



Metodi AF e Mash

Strumenti Formali per la Bioinformatica - A.A. 2023/24

Alessandro Ricchetti
Marco Cappiello

Overview

1

Limiti metodi basati su
allineamento

3

Classificazione metodi

5

Mash (sketch, dist, screen)

2

Confronto tra metodi basati su
allineamento e metodi AF

4

Benchmark

6

Esecuzione



Limiti metodi basati su allineamento

01

Complessità computazionale

Il calcolo di un allineamento accurato di sequenze multiple è un problema NP-hard.

03

Dimensioni delle sequenze

Allineare due sequenze di DNA lunghe milioni di nucleotidi non è pratico.

02

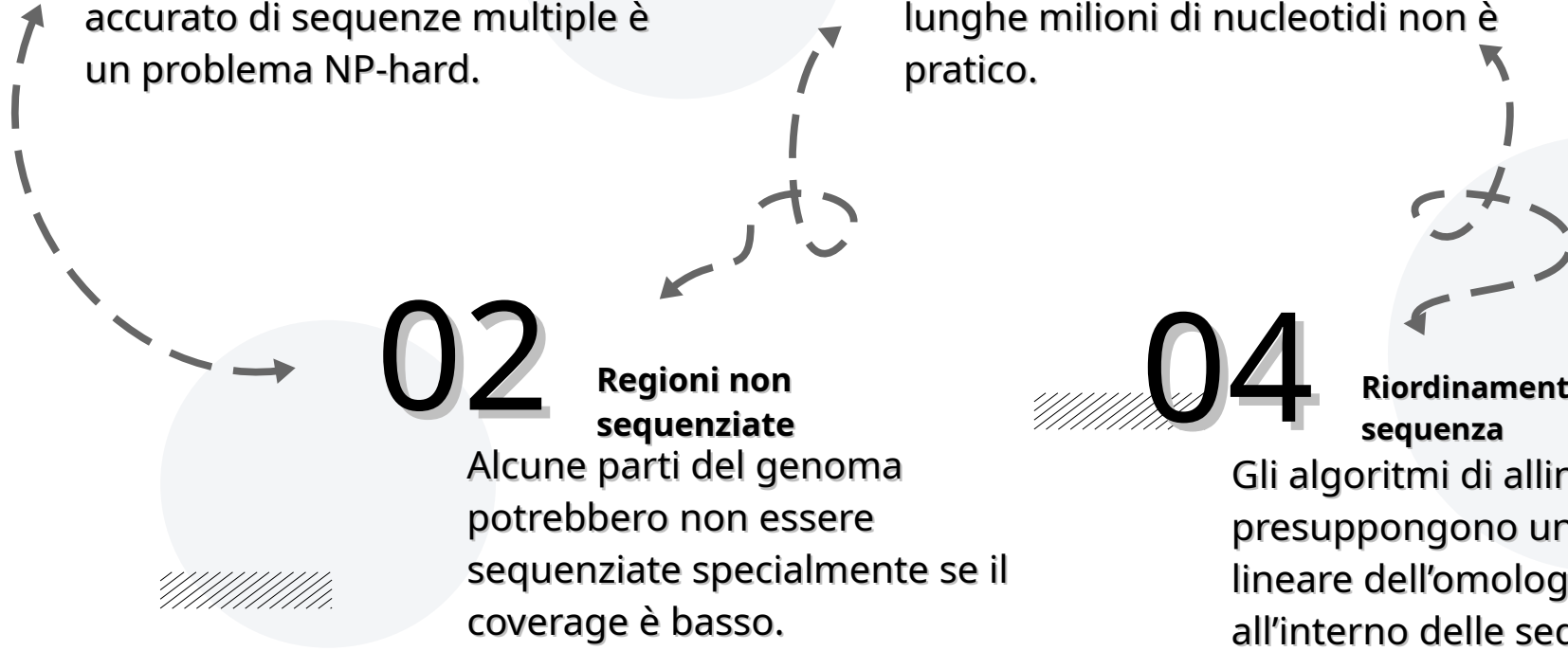
Regioni non sequenziate

Alcune parti del genoma potrebbero non essere sequenziate specialmente se il coverage è basso.

04

Riordinamenti di sequenza

Gli algoritmi di allineamento presuppongono un ordine lineare dell'omologia conservato all'interno delle sequenze.



