UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA INFORMATICA

"CLASSIFICAZIONE TASSONOMICA: COME LA DIMENSIONE DEL DATABASE INFLUENZA LA PRECISIONE CON KRAKEN2"

Relatore: Prof. Comin Matteo

Laureando: Leonardo Lazzaro

ANNO ACCADEMICO 2021 – 2022

Data di laurea 25/11/2022

Alla mia famiglia
e alle persone che hanno preso parte a questo percorso

La metagenomica è la disciplina che si occupa dello studio di sequenze geniche ottenute direttamente dall'ambiente dove più microorganismi convivono, con l'obiettivo di caratterizzarne la diversità tassonomica. I software di classificazione tassonomica etichettano sequenze di DNA utilizzando come riferimento database contenenti informazioni genomiche. Esistono collezioni di dati genici specifici per particolari entità biologiche o database ad ampio spettro che raccolgono il DNA di diversi gruppi di microorganismi. La dimensione del database utilizzato può influenzare la validità delle classificazioni tassonomiche, sia dal punto di vista di disponibilità di memoria nel sistema, sia per il dettaglio delle sequenze presenti al suo interno. In questo studio viene analizzata la validità delle classificazioni tassonomiche di tre dataset: l'SRR1804065, contenente DNA umano, il CAMI2 Marine di origine marina ed infine un dataset simulato contenente sequenze genomiche di 10 virus e 40 batteri. Ogni dataset viene classificato attraverso il software Kraken2, affiancato dai maggiori database realizzati per quest'ultimo, tra i quali Standard, PlusPF e PlusPFP. Per poter valutare i risultati sono state misurate la sensitività, precisione ed F_1-measure per ogni classificazione tassonomica effettuata, sia a livello di geni che di specie.

Indice

1 Introduzione

Durante i due anni di pandemia SARS-CoV-2, il sequenziamento metagenomico ha permesso agli Stati di tutto il mondo di rispondere efficacemente alla necessità di rilevare il virus negli esseri umani, attraverso tecniche minimamente invasive, come i tamponi naso-faringei. Il sequenziamento del genoma del virus è stato eseguito in appena 6 giorni e ciò ha consentito agli esperti di poterne studiare la natura, monitornarne l'evoluzione e le mutazioni[1].

Nel febbraio del 2015, sul New York Times è stato pubblicato un articolo che riportava la scoperta, da parte di un gruppo di ricercatori della Weill Cornell Medical College, di tracce di materiale genetico del batterio Yersinia Pestis (agente eziologico della peste bubbonica) nei pressi della metropolitana di New York. Viene poco dopo ritrattata la versione dallo stesso team, ipotizzando che i geni raccolti potessero appartenere ad un diverso batterio[2].

A causa dei limiti tecnologici e la ridotta potenza di calcolo, la capacità dello studio genomico si fermava al laboratorio. Come spiegato in The New Science of Metagenomics, «Storicamente, lo studio dei microbi si è concentrato prevalentemente su singole specie in colture di laboratorio, per cui la comprensione delle comunità microbiche è in ritardo rispetto a quella dei loro singoli membri». Nasce attorno al 1980, grazie al progresso scientifico-tecnologico, la metagenomica, disciplina che si occupa dello studio di sequenze geniche ottenute direttamente dall'ambiente dove più microorganismi convivono, con l'obiettivo di caratterizzarne la diversità tassonomica. Come spiegato in *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*, «La radice "Meta" significa che questa nuova scienza cerca di comprendere la biologia a livello aggregato, trascendendo il singolo organismo per concentrarsi sui geni della comunità e su come essi possano influenzare le attività reciproche nel servire le funzioni collettive. Allo stesso tempo, vi è anche la necessità di sviluppare metodi computazionali che massimizzino la comprensione della composizione genetica e delle attività di comunità così complesse che possono essere solo campionate, ma mai completamente caratterizzate» [1]. Questi processi vengono effettuati utilizzando software di classificazione tassonomica che, affiancati da database contenenti le informazioni genomiche dei microrganismi, etichettano i campioni sequenziati di DNA in esame.

La diversità dei dati presenti, il livello di dettaglio e la complessità del database sono parametri da mettere a punto per ottenere prestazioni accettabili e risultati affidabili, in relazione a ciò che il problema di sequenziamento richiede.

La validità della classificazione metagenomica è quindi un fattore di estrema rilevanza. In primis, per l'affidabilità intrinseca dei risultati ottenuti, i quali dovranno essere studiati e analizzati. Ma non è da sottovalutare, visto il periodo storico che stiamo attraversando, la risonanza mediatica di questa disciplina, che necessita di elevata accuratezza per evitare conseguenze circa la fiducia nella

[1] https://www.izsvenezie.it/sequenziato-genoma-virus-sars-cov-2-veneto/

[2] https://www.nytimes.com/2015/02/07/nyregion/bubonic-plague-in-the-subway-system-dont-worry-about-it.html

classe scientifica da parte della popolazione mondiale.

Viene proposta un'analisi della bontà della classificazione tassonomica rispetto a 22 dei maggiori database, di diverse dimensioni. Verrà preso in esame il software Kraken2 e per valutare i risultati sono state selezionate tre metriche: la precisione, la sensitività e la F_1-measure.

Definiamo per precisione il rapporto tra il numero delle classificazioni corrette di un genoma sul totale delle volte che il modello lo prevede. Per sensitività intendiamo invece la frequenza con cui il software etichetta correttamente un genoma nell'insieme delle sequenze geniche analizzate. Dunque un modello con un'elevata sensitività riporta un basso tasso di falsi negativi. Infine la F_1-measure, è la media armonica tra precisione e sensitività. Tale metrica può essere compresa in un range tra 0 e 1, dove l'estremo superiore indica la massima accuratezza del modello.

2 Setup sperimentale

2.1 Il software: Kraken2

Il software di classificazione tassonomica che verrà utilizzato è Kraken2, seconda versione della famiglia di programmi Kraken. Quest'ultimo assegna etichette tassonomiche a sequenze di DNA, esaminando le k-mers di una sequenza genetica e utilizza queste informazioni per interrogare il database di riferimenti. Come indicato in *Improved metagenomic analysis with Kraken2*, rispetto alla prima versione, Kraken2 riduce dell'85% l'utilizzo di memoria durante la classificazione, permettendo di sfruttare grandi quantità di referenze genomiche, mantenendo un'alta precisione. Viene inoltre quintuplicata la velocità di elaborazione rispetto a Kraken.

Kraken2 implementa un algoritmo ad utilizzo intensivo di memoria, in grado di associare alle sottostringhe genomiche, in gergo k-mers, il corrispondente lowest common ancestor (LCA). Il requisito di un'elevata disponibilità di memoria nella prima versione del software (si superavano facilmente i 100 Gb di RAM) ha portato molti ricercatori ad utilizzare database a sensibilità ridotta o utilizzare subset dei database stessi.

Con Kraken2, gli sviluppatori introducono un hash-table compatta e probabilistica per mappare le coppie k-mers/LCA. Inoltre memorizza soltanto l-mers, detti anche minimizers, di lunghezza minore o uguale alle k-mers, al contrario di Kraken, che memorizzava tutte le k-mers.

La compact hash-table (CHT) di Kraken2 è molto simile ad una hash-table tradizionale che utilizza il linear probing[3] per la risoluzione delle collisioni, con alcune modifiche. Come riportato da Wood, D.E., Lu, J. & Langmead, B. in Improved metagenomic analysis with Kraken2 [2], viene utilizzato un array di celle hash da 32 bit per le coppie chiave-valore. In base agli identificatori da rappresentare presenti nella libreria di riferimenti, varia il numero di bit necessari per ogni cella, che vengono memorizzati nei bit meno significativi di queste ultime. Nel restante spazio vengono scritti i bit più significativi del codice hash della chiave, eseguendo un'operazione di riduzione di dimensionalità e creando un codice hash compatto. La ricerca di una chiave K nella CHT comincia con il calcolo del codice hash della chiave, h(K). Quindi viene eseguita una scansione lineare dell'array iniziando dalla posizione $h(K)\%|T|$, con T numero di celle dell'array, fino alla chiave corrispondente.

Kraken2, utilizzando codici hash compatti da 32 bit offre, rispetto ai 96 bit della prima versione, un notevole risparmio in termini di memoria.

