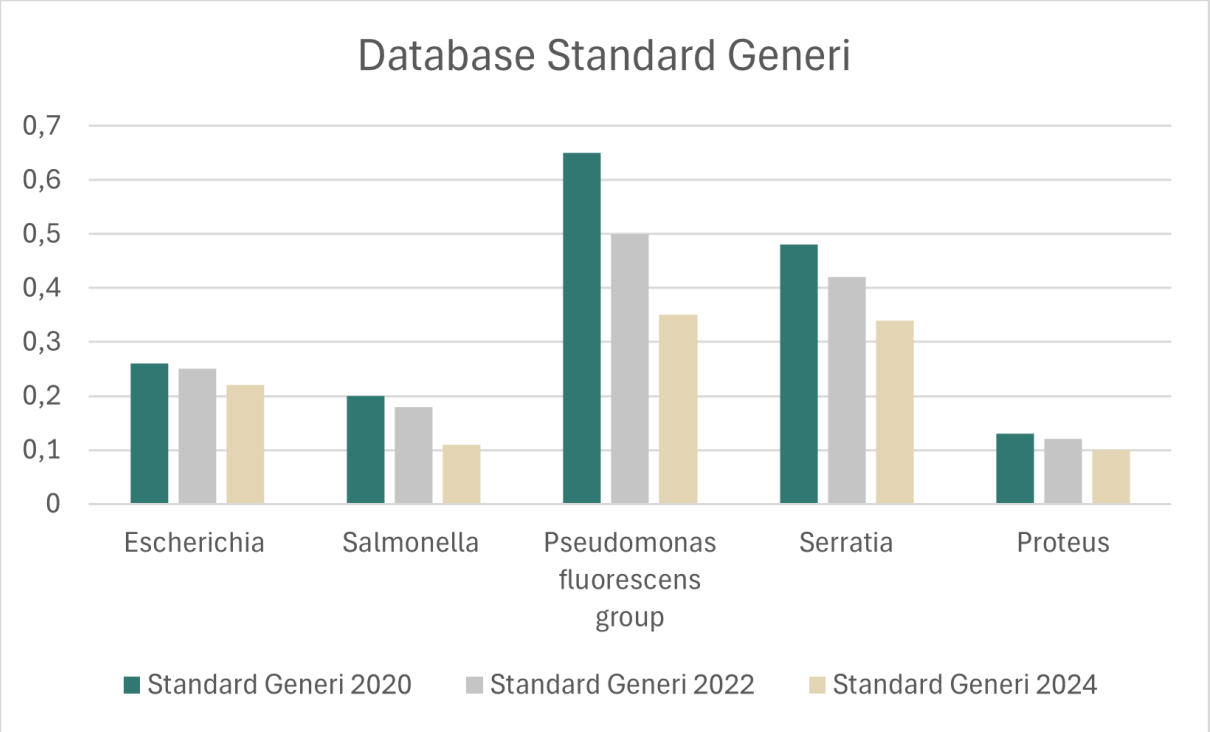

Figura 5.11: Grafico associato

Standard Specie			
Genoma	**2020**	**2022**	**2024**
Escherichia coli	0,17	0,17	0,15
Salmonella enterica	0,16	0,14	0,08
Pseudomonas fluorescens	0,25	0,19	0,11
Serratia marcescens	0,08	0,07	0,05
Proteus mirabilis	0,04	0,03	0,03