Alignment-based vs Alignment-free

- Questi metodi presuppongono che le regioni omologhe siano contigue (gap)
- Calcola tutti i possibili confronti a coppie di sequenze
- Richiede modelli di sostituzione/evoluzione
- Utilizza algoritmi di inferenza con complessità di almeno $O(n^2)$; meno efficienti in termini di tempo



- Non presuppone la contiguità delle regioni omologhe
- Basato su occorrenze di sotto-sequenze
- Non è necessario l'utilizzo di modelli evolutivi
- Algoritmi di inferenza tipicamente $O(n^2)$ o meno

Classificazione metodi - 1/3

**Exact k-mer
count**

AAF

AFKS

alfpy

FFP

CAFÈ

jD2Stat

**Inexact k-mer
count**

Spaced


**Teoria
dell'informazione**

LZW-
Kernel

**k-mer
count**

kWIP

Classificazione metodi - 2/3



**Maximal length
of exact
common
substring**

ALFRED-G

kr

kmacs

Underlying
Approach



**micro-
alignments**

andi

co-phylog

FS
WM

Multi-
SpaM

phylonium

**SNP
count**

kSNP3



Classificazione metodi - 3/3

Number of word matches

mash

Skmer

Slope-
SpaM



**Variable length-
word count**

EP-sim



**Return time
distribution**

RTD-
Phylogeny



Benchmark - 1/3

Protein classification			
High sequence		Low sequence	
AKFS	0.798	AKFS	0.742
alfpy	0.778	alfpy	0.739
lzw-ncd	0.760	EP-sim	0.705

Gene Trees	
jD2Star	0.3296
AFKS	0.3414
alfpy	0.3426

Regulatory Sequences	
alfpy	0.736
rtd-phylogeny	0.614
café	0.601

Benchmark - 2/3

Genome-based phylogeny					
Assembled					
Mitochondria		E.coli		Plants	
mash	0.05	phylonium	0.04	mash	0.09
FSWM	0.05	andi	0.08	co-phylog	0.09
spaced	0.05	co-phylog	0.12	Multi-SpaM	0.09
Unassembled					
E.coli			Plants		
andi	0.21		mash	0.14	
co-phylog	0.21		Skmer	0.25	
mash	0.27		co-phylog	0.27	

Benchmark - 3/3

Horizontal Gene Transfer					
Simulated genomes		E.coli/Shigella		Yersinia	
mash	0.05	andi	0.08	Skmer	0.00
AFKS	0.05	mash	0.12	AFKS	0.00
cafe	0.05	FSWM	0.17	alfpy	0.00