Adottando questo approccio nasce una problematica sulle collisioni tra due chiavi distinte, che condividono lo stesso codice hash compatto. Esse possono essere trattate come identiche se le loro posizioni iniziali di ricerca sono vicine abbastanza da trovare un codice hash memorizzato corrispondente prima di una cella vuota. Viene conferita la natura probabilistica di questa struttura dati,

[3] Quando la funzione di hash causa una collisione, mappando una nuova chiave in una cella della hash table già occupata da un'altra chiave, la tecnica di linear probing cerca la cella libera più vicina e vi inserisce la nuova chiave

ottenendo due possibili risultati: una chiave che non è stata inserita può essere riportata come presente nella tabella, oppure i valori di due chiavi possono essere confusi l'uno con l'altro. Nel primo caso siamo di fronte ad un falso positivo, il secondo invece assegna un LCA meno specifico al minimizzatore [3]. Vi è una probabilità minore dell'1% che questo accada, con un fattore di carico impostato al 70% (di default su Kraken2). Viene offerta la possibilità di compensare alla probabilità d'errore sfruttando la *Kraken's confidence scoring thresholds*, in grado di scartare risultati a bassa fiducia e migliorare la precisione.

La prima versione di Kraken utilizza la rappresentazione tassonomica inserita dall'utente, mentra Kraken2 effettua modifiche trovando un insieme minimo di nodi, assegnando ad ognuno un ID sequenzialmente crescente, attraverso una ricerca breadth-first (BFS) a partire dalla radice. La rappresentazione interna della tassonomia consente a Kraken2 di utilizzare il numero minimo di bit per la memorizzazione dei numeri ID della tassonomia, offrendo il massimo spazio per i codici hash compatti e riducendo la probabilità di errori dovuti alla CHT.

2.2 Hardware utilizzato

Le prove sperimentali di classificazione tassonomica, utilizzando il software Kraken2, sono state eseguite utilizzando il cluster di calcolo Blade, installato presso il dipartimento DEI dell'Università di Padova. Il sistema è fornito di 12 server, 16 processori di diversa potenza per un totale di 304 CPUs, 9 Terabyte di RAM e 15 GPUs, ripartiti nel seguente modo[4]:

- runner-[01-03]: 3 nodi con 48 CPUs (4x Intel(R) Xeon(R) Gold 5118 CPU @ 2.30/3.20GHz), 1.5TB RAM
- runner-[04-06]: 3 nodi con 72 CPUs (4x Intel(R) Xeon(R) Gold 5220 CPU @ 2.20/3.90GHz), 2TB RAM, 1 Nvidia Quadro P2000 GPU
- runner-[07-09]: 3 nodi con 96 CPUs (4x Intel(R) Xeon(R) Gold 6252N CPU @ 2.30/3.60GHz, 3TB RAM
- gpu1: 24 CPUs (2x Intel(R) Xeon(R) Gold 5118 CPU @ 2.30/3.20GHz), 1TB RAM, 6 Nvidia Titan RTX GPUs
- gpu[2-3]: 2 nodi da 32 CPUs (2x Intel(R) Xeon(R) Gold 5218 CPU @ 2.30/3.90GHz), 1.5TB RAM, 8 Nvidia RTX 3090 GPUs

2.3 I database genomici

I database genomici utilizzati sono stati scaricati dalla repository ufficiale del software «Kraken 2 / Bracken Refseq indexes». Nella tabella[5]:

[4] https://clusterdeiguide.readthedocs.io/en/latest/Overview.html

[5] La tabella è presente all'indirizzo: https://benlangmead.github.io/aws-indexes/k2 seguente, sono riportati i 12 database

Collection	Contains	Date	Archive size (GB)	Index size (GB)
Viral	viral	9/8/2022	0.4	0.5
MinusB3	archaea, viral, plasmid, human1, UniVec_Core	9/26/2022	5.9	8.5
Standard3	archaea, bacteria, viral, plasmid, human1, UniVec_Core	9/26/2022	46	60
Standard-83	Standard with DB capped at 8 GB	9/26/2022	5.5	7.5
Standard-163	Standard with DB capped at 16 GB	9/26/2022	11	15
PlusPF	Standard plus protozoa & fungi	9/8/2022	49	64
PlusPF-8	PlusPF with DB capped at 8 GB	9/8/2022	5.5	7.5
PlusPF-16	PlusPF with DB capped at 16 GB	9/8/2022	11	15
PlusPFP	Standard plus protozoa, fungi & plant	9/8/2022	99	129
PlusPFP-8	PlusPFP with DB capped at 8 GB	9/8/2022	5.1	7.5
PlusPFP-16	PlusPFP with DB capped at 16 GB	9/8/2022	11	15
EuPathDB462	Eukaryotic pathogen genomes with contaminants removed	11/13/2020	26.4	34.1
Standard	archaea, bacteria, viral, plasmid, human1, UniVec_Core	6/7/2022	44	58
PlusPF	Standard plus protozoa & fungi	6/7/2022	47	61
PlusPFP	Standard plus protozoa, fungi & plant	6/7/2022	55	129
Standard	archaea, bacteria, viral, plasmid, human1, UniVec_Core	5/17/2021	38.6	50.1
PlusPF	Standard plus protozoa & fungi	5/17/2021	41.0	53.2
Standard	archaea, bacteria, viral, plasmid, human1, UniVec_Core	12/2/2020	36.0	46.8
PlusPF	Standard plus protozoa & fungi (fixed from 12/2/20 version3)	1/27/2021	38.4	49.8
PlusPFP	Standard plus protozoa, fungi & plant (fixed from 12/2/20 version3)	1/27/2021	71.8	96.3
Standard	archaea, bacteria, viral, plasmid, human1, UniVec_Core	9/19/2020	36.0	47.0
PlusPF	Standard plus protozoa & fungi	9/19/2020	37.0	48.0
PlusPFP	Standard plus protozoa, fungi & plant	9/19/2020	66.5	90.0

Table 1: Database utilizzati

2.4 I dataset metagenomici

I dataset metagenomici che verranno classificati sono tre. Abbiamo a disposizione il dataset SRR1804065, campione del *Human Microbiome Project* (HMP), scaricabile dal sito NCBI; oltre a questo, un dataset simulato, con genomi di 40 batteri e 10 virus. Infine l'ultimo contenente sequenze genetiche di organismi marini.

I dataset presi in esame sono in formato FASTQ, basato su testo che permette la memorizzazione di sequenze biologiche e il loro punteggio di qualità utilizzando un singolo carattere ASCII ciascuno.

Le sequenze in questi files sono inoltre paired-end, ciò rende il processo di sequenziamento più sensibile e accurato rispetto a sequenziamenti single-end, poiché facilita le operazioni di *alignment*, consentendo il rilevamento di eventuali delezioni, duplicazioni o inserzioni nel DNA.

Dataset	**No. di specie**
Simulated (x10)	40 Batteri, 10 Virus
CAMI2 Marine	4956
SRR1804065	2537

Table 2: Numero di specie presenti nei dataset analizzati [4]

3 Prove sperimentali

La classificazione tassonomica dei dataset reali e simulati è stata effettuata utilizzando lo script *bash*:

```
./kraken2 --db $db_path --threads 16 \
--paired $file1 $file2 > result.out
```

Mentre per il dataset marino, non essendo *paired end*, lo script viene modificato nel modo seguente:

```
./kraken2 --db $db_path --threads 16 \
$file1 > result.out
```

```
C	SRR1804065.1	28116	100|100 A:1 816:61 28116:4 |:| 28116:11 816:3 28116:10 816:12 28116:7 816:1 0:22
C	SRR1804065.11	292800	100|100 0:66 |:| 0:1 292800:5 0:13 292800:3 0:44
C	SRR1804065.17	909656	100|100 0:66 |:| 909656:26 976:8 2:32
C	SRR1804065.18	28116	100|100 A:1 816:65 |:| 816:14 28116:5 816:5 28116:34 816:4 28116:4
C	SRR1804065.21	815	100|100 0:66 |:| 2:28 815:10 0:28
C	SRR1804065.25	718255	100|100 A:1 718255:65 |:| 0:28 718255:1 0:5 718255:32
C	SRR1804065.28	28116	100|100 28116:1 816:65 |:| 816:3 976:4 816:59
C	SRR1804065.33	817	100|100 A:1 0:4 817:4 0:6 817:49 816:2 |:| 816:12 817:41 816:6 817:1 816:3 817:3
C	SRR1804065.37	28116	100|100 816:66 |:| 816:39 28116:4 816:5 28116:1 816:4 28116:4 816:5 28116:1 816:3
C	SRR1804065.42	1679	100|100 1679:11 216816:5 1678:1 216816:45 1678:2 216816:2 |:| 1679:11 216816:5 1679:7 1678:5 1679:1 216816:8 1678:29
C	SRR1804065.43	816	100|100 816:31 171549:3 816:32 |:| 816:9 815:1 816:6 815:2 816:43 815:3 816:2
C	SRR1804065.46	28116	100|100 A:1 816:65 |:| 816:55 28116:7 816:4
C	SRR1804065.48	2093857	100|100 2093857:66 |:| 2093857:66
C	SRR1804065.50	28116	100|100 816:15 0:3 816:5 0:7 816:1 0:5 816:2 0:4 28116:4 816:9 28116:11 |:| 816:25 815:4 816:10 28116:19 816:8
C	SRR1804065.57	171549	100|100 A:1 171549:65 |:| 171549:66
C	SRR1804065.58	357276	100|100 909656:60 357276:6 |:| 0:1 909656:5 0:2 909656:4 357276:10 815:2 357276:2 909656:1 357276:30 909656:9
C	SRR1804065.69	28116	100|100 28116:7 816:9 28116:7 816:5 28116:2 816:36 |:| 28116:1 816:12 28116:5 816:32 28116:16
C	SRR1804065.71	816	100|100 816:66 |:| 816:66
C	SRR1804065.72	171549	100|100 171549:66 |:| 171549:66
C	SRR1804065.75	816	100|100 A:1 0:65 |:| 816:2 2:3 816:61
C	SRR1804065.76	657313	100|100 657313:10 0:56 |:| 0:66
C	SRR1804065.82	357276	100|100 A:1 909656:54 357276:6 909656:4 357276:1 |:| 909656:39 0:26 357276:1
C	SRR1804065.87	909656	100|100 0:66 |:| 815:7 171549:11 2:8 1:7 2:9 909656:5 0:4 816:4 0:11
```