Figura 5.12: Database Standard - Valori percentuali delle specie in esame

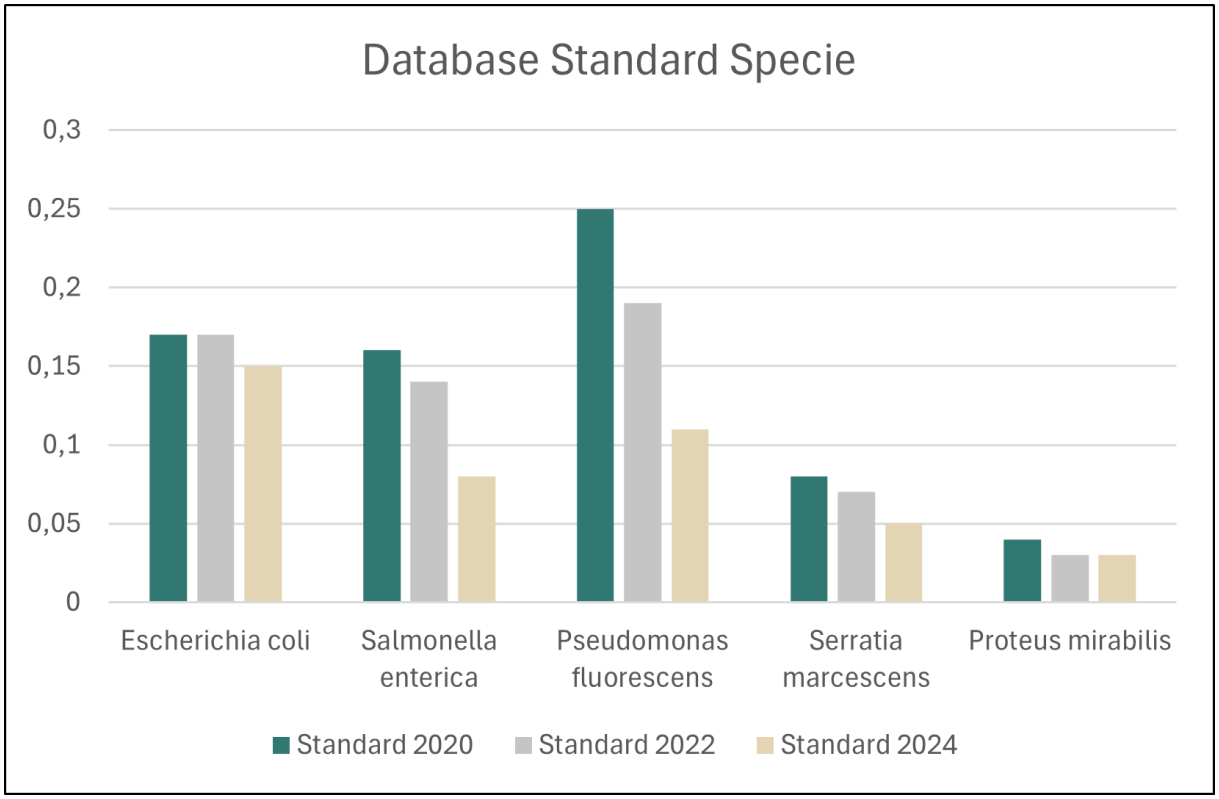

Figura 5.13: Grafico associato

5.2.2 Database Plus PFP

PlusPFP Generi			
Genoma	**2020**	**2022**	**2024**
Escherichia	0,14	0,11	0,1
Salmonella	0,11	0,08	0,05
Pseudomonas fluorescens group	0,34	0,22	0,15
Serratia	0,25	0,19	0,14
Proteus	0,07	0,05	0,04