Mash



Sketch

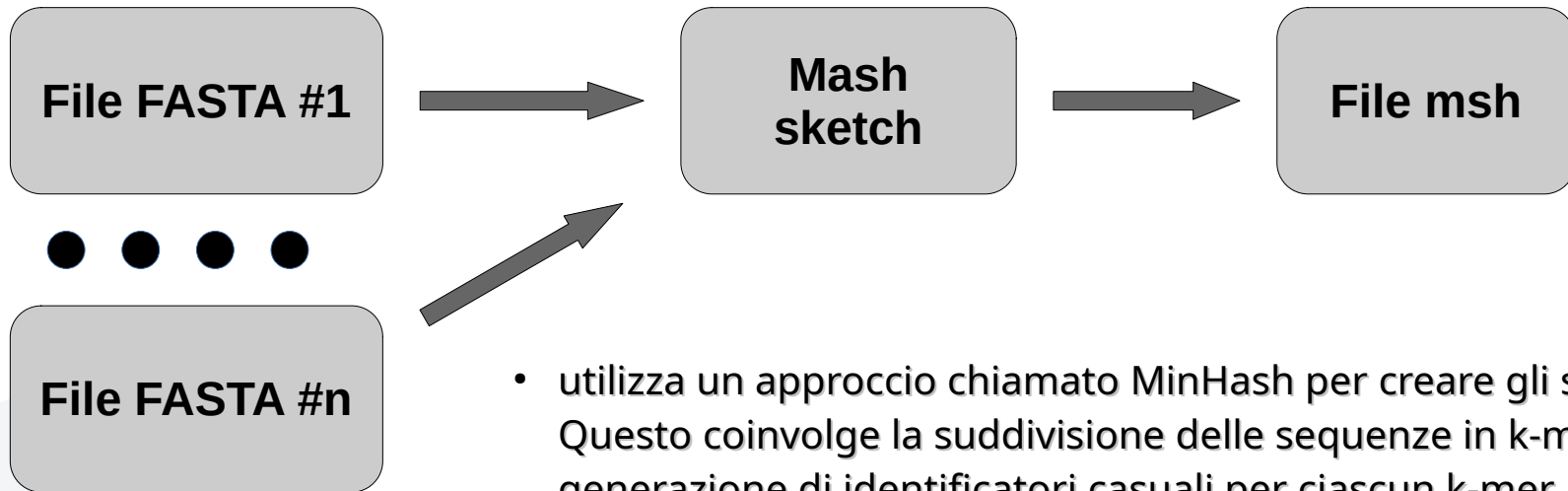


Dist



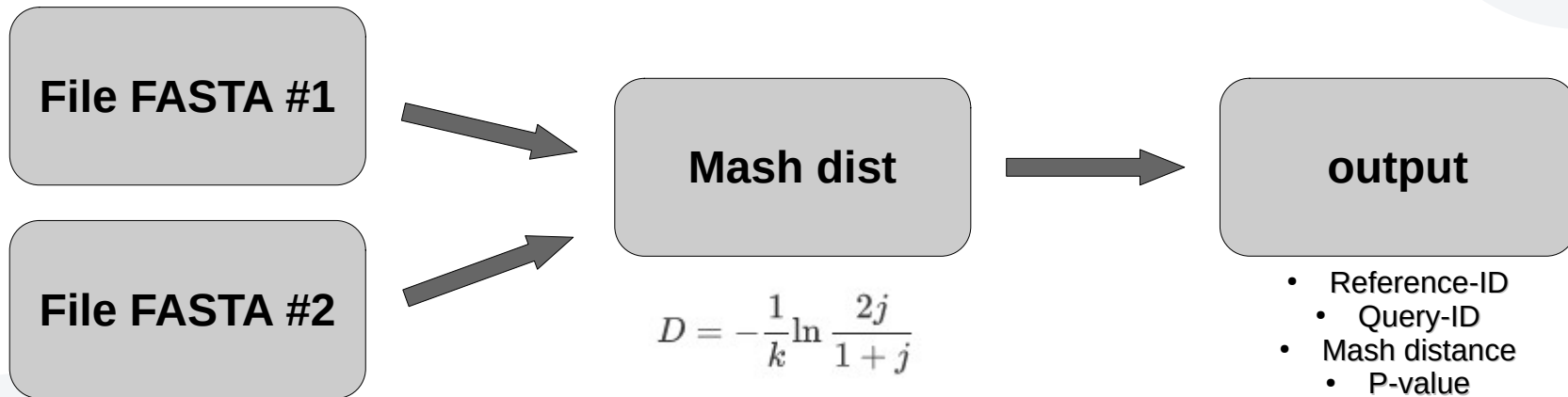
Screen

Mash sketch

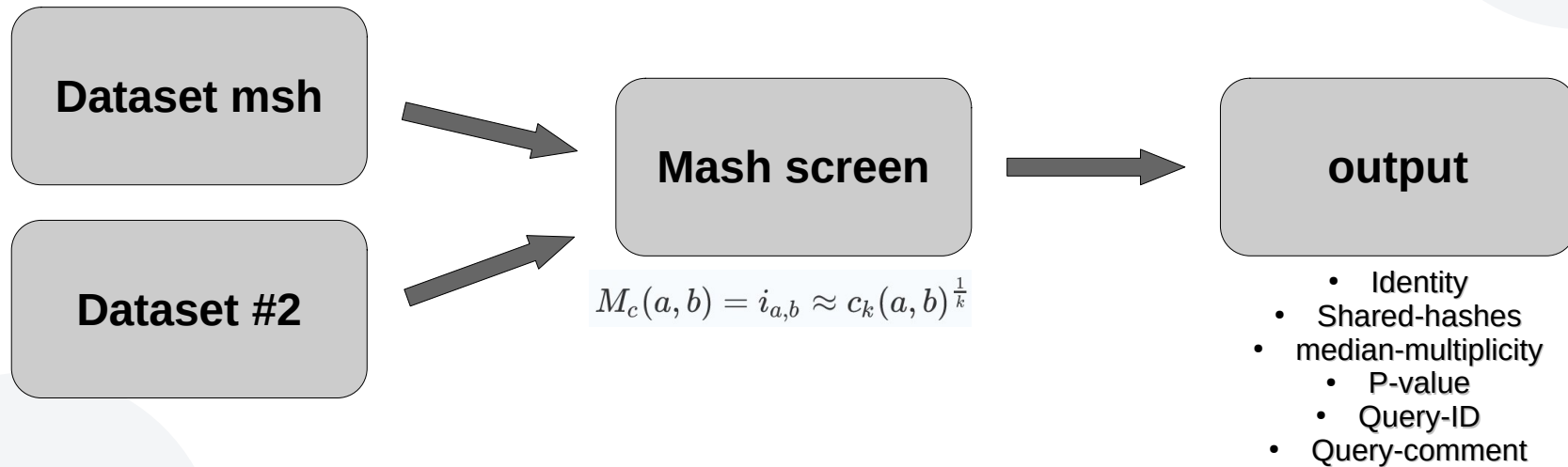


- utilizza un approccio chiamato MinHash per creare gli sketch. Questo coinvolge la suddivisione delle sequenze in k-mer e la generazione di identificatori casuali per ciascun k-mer.
- calcola gli sketch in modo efficiente, con un tempo di calcolo approssimativo di $O(n \log s)$ per uno sketch di dimensioni s su un genoma di dimensioni n .

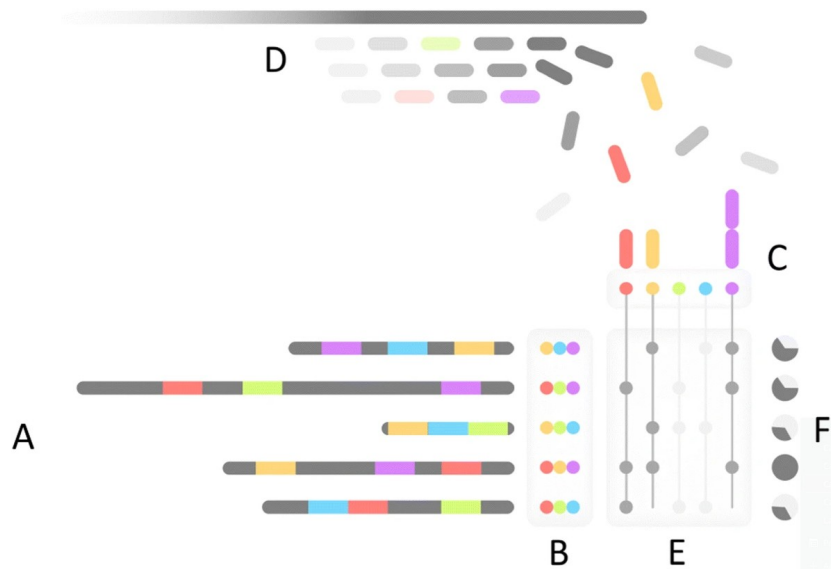
Mash dist



Mash screen - 1/2



Mash screen - 2/2



(A) sequenza di riferimento

(B) hash più piccoli

(C) hash degli sketch usati come chiavi per una map
contenente conteggio per ciascun hash

(D) hashing dei k-mer

(E) conteggi della mappa vengono interrogati per ogni
sketch

(F) stima del contenimento per ogni
costituente

Esecuzione - 1/3

Si confrontano due genomi di Escherichia coli con mash dist:
> gi|49175990|ref |NC000913.2|Escherichiacolistr.K-12substr.M G1655
> gi|47118301|dbj|BA000007.2|EscherichiacoliO157:H7str.SakaiDNA

./mash dist genome1.fna genome2.fna

Reference-ID	Query-ID	Mash distance	P-value	Hash corrispondenti
genome1.fna	genome2.fna	0.0222766	0	456/1000

Esecuzione - 2/3

Si utilizza un altro genoma di E.coli:

> *gi|682117612|gb|CP009273.1|EscherichiacoliBW25113*

Si effettua prima lo sketch dei due genomi in modo tale da creare uno sketch combinato:

```
./mash sketch -o reference genome1.fna genome2.fna
```