Figure 1: Esempio di file .out

Lo script esegue il software Kraken2, che richiede in input il database genomico e il dataset da classificare. Per ogni dataset genomico è stata effettuata la classificazione con ognuno dei database presenti nella tabella 1. Una volta ottenuti i file di output (con estensione .out), vengono convertiti in file con estensione .res, conservando, delle cinque colonne presenti per ogni riga, solamente la *Sequence ID* e la *Taxonomy ID*, attraverso il comando:

```
cut -f 2-3 result.out > result.res
```

```
SRR1804065.11   292800
SRR1804065.17   909656
SRR1804065.18   28116
SRR1804065.21   815
SRR1804065.25   718255
SRR1804065.28   28116
SRR1804065.33   817
SRR1804065.37   28116
SRR1804065.42   1679
SRR1804065.43   816
SRR1804065.46   28116
SRR1804065.48   2093857
SRR1804065.50   28116
SRR1804065.57   171549
SRR1804065.58   357276
SRR1804065.69   28116
SRR1804065.71   816
SRR1804065.72   171549
SRR1804065.75   816
SRR1804065.76   657313
SRR1804065.82   357276
SRR1804065.87   909656
```

Figure 2: Esempio di file .res

Una volta ottenuto il file .res, viene eseguito l'ultimo script che ne valuta la precisione in termini di specie e geni. Lo script richiede in input il file *nodes.dmp*, contenente la tassonomia del database completo di Kraken2 ed il file *ground truth*, in cui sono presenti informazioni tassonomiche ritenute vere, utilizzate per validare i dati ottenuti in output dal software Kraken2.

Di seguito viene riportato lo script che valuta la precisione, sensitività ed F_1 in termini di geni:

```
#!/bin/bash

set -e

echo "Evaluation at species level"

while IFS= read -r line
do

        for tool in kraken2
        do
                for db in Viral MinusB Standard Standard8 Standard16 PlusPF PlusPF16 PlusPFP PlusPFP8 PlusPFP16 EuPathDB48
                do
                        echo "Evaluation of REAL_$db.res:"
                        ./evaluate_calls nodes.dmp genus ground_truth.tsv result.res
                done

                for db in Viral MinusB Standard Standard8 Standard16 PlusPF PlusPF16 PlusPFP PlusPFP8 PlusPFP16 EuPathDB48
                do
                        echo "Evaluation of SMLTD_$db.res:"
                        ./evaluate_calls nodes.dmp genus ground_truth.tsv result.res
                done

```

```
                for db in Viral MinusB Standard Standard8 Standard16 PlusPF PlusPF16 PlusPFP PlusPFP8 PlusPFP16 EuPathDB48
                do
                        echo "Evaluation of MARINE_$db.res:"
                        ./evaluate_calls nodes.dmp genus ground_truth.tsv result.res
                done
        done
done < selected_run.csv
```

E lo script che valuta la precisione, sensitività ed F_1 in termini di specie:

```
#!/bin/bash

set -e

echo "Evaluation at species level"

while IFS= read -r line
do

        for tool in kraken2
        do
                for db in Viral MinusB Standard Standard8 Standard16 PlusPF PlusPF16 PlusPFP PlusPFP8 PlusPFP16 EuPathDB48
                do
                        echo "Evaluation of REAL_$db.res:"
                        ./evaluate_calls nodes.dmp species ground_truth.tsv result.res
                done

                for db in Viral MinusB Standard Standard8 Standard16 PlusPF PlusPF16 PlusPFP PlusPFP8 PlusPFP16 EuPathDB48
                do
                        echo "Evaluation of SMLTD_$db.res:"
                        ./evaluate_calls nodes.dmp species ground_truth.tsv result.res
                done

                for db in Viral MinusB Standard Standard8 Standard16 PlusPF PlusPF16 PlusPFP PlusPFP8 PlusPFP16 EuPathDB48
                do
                        echo "Evaluation of MARINE_$db.res:"
                        ./evaluate_calls nodes.dmp species ground_truth.tsv result.res
                done
        done
done < selected_run.csv
```

In output si ottengono diverse metriche riguardanti la classificazione tassonomica, tra cui la sensitività, precisione ed F_1-measure rispettivamente evidenziate in verde.

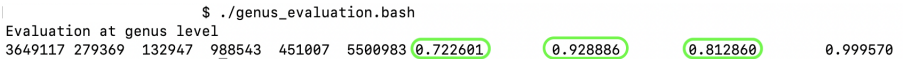

Figure 3: Output dello script di valutazione

4 Risultati

4.1 Viral

Il database Viral contiene le sequenze di DNA ed RNA dell'insieme dei virus, per una dimensione totale di 0.5 GB. Di seguito le metriche rilevate nella classificazione tassonomica dei 3 dataset:

Table 3: Metriche ottenute a livello di geni - Database Viral

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.000005	0.001046	0.000009
Simulato	0.199535	0.965218	0.330705
CAMI2 Marine	0.000001	0.000021	0.000002

Table 4: Metriche ottenute a livello di specie - Database Viral

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.000006	0.001187	0.000012
Simulato	0.171808	0.956872	0.291310
CAMI2 Marine	0.000010	0.000230	0.000018

La diversa natura dei dataset SRR1804065 e CAMI2 Marine rispetto ai dati presenti nel database Viral, non permette di evidenziare alcun dato significativo.

La classificazione tassonomica ha prodotto corrispondenze significative per il dataset Simulato, che ottiene ottime precisioni sia sui geni che sulle specie. L'F_1-score si attesta al 33% a livello di geni e al 29% a livello di specie, rimarcando come nonostante le precisioni siano elevate, l'accuratezza della classificazione tassonomica sia al contrario mediocre.

4.2 MinusB

Il database MinusB è una collezione di insiemi di geni riguardanti archèobatteri, batteri inizialmente individuati negli ambienti più estremi della Terra, ma presenti anche in ambienti più ospitali, virus, plasmidi, il set Human[1], uno dei più recenti modelli metabolici umani e infine UniVec_Core, un insieme di sequenze di acidi nucleici di origine vettoriale. Il database ha una dimensione di 8.5 GB. Le metriche rilevate nella classificazione tassonomica dei 3 dataset:

Table 5: Metriche ottenute a livello di geni - Database MinusB

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.011299	0.232419	0.021551
Simulato	0.324262	0.821019	0.464908
CAMI2 Marine	0.086983	0.685520	0.154377

Table 6: Metriche ottenute a livello di specie - Database MinusB

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.006796	0.105877	0.012773
Simulato	0.280155	0.773112	0.411275
CAMI2 Marine	0.078675	0.632941	0.139954

Il database MinusB non è consigliato per i dataset SRR1804065 e CAMI2 Marine, le cui classificazioni tassonomiche ottengono valori bassi circa le metriche.

Le metriche sulla classificazione tassonomica del dataset Simulato evidenziano che, nonostante le precisioni siano minori rispetto al database Viral, vi è un guadagno in termini di accuratezza, come evidenziato dall'F_1-score.