Figura 5.14: Database PlusPFP - Valori percentuali dei generi in esame

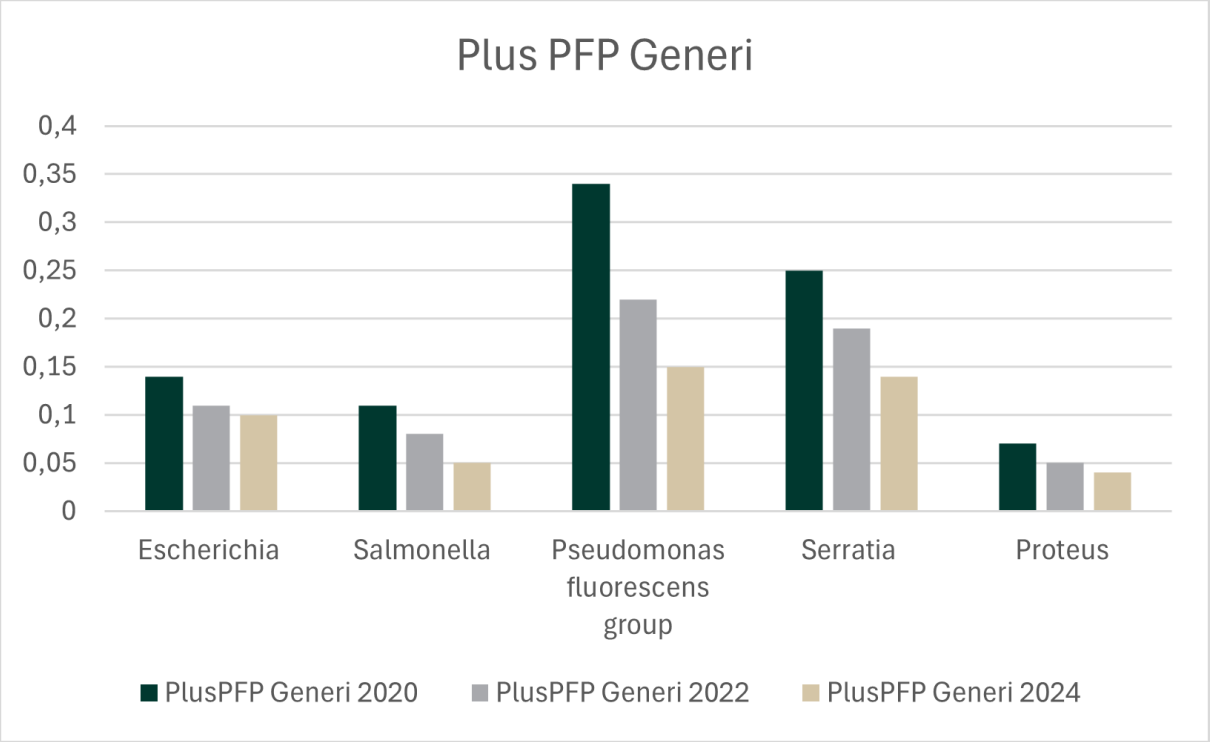

Figura 5.15: Grafico associato

PlusPFP Speice			
Genoma	**2020**	**2022**	**2024**
Escherichia coli	0,09	0,08	0,07
Salmonella enterica	0,08	0,06	0,03
Pseudomonas fluorescens	0,13	0,08	0,05
Serratia marcescens	0,04	0,03	0,02
Proteus mirabilis	0,02	0,01	0,01

Figura 5.16: Database PlusPFP - Valori percentuali delle specie in esame

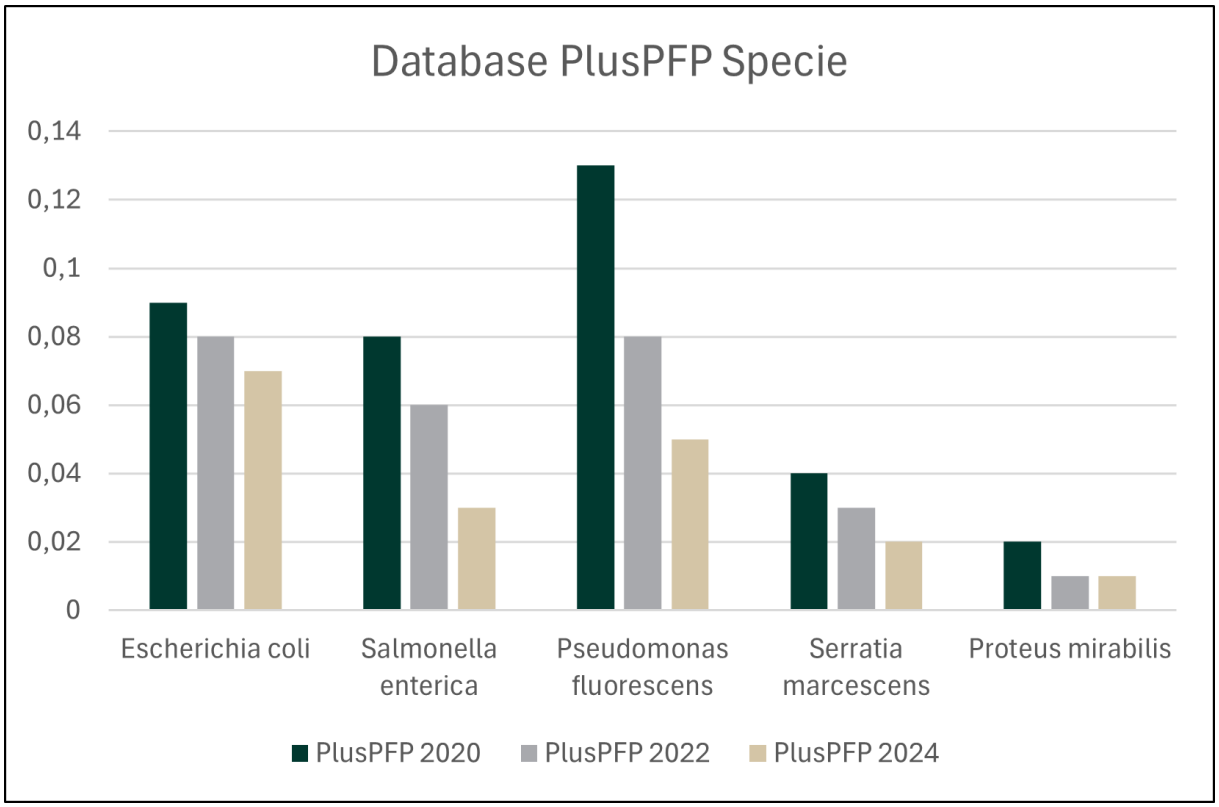

Figura 5.17: Grafico associato

5.2.3 Analisi in merito ai valori percentuali

Il trend generale suggerito da tutti i grafici (es. figura 5.17) che rappresentano l'evoluzione negli anni della percentuale di letture assegnate ad un determinato taxa, è la costante diminuzione di questo valore. Il fenomeno in questione è dovuto principalmente all'aggiunta di nuove specie nel database. Ciò rende la classificazione a questo livello più difficile, perchè i k-mer che in precedenza erano discriminanti per un determinato taxon ora potrebbero essere condivisi da un numero maggiore di specie.
Una crescita del database di questo tipo porta il software Kraken2 a non essere più in grado di completare una classificazione specifica, ottenendo così risultati a livelli tassonomici superiori (solitamente passando al livello tassonomico precedente - il genere)[3].
Di conseguenza, l'espansione del database non si traduce necessariamente in una classificazione più precisa. Al contrario, porta ad avere una quantità maggiore di dati superflui da confrontare, ed una necessità sempre crescente di risorse per poter usufruire di questi essenziali strumenti bioinformatici.