Poi si stima la distanza da ciascuna query con il terzo genoma:

```
./mash dist reference.msh genome3.fna
```

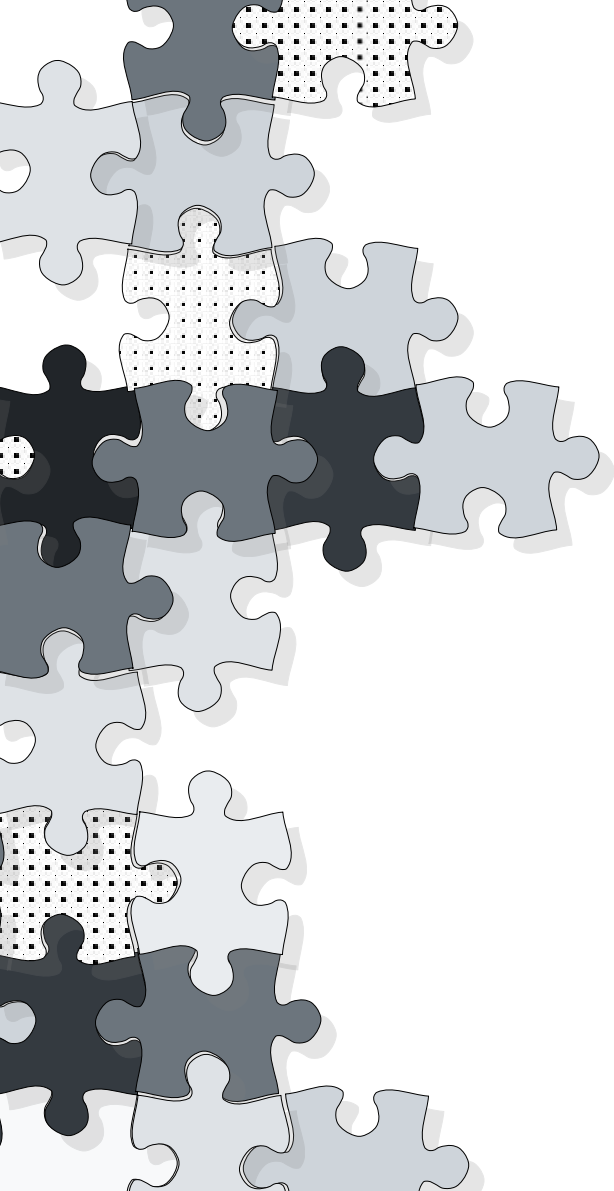
Reference-ID	Query-ID	Mash distance	P-value	Hash corrispondenti
genome1.fna	genome3.fna	0	0	1000/1000
genome2.fna	genome3.fna	0.0222766	0	456/1000

Esecuzione - 3/3

Si può analizzare il dataset scelto andando a rimuovere la ridondanza con -w con i genomi RefSeq:

```
./mash screen -w -p 12 refseq.genomes.k21s1000.msh ERR024951.fastq
```

identity	shared-hashes	median-multiplicity	P-value	query-ID
0.99957	991/1000	24	0	GCF 002054545.1 ASM205454v1 genomic.fna.gz
0.99899	979/1000	26	0	GCF 000841985.1 ViralProj14228 genomic.fna.gz
0.998844	976/1000	101	0	GCF 900086185.1 12082 4 85 genomic.fna.gz
0.923964	190/1000	49	0	GCF 000900935.1 ViralProj181984 genomic.fna.gz
0.900615	111/1000	100	0	GCF 001876675.1 ASM187667v1 genomic.fna.gz
0.887722	82/1000	31	3.16322e-233	GCF 001470135.1 ViralProj306294 genomic.fna.gz
0.873204	58/1000	22	1.8212e-156	GCF 000913735.1 ViralProj227000 genomic.fna.gz
0.868675	52/1000	57	6.26251e-138	GCF 001744215.1 ViralProj344312 genomic.fna.gz
0.862715	45/1000	1	1.05185e-116	GCF 001882095.1 ViralProj353688 genomic.fna.gz
0.856856	39/1000	21	6.70643e-99	GCF 000841165.1 ViralProj14230 genomic.fna.gz



“

Grazie per l'attenzione

”