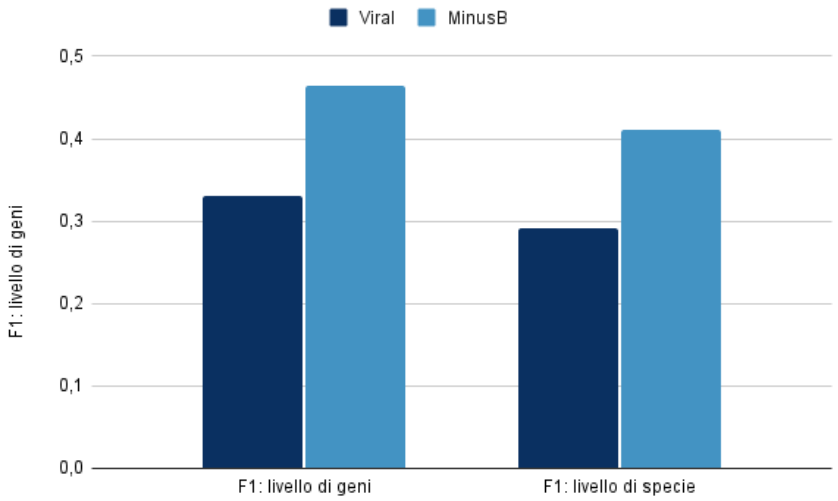

Figure 4: Confronto tra il database Viral e MinusB rispetto al dataset simulato (virus e batteri)

4.3 Database Standard

4.3.1 Standard

Il database Standard ha una dimensione totale di 60 GB e contiene le informazioni genetiche di archèobatteri, batteri, virus, plasmidi, human[1] e UniVec_Core. Le metriche ottenute sono:

Table 7: Metriche ottenute a livello di geni - Database Standard

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.722601	0.928886	0.812860
Simulato	0.973690	0.993822	0.983653
CAMI2 Marine	0.640682	0.812440	0.716410

Table 8: Metriche ottenute a livello di specie - Database Standard

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.465398	0.741643	0.571910
Simulato	0.776065	0.982972	0.867350
CAMI2 Marine	0.531782	0.752057	0.623023

Il database ottiene ottimi risultati nella classificazione dei 3 dataset, con particolarmente elevate percentuali per il dataset simulato. L'F_1-score evidenzia una notevole accuratezza nella classificazione a livello di geni, mentre fatica maggiormente a livello di specie.

4.3.2 Standard-8

Il database Standard-8 è un sottoinsieme del database completo Standard, tale che la dimensione abbia un tetto massimo di 8 GB. Più precisamente il database pesa 7.5 GB. Le metriche ottenute sono:

Table 9: Metriche ottenute a livello di geni - Database Standard-8

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.656019	0.910667	0.762648
Simulato	0.948025	0.990809	0.968945
CAMI2 Marine	0.214910	0.700065	0.328863

Table 10: Metriche ottenute a livello di specie - Database Standard-8

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.369161	0.727900	0.489877
Simulato	0.679675	0.982863	0.803623
CAMI2 Marine	0.170674	0.632813	0.268840

Il "figlio minore" del database Standard ottiene risultati analoghi a quest'ultimo. Le maggiori variazioni sono evidenziate dalle metriche calcolate sul dataset di origine CAMI2 Marine, con una differenza di F_1-score di oltre 30 punti percentuali, sia a livello di geni che di specie. In termini di risparmio di memoria rispetto

a Standard, la classificazione tassonomica è ottima sia sul dataset SRR1804065 che sul CAMI2 Marine.

4.3.3 Standard-16

Il database Standard-16 è un sottoinsieme del database completo Standard, tale che la dimensione abbia un tetto massimo di 16 GB. Più precisamente il database pesa 15 GB. Le metriche ottenute sono:

Table 11: Metriche ottenute a livello di geni - Database Standard-16

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.693456	0.924910	0.792632
Simulato	0.966950	0.993173	0.979886
CAMI2 Marine	0.346395	0.751874	0.474284

Table 12: Metriche ottenute a livello di specie - Database Standard-16

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.423404	0.747347	0.540558
Simulato	0.733681	0.985309	0.841078
CAMI2 Marine	0.277841	0.687858	0.395807

Standard-16 si avvicina maggiormente ai risultati ottenuti dal database Standard, riuscendo addirittura a migliorarne la precisione. Si riconferma il trend evidenziato per Standard-8: i dataset SRR1804065 e Simulato producono analoghe classificazioni tassonomiche affiancate a Standard-16, nonostante il risparmio di oltre 45 GB di memoria. Per quanto riguarda il dataset CAMI2 Marine, vi è un guadagno in termini di F_1-score rispetto a Standard-8, ma rimane una differenza di circa 20 punti percentuali rispetto alla versione originale del database.

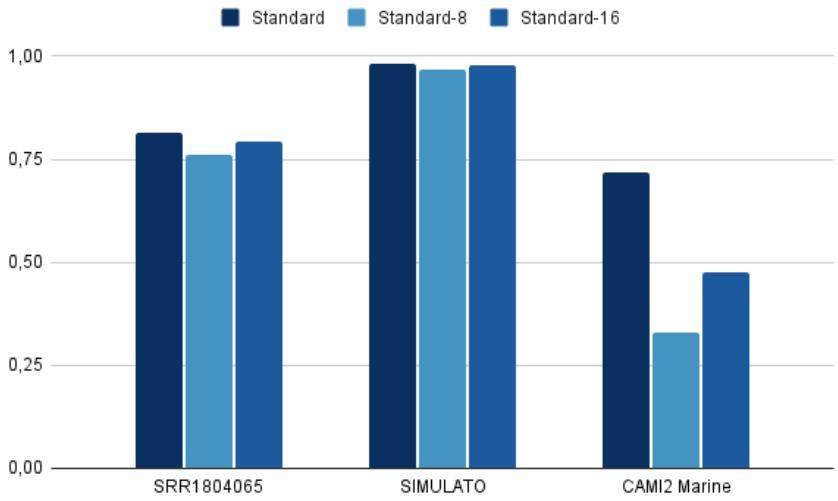

Figure 5: Confronto tra i database della famiglia Standard: precisione sui geni

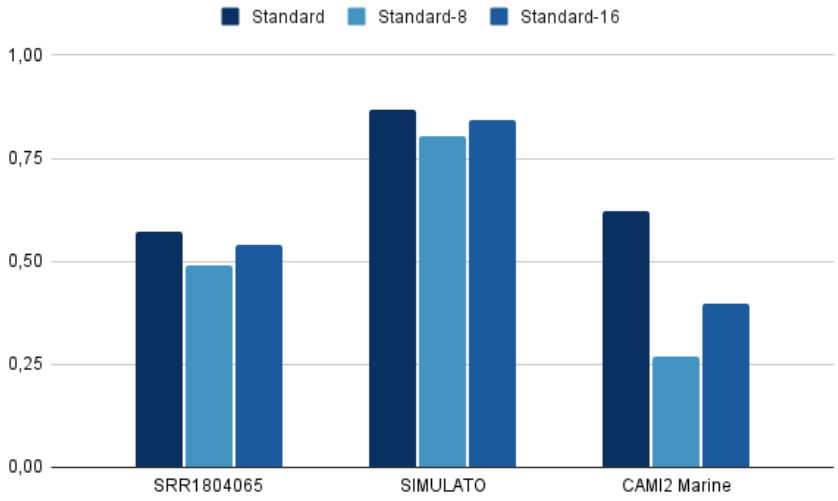

Figure 6: Confronto tra i database della famiglia Standard: precisione sulle specie

4.3.4 Standard (6/7/2022)

Table 13: Metriche ottenute a livello di geni - Database Standard (6/7/2022)

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.723455	0.928783	0.813361
Simulato	0.973981	0.993827	0.983804
CAMI2 Marine	0.643828	0.817662	0.720407

Table 14: Metriche ottenute a livello di specie - Database Standard (6/7/2022)

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.475128	0.747390	0.580942
Simulato	0.781781	0.983213	0.871003
CAMI2 Marine	0.539739	0.758494	0.630687

4.3.5 Standard (17/5/2021)

Table 15: Metriche ottenute a livello di geni - Database Standard (17/5/2021)

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.730214	0.932852	0.819188
Simulato	0.968638	0.995847	0.982054
CAMI2 Marine	0.651088	0.837255	0.732528

Table 16: Metriche ottenute a livello di specie - Database Standard (17/5/2021)

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.493393	0.754166	0.596525
Simulato	0.781231	0.986000	0.871752
CAMI2 Marine	0.552880	0.782565	0.647971

4.3.6 Standard (27/1/2021)

Table 17: Metriche ottenute a livello di geni - Database Standard (27/1/2021)