Capitolo 6

Conclusioni

La crescita del database di riferimenti non è sempre vantaggiosa per l'accuratezza della classificazione. Riempire la base di dati con specie non presenti nel campione analizzato può compromettere i risultati, specialmente ai livelli più specifici dell'albero tassonomico.
Come discusso nel capitolo precedente, l'aumento della dimensione del database porta ad una diminuzione delle classificazioni a livello di specie, mentre aumenta la frazione di classificazioni a livello di genere. Questo è dovuto all'approccio LCA, dove sequenze condivise vengono assegnate all'antenato comune più basso tra i taxa corrispondenti [3].
In aggiunta al problema portato dalla semplice crescita, l'incremento di specie nei database non è uniforme [5]. Ciò porta ad una rappresentazione degli organismi non omogenea attraverso l'albero della vita e ulteriori imprecisioni nella classificazione tassonomica.
Dunque la scelta del database di riferimento è cruciale. I ricercatori dovrebbero prestare attenzione alla selezione della base di dati e operare una scelta ponderata a partire dalle specie attese nel campione da analizzare. Un'indagine esplorativa preliminare di quest'ultimo può essere utile per adattare il database alle sue caratteristiche specifiche. La soluzione ottimale, sebbene impossibile, sarebbe utilizzare il più piccolo dataset che contiene tutte le specie presenti nel campione.

Elenco delle figure

2.1 Funzionamento Kraken 2 4
2.2 Albero tassonomico 5
2.3 Database di riferimenti presi in analisi 7

3.1 Parametri per la configurazione di Kraken 2 9
3.2 Riga del database che rappresenta il genere Escherichia, presa come esempio . 10

4.1 Aumento del numero di specie, generi e famiglie osservati[3] 12
4.2 Rapporto ceppi-specie e rapporto specie-generi[3] 12
4.3 Indice di diversità di Simpson per specie e per generi [3] 13

5.1 Crescita numero minimizers totale Standard e PlusPFP 15
5.2 Database Standard - Numero minimizers dei generi in esame 16
5.3 Grafico associato 16
5.4 Database Standard - Numero minimizers delle specie in esame 17
5.5 Grafico associato 17
5.6 Database PlusPFP - Numero minimizers dei generi in esame 18
5.7 Grafico associato 18
5.8 Database PlusPFP - Numero minimizers delle specie in esame 19
5.9 Grafico associato 19
5.10 Database Standard - Valori percentuali dei generi in esame 21
5.11 Grafico associato 21
5.12 Database Standard - Valori percentuali delle specie in esame 22
5.13 Grafico associato 22
5.14 Database PlusPFP - Valori percentuali dei generi in esame 23
5.15 Grafico associato 23
5.16 Database PlusPFP - Valori percentuali delle specie in esame 24
5.17 Grafico associato 24

Bibliografia

[1] D. E. Wood, J. Lu e B. Langmead, «Improved metagenomic analysis with Kraken 2,» *Genome biology*, vol. 20, pp. 1–13, 2019.

[2] K. D. Pruitt, T. Tatusova e D. R. Maglott, «NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins,» *Nucleic acids research*, vol. 33, n. suppl_1, pp. D501–D504, 2005.

[3] D. J. Nasko, S. Koren, A. M. Phillippy e T. J. Treangen, «RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification,» *Genome biology*, vol. 19, pp. 1–10, 2018.

[4] K. D. Pruitt, T. Tatusova, W. Klimke e D. R. Maglott, «NCBI Reference Sequences: current status, policy and new initiatives,» *Nucleic acids research*, vol. 37, n. suppl_1, pp. D32–D36, 2009.

[5] M. Gray, Z. Zhao e G. L. Rosen, «How Scalable Are Clade-Specific Marker K-Mer Based Hash Methods for Metagenomic Taxonomic Classification?» *Frontiers in Signal Processing*, vol. 2, 2022, issn: 2673-8198. doi: 10.3389/frsip.2022.842513. indirizzo: https://www.frontiersin.org/journals/signal-processing/articles/10.3389/frsip.2022.842513.