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.732624	0.930387	0.819747
Simulato	0.970407	0.997762	0.983894
CAMI2 Marine	0.672536	0.852437	0.751875

Table 18: Metriche ottenute a livello di specie - Database Standard (27/1/2021)

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.519467	0.770008	0.620398
Simulato	0.783233	0.988628	0.874026
CAMI2 Marine	0.576783	0.805829	0.672334

4.3.7 Standard (19/9/2020)

Table 19: Metriche ottenute a livello di geni - Database Standard (19/9/2020)

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.752645	0.933581	0.833406
Simulato	0.969628	0.998082	0.983649
CAMI2 Marine	0.673178	0.853771	0.752795

Table 20: Metriche ottenute a livello di specie - Database Standard (19/9/2020)

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.561810	0.799822	0.660014
Simulato	0.783704	0.989119	0.874511
CAMI2 Marine	0.577292	0.806804	0.673019

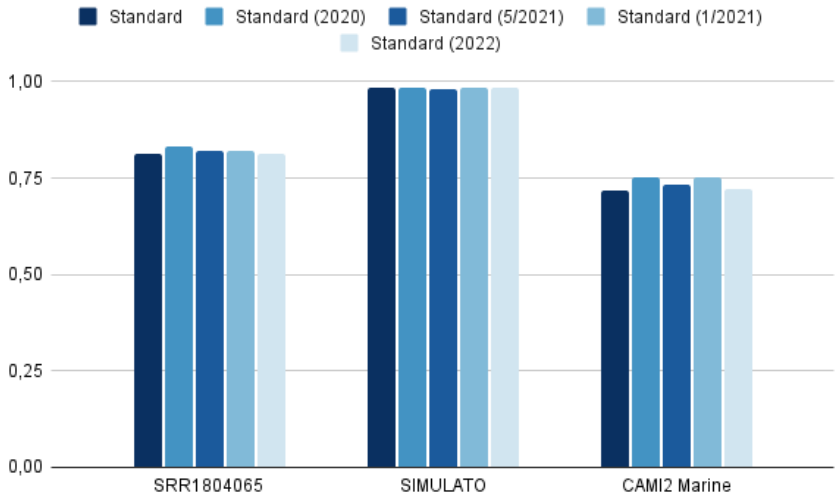

Figure 7: Database Standard a confronto (livello di geni)

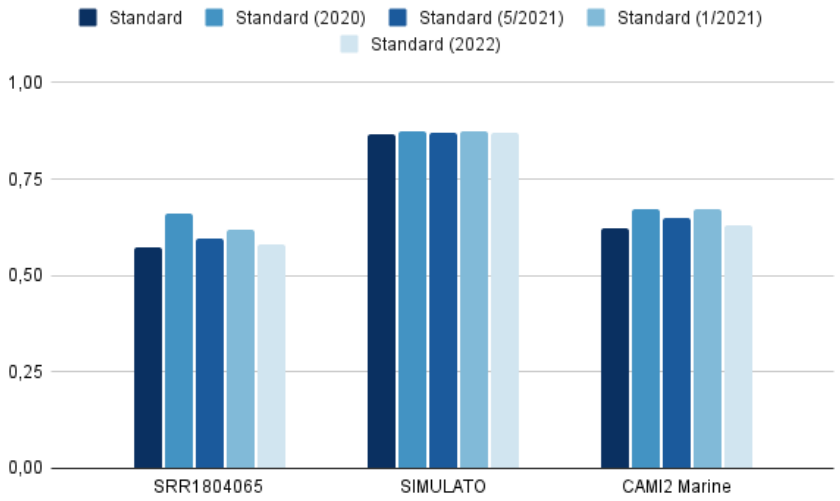

Figure 8: Database Standard a confronto (livello di specie)

Il confronto con le versioni precedenti del database Standard evidenzia un leggero peggioramento dell'F_1-measure sull'edizione più recente, sia a livello di geni che di specie. Le prestazioni migliori le ottiene il database Standard datato 2020 su tutti e tre i dataset.

4.4 PlusPF

4.4.1 PlusPF

Il database PlusPF memorizza i dati del database Standard, aggiungendo le sequenze genetiche di protozoi e funghi. La sua dimensione è di 64 GB totali. Le metriche ottenute sono:

Table 21: Metriche ottenute a livello di geni - Database PlusPF

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.722719	0.928785	0.812896
Simulato	0.973701	0.993849	0.983672
CAMI2 Marine	0.638921	0.808703	0.713855

Table 22: Metriche ottenute a livello di specie - Database PlusPF

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.465461	0.741424	0.571892
Simulato	0.776504	0.982901	0.867596
CAMI2 Marine	0.530491	0.748236	0.620824

Nonostante il maggior peso in memoria di 4 GB, ottiene risultati analoghi al database Standard, sia a livello di geni che di specie. L'F_1-score per i tre dataset è comunque maggiore per il database PlusPF, ma con differenze infinitesimali.

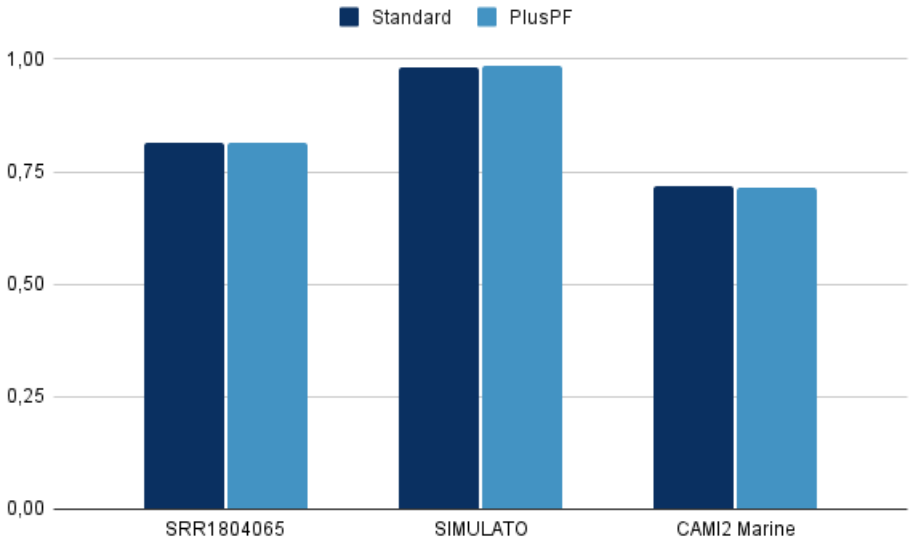

Figure 9: Precisioni sui geni, ottenute dai database Standard e PlusPF

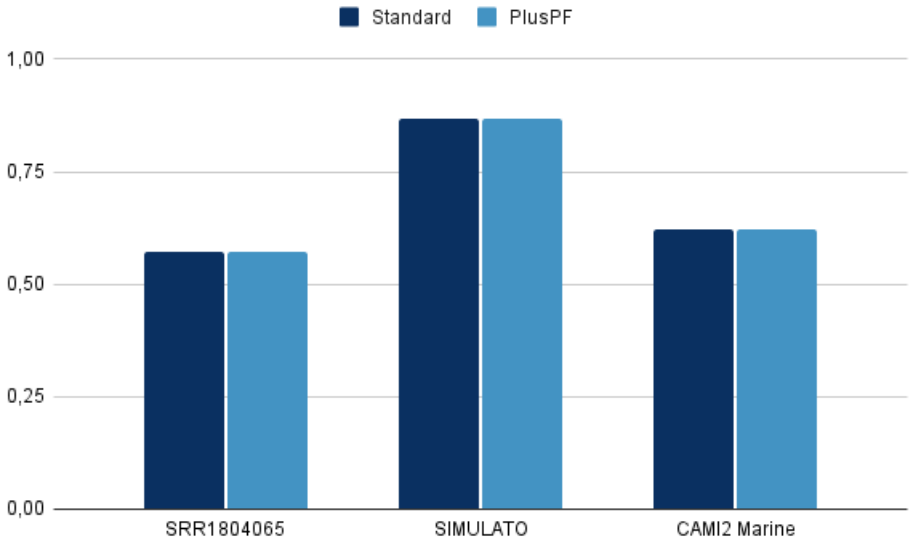

Figure 10: Precisioni sulle specie, ottenute dai database Standard e PlusPF

4.4.2 PlusPF-16

PlusPF-16 è un sottoinsieme del database PlusPF tale che la dimensione abbia un tetto massimo di 16 GB. Più precisamente il database pesa 15 GB. Le metriche ottenute sono:

Table 23: Metriche ottenute a livello di geni - Database PlusPF-16

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.691160	0.923902	0.790761
Simulato	0.966092	0.992870	0.979298
CAMI2 Marine	0.327684	0.710222	0.448458

Table 24: Metriche ottenute a livello di specie - Database PlusPF-16

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.419448	0.746003	0.536976
Simulato	0.729844	0.985033	0.838452
CAMI2 Marine	0.262561	0.644395	0.373101

La versione da 16 GB del database PlusPF ottiene risultati simili alla controparte della famiglia Standard. Mettendo a confronto le metriche ottenute tra il database PlusPF e la versione da 16GB possiamo evidenziare l'affinità dello score F_1, sia per il dataset SRR1804065, sia per il simulato. La differenza maggiore è rimarcata nello score F_1 della classificazione del dataset CAMI2 Marine, con una perdita di accuratezza di oltre 30 punti percentuali.

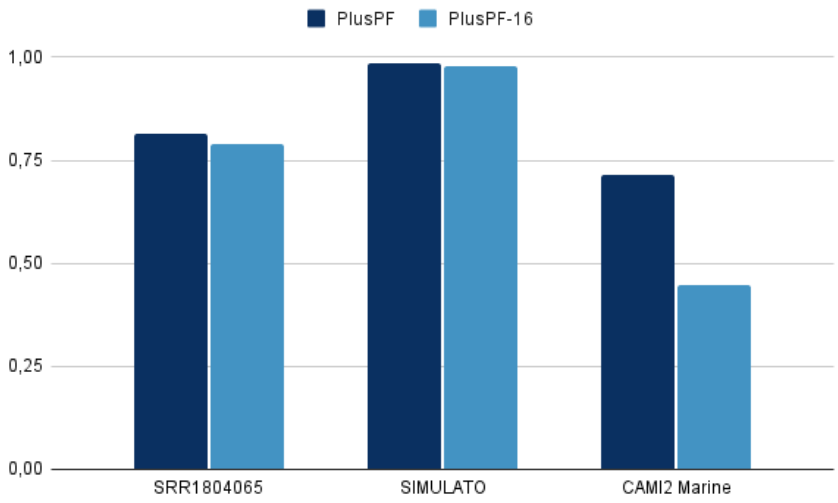

Figure 11: Database della famiglia PlusPF a confronto (F_1 a livello di geni)

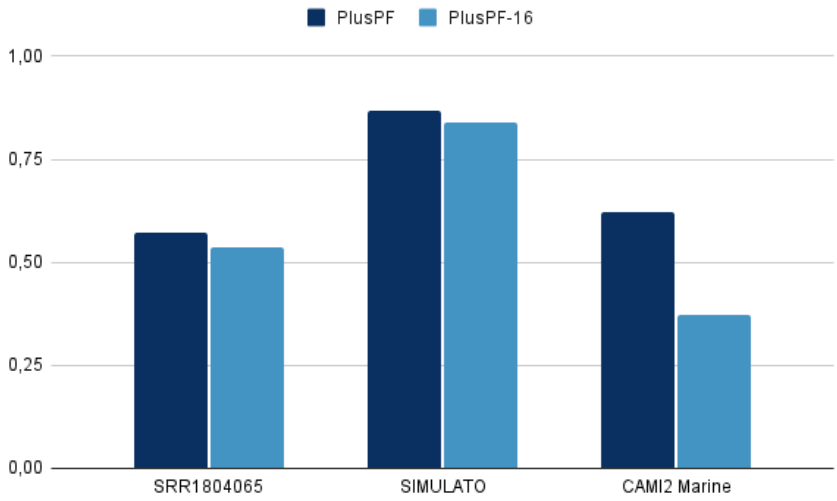

Figure 12: Database della famiglia PlusPF a confronto: precisione sulle specie

4.4.3 PlusPF (7/6/2022)

Table 25: Metriche ottenute a livello di geni - Database PlusPF (7/6/2022)

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.723465	0.928776	0.813364
Simulato	0.973975	0.993822	0.983798
CAMI2 Marine	0.642248	0.814106	0.718037

Table 26: Metriche ottenute a livello di specie - Database PlusPF (7/6/2022)

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.475126	0.747366	0.580933
Simulato	0.781774	0.983207	0.870996
CAMI2 Marine	0.538522	0.754876	0.628604

4.4.4 PlusPF (17/5/2021)

Table 27: Metriche ottenute a livello di geni - Database PlusPF (17/5/2021)

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.730240	0.932836	0.819198
Simulato	0.968629	0.995835	0.982044
CAMI2 Marine	0.649446	0.833435	0.730026

Table 28: Metriche ottenute a livello di specie - Database PlusPF (17/5/2021)

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.493393	0.754117	0.596510
Simulato	0.781222	0.985986	0.871741
CAMI2 Marine	0.551563	0.778655	0.645725

4.4.5 PlusPF (27/1/2021)

Table 29: Metriche ottenute a livello di geni - Database PlusPF (27/1/2021)

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.732634	0.930407	0.819760
Simulato	0.970396	0.997751	0.983884
CAMI2 Marine	0.670786	0.849168	0.749509

Table 30: Metriche ottenute a livello di specie - Database PlusPF (27/1/2021)

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.519464	0.770006	0.620395
Simulato	0.783220	0.988617	0.874013
CAMI2 Marine	0.575345	0.802439	0.670177

4.4.6 PlusPF (19/9/2020)

Table 31: Metriche ottenute a livello di geni - Database PlusPF (19/9/2020)

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.752639	0.933559	0.833393
Simulato	0.969620	0.998074	0.983641
CAMI2 Marine	0.672541	0.852474	0.751893

Table 32: Metriche ottenute a livello di specie - Database PlusPF (19/9/2020)

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.561809	0.799807	0.660008
Simulato	0.783692	0.989109	0.874500
CAMI2 Marine	0.576769	0.805475	0.672201

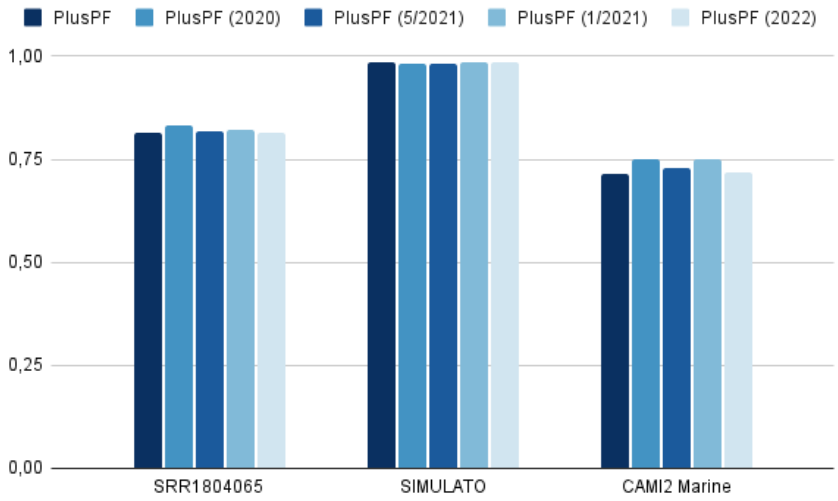

Figure 13: Database PlusPF a confronto (livello di geni)

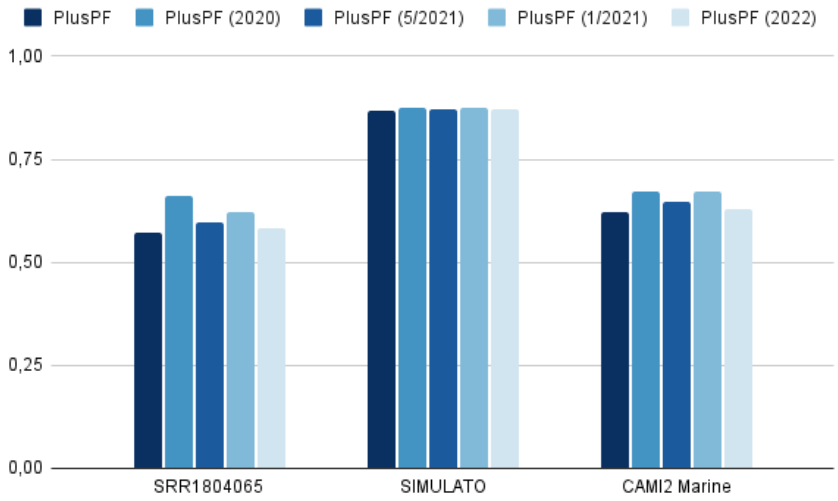

Figure 14: Database PlusPF a confronto (livello di specie)

Come per il database Standard, l'ultima versione di PlusPF peggiora l'F_1-measure rispetto alle versioni degli anni precedenti. Il database PlusPF datato 2020 risulta sia per il dataset SRR1804065, sia per il simulato, il più prestazionale, sia a livello di geni che di specie, mentre ottiene risultati analoghi al PlusPF (1/2021) per quanto riguarda la classificazione tassonomica di CAMI2 Marine.

4.5 PlusPFP

4.5.1 PlusPFP

Il database PlusPFP contiene lo Standard, aggiungendo le informazioni genetiche di protozoi, funghi e piante. La sua dimensione totale è di 129 GB. Le metriche ottenute sono:

Table 33: Metriche ottenute a livello di geni - Database PlusPFP

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.722902	0.928457	0.812886
Simulato	0.973611	0.993599	0.983504
CAMI2 Marine	0.603566	0.739063	0.664477

Table 34: Metriche ottenute a livello di specie - Database PlusPFP

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.465446	0.740826	0.571703
Simulato	0.776420	0.982616	0.867433
CAMI2 Marine	0.503247	0.678052	0.577716

Confrontando i 3 database maggiori, Standard, PlusPF e PlusPFP si può evidenziare come le metriche ottenute sui dataset SRR1804065 e Simulato siano pressoché equivalenti. Questo non si riconferma per il dataset CAMI2 Marine, dove si può apprezzare la differenza maggiore in termini di F_1-score: PlusPFP risulta il database meno accurato dei tre, sia a livello di geni che di specie.

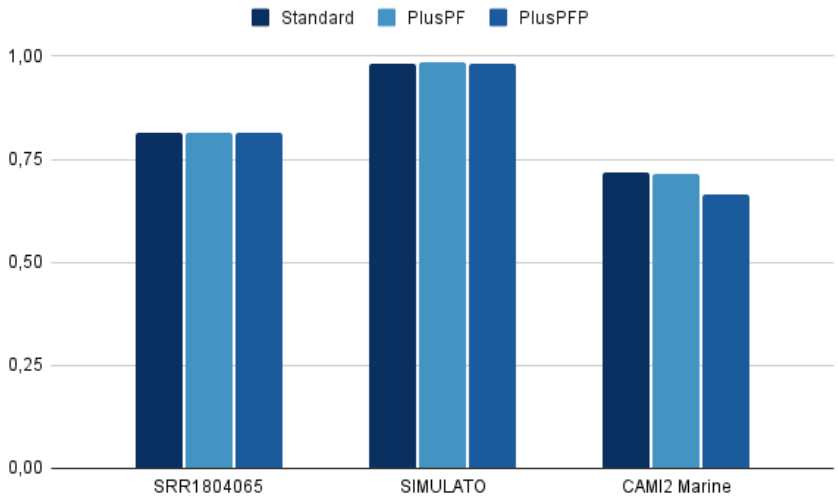

Figure 15: Standard, PlusPF e PlusPFP a confronto: precisione sui geni

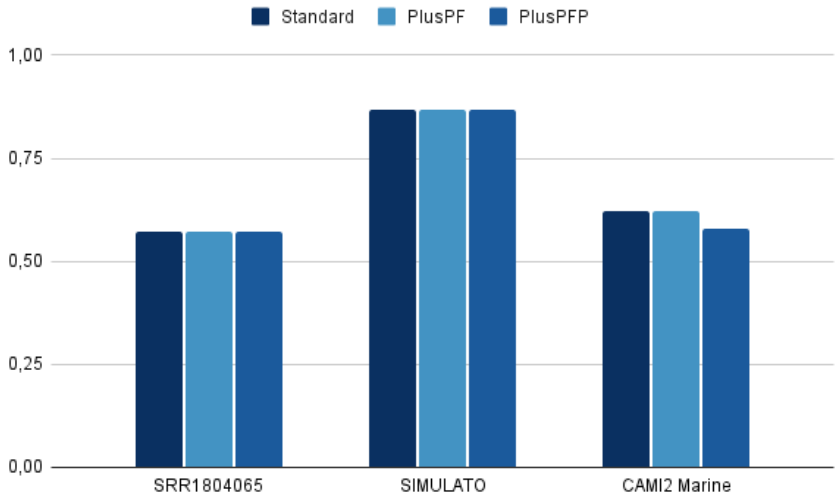

Figure 16: Standard, PlusPF e PlusPFP a confronto: precisione sulle specie

4.5.2 PlusPFP-8

PlusPFP-8 è un sottoinsieme del database PlusPFP tale che la dimensione abbia un tetto massimo di 8 GB. Più precisamente il database pesa 7.5 GB. Le metriche ottenute sono:

Table 35: Metriche ottenute a livello di geni - Database PlusPFP-8

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.541236	0.884050	0.671415
Simulato	0.827192	0.984918	0.899191
CAMI2 Marine	0.106696	0.475179	0.174263

Table 36: Metriche ottenute a livello di specie - Database PlusPFP-8

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.269400	0.680361	0.385969
Simulato	0.545548	0.974072	0.699389
CAMI2 Marine	0.084133	0.409135	0.139567

4.5.3 PlusPFP-16

PlusPFP-16 è un sottoinsieme del database PlusPFP tale che la dimensione abbia un tetto massimo di 16 GB. Più precisamente il database pesa 15 GB. Le metriche ottenute sono:

Table 37: Metriche ottenute a livello di geni - Database PlusPFP-16

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.645180	0.906454	0.753819
Simulato	0.939031	0.989354	0.963536
CAMI2 Marine	0.187385	0.555761	0.280271

Table 38: Metriche ottenute a livello di specie - Database PlusPFP-16

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.356779	0.721419	0.477439
Simulato	0.664530	0.980848	0.792283
CAMI2 Marine	0.148661	0.487685	0.227862

La versione originale ottiene uno score-F_1 eccellente per il dataset Simulato e le versioni con tetto ad 8GB e 16GB non si discostano di molto, considerando il grande risparmio in termini di memoria occupata. Le variazioni più marcate sono evidenziate dalle metriche delle classificazioni tassonomiche del dataset CAMI2 Marine. Le versioni da 8GB e 16GB ottengono un F_1-score veramente scadente rispetto alla versione da 129GB.

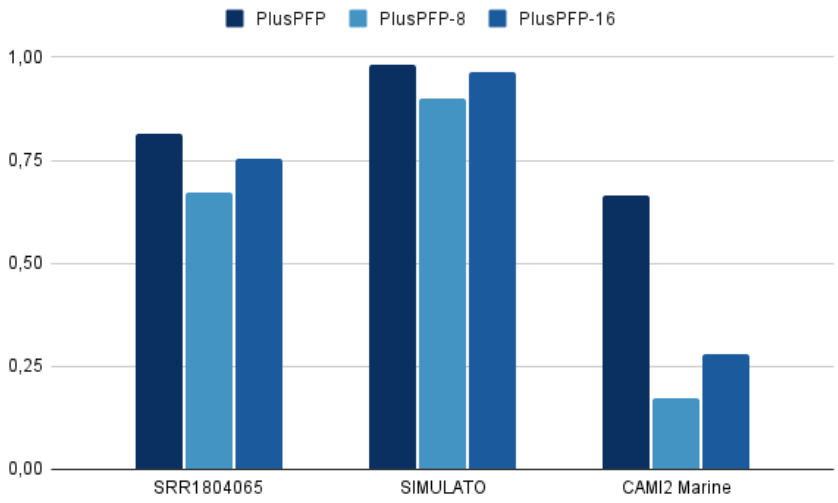

Figure 17: Precisione sui geni dei database PlusPFP a confronto

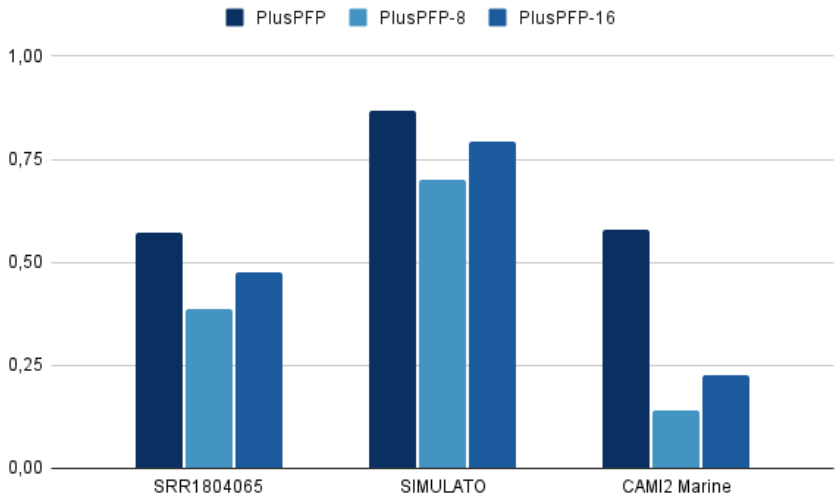

Figure 18: Precisione sulle specie dei database PlusPFP a confronto

4.5.4 PlusPFP (7/6/2022)

Table 39: Metriche ottenute a livello di geni - Database PlusPFP (7/6/2022)

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.723502	0.928639	0.813335
Simulato	0.973996	0.993839	0.983817
CAMI2 Marine	0.648172	0.828018	0.727140

Table 40: Metriche ottenute a livello di specie - Database PlusPFP (7/6/2022)

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.475149	0.747168	0.580890
Simulato	0.782202	0.983123	0.871229
CAMI2 Marine	0.543290	0.769105	0.636770

4.5.5 PlusPFP (27/1/2021)

Table 41: Metriche ottenute a livello di geni - Database PlusPFP (27/1/2021)

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.732837	0.930317	0.819853
Simulato	0.970320	0.997518	0.983731
CAMI2 Marine	0.642005	0.791250	0.708857

Table 42: Metriche ottenute a livello di specie - Database PlusPFP (27/1/2021)

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.519458	0.769638	0.620272
Simulato	0.783157	0.988346	0.873868
CAMI2 Marine	0.551692	0.742618	0.633073

4.5.6 PlusPFP (19/9/2020)

Table 43: Metriche ottenute a livello di geni - Database PlusPFP (19/9/2020)

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.752694	0.933510	0.833408
Simulato	0.969542	0.997845	0.983490
CAMI2 Marine	0.645251	0.797413	0.713308

Table 44: Metriche ottenute a livello di specie - Database PlusPFP (19/9/2020)

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.561803	0.799684	0.659962
Simulato	0.783630	0.988846	0.874358
CAMI2 Marine	0.554465	0.748774	0.637134

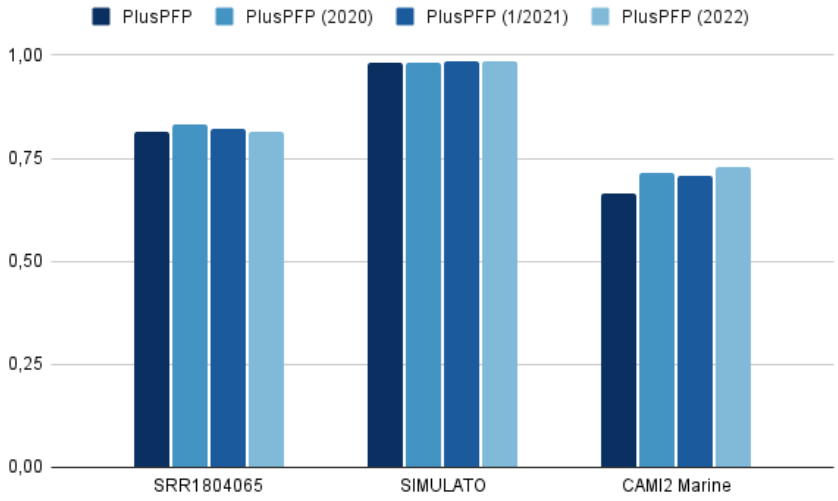

Figure 19: Database PlusPFP a confronto (livello di geni)

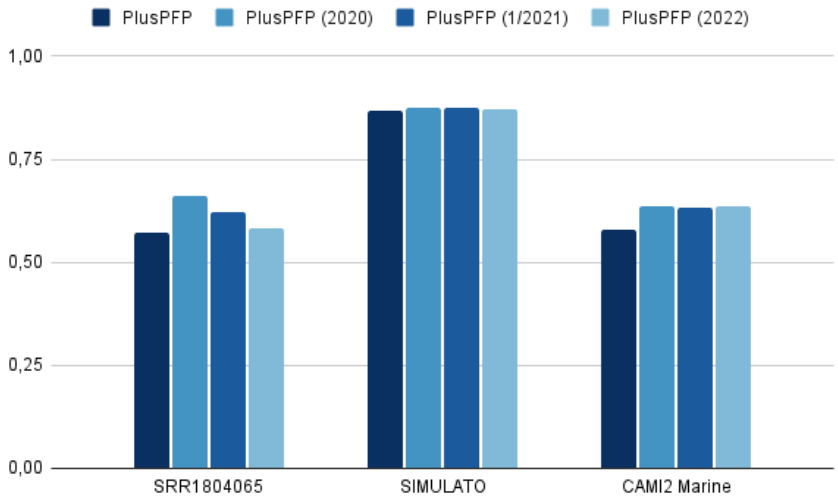

Figure 20: Database PlusPFP a confronto (livello di specie)

Il confronto dei database PlusPFP degli anni precedenti evidenzia nuovamente un peggioramento della F_1-measure per versione più recente. Il database PlusPFP datato 2020 ottiene i migliori risultati nella classificazione del dataset SRR1804065, sia a livello di geni che di specie. Per il dataset simulato i risultati sono molto simili per tutte le versioni, mentre per il dataset CAMI2 Marine brilla il database rilasciato il 7 Giugno 2022.

4.6 EuPathDB48

Il database EuPathDB48 contiene le sequenze genetiche di patogeni eucarioti senza contaminanti. La dimensione è pari a 34.1 GB. Vista la natura diametralmente diversa delle informazioni genetiche contenute in questo database rispetto ai dataset classificati, non sono state prodotte corrispondenze significative. Le metriche ottenute sono:

Table 45: Metriche ottenute a livello di geni - Database PlusPFP-16

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.000000	0.000111	0.000001
Simulato	0.000000	0.000000	0.000000
CAMI2 Marine	0.000003	0.000669	0.000005

Table 46: Metriche ottenute a livello di specie - Database PlusPFP-16

Dataset	Sensitività	Precisione	F_1
SRR1804065	0.000000	0.000048	0.000000
Simulato	0.000000	0.000000	0.000000
CAMI2 Marine	0.000003	0.000668	0.000005

5 Conclusione

La ricerca presenta un'analisi di precisione, sensitività e F_1-score di classificazioni tassonomiche, eseguite con il software Kraken2, affiancato da ventidue dei maggiori database realizzati per questo software.

A livello di geni, le prove sperimentali hanno evidenziato che la versione del database PlusPFP, rilasciata nel 2020, restituisce i migliori risultati in termini di F_1-score per il dataset SRR1804065. Il database Standard (1/2021) è il più performante per il dataset Simulato, mentre per il CAMI2 Marine è lo Standard (2020).

In merito all'indagine di specie, il database Standard (2020) ottiene i risultati assoluti migliori per ognuno dei 3 dataset classificati.

Relativamente ai database sopracitati con tetto di 8GB e 16GB, si ottengono risultati comparabili alle versioni di dimensione originale, sia per il dataset SRR1804065 che per il simulato. Tale trend non viene riconfermato nel CAMI2 Marine, contenente oltre 4000 specie, in cui si possono apprezzare le maggiori differenze tra le varie versioni dei database, favorendo quelle di dimensione maggiore.

Le prove sperimentali infine, hanno evidenziato un leggero peggioramento dell'F_1-measure nelle versioni più recenti dei database Standard, PlusPF e PlusPFP, favorendo di molto le *release* più datate, in particolare quelle rilasciate nel 2020.

Bibliografia

[1] National Research Council. *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. The National Academies Press, Washington, DC, 2007. ISBN 978-0-309-10676-4. doi: 10.17226/11902. URL https://nap.nationalacademies.org/catalog/11902/the-new-science-of-metagenomics-revealing-the-secrets-of-our.

[2] Wood D.E., Lu J., and Langmead B. Improved metagenomic analysis with kraken 2. *Genome Biol*, 2019. doi: https://doi.org/10.1186/s13059-019-1891-0. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1891-0#citeas.

[3] Michael Roberts, Wayne Hayes, Brian R. Hunt, Stephen M. Mount, and James A. Yorke. Reducing storage requirements for biological sequence comparison. *Bioinformatics*, 20(18):3363–3369, 07 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth408. URL https://doi.org/10.1093/bioinformatics/bth408.

[4] Davide Storato and Matteo Comin. K2mem: Discovering discriminative k-mers from sequencing data for metagenomic reads classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(1):220–229, 2022. doi: 10.1109/TCBB.2021.3117406.

Indice delle immagini

Indice delle tabelle