Encryption and compression for collections of text files*

Dott. Ferdinando Montecuollo
Università degli Studi della Campania
«Luigi Vanvitelli»

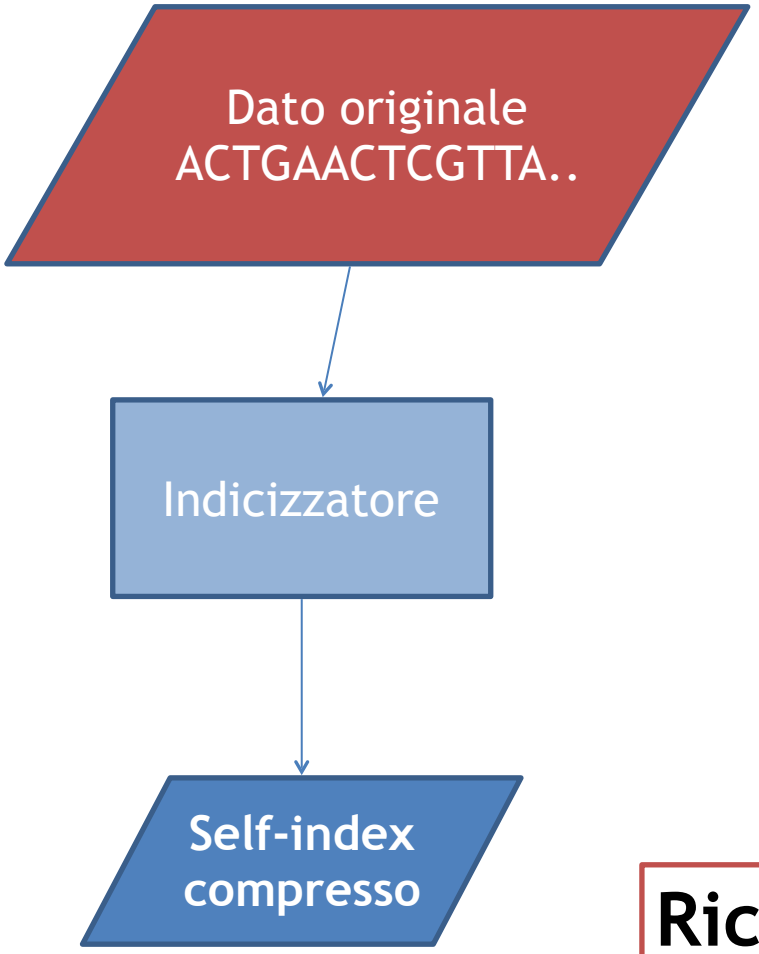
*F. Montecuollo, G. Schmid and R. Tagliaferri, *E²FM: an encrypted and compressed full-text index for collections of genomic sequences*, Bioinformatics, 33(18), 2017

Gestione di dati genomici: una sfida aperta per la Bioinformatica

Gestione di dati genomici: una sfida aperta per la Bioinformatica

- Piattaforme NGS e tecnologie high-throughput
  - Nuove sfide nel campo della memorizzazione e del trattamento dei dati genomici
    - Enorme mole di informazioni
    - L'informazione genetica è altamente *sensibile*
  - Necessità di nuovi algoritmi e sistemi che coniughino:
    - basso costo di memorizzazione;
    - efficienza nelle ricerche;
    - **confidenzialità**.

Compressione e indicizzazione



Sostituisce il dato originario

Ridotta occupazione di m

Rapporto di compressione ~=30/100=0.30

Ricerche efficienti

Un esempio: FM-Index

Ricerche su dati compressi
Non garantisce la confidenzialità!

Confidenzialità

Confidenzialità

- Obiettivo
  - Impedire l’accesso ai dati da parte di soggetti non autorizzati

Confidenzialità

- Obiettivo
  - Impedire l'accesso ai dati da parte di soggetti non autorizzati
- L'informazione genetica
  - potrebbe consentire di risalire all'identità dell'individuo anche quando non associata a dati anagrafici
  - sfruttabile per fini discriminatori e non legittimi
    - Una compagnia di assicurazioni potrebbe negare la stipula di una polizza sulla salute in base alla predisposizione a sviluppare alcuni tipi di malattie
    - Un'azienda potrebbe non assumere un lavoratore desumendo dal suo DNA un'aggressività superiore alla media
- Necessarie:
  - Nuove normative
  - Uso di **tecniche di cifratura**

La nostra attività di ricerca

La nostra attività di ricerca

- Obiettivo
  - Nuove tecniche **integrate** di compressione e cifratura del genoma
    - Confidenzialità
    - Basso costo di memorizzazione
    - Ricerche efficienti sui dati compressi e cifrati

La nostra attività di ricerca

- Obiettivo
  - Nuove tecniche **integrate** di compressione e cifratura del genoma
    - Confidenzialità
    - Basso costo di memorizzazione
    - Ricerche efficienti sui dati compressi e cifrati
- Risultato
  - Un nuovo self-index per la memorizzazione cifrata di collezioni genomiche
    - E^2FM

L’approccio scelto

L’approccio scelto

- Trasformata di Burrows e Wheeler (BWT)
  - Data una sequenza di caratteri, ne produce una permutazione reversibile che ne migliora la comprimibilità
  - Risultato fortemente dipendente dall’ordinamento

L’approccio scelto

- Trasformata di Burrows e Wheeler (BWT)
  - Data una sequenza di caratteri, ne produce una permutazione reversibile che ne migliora la comprimibilità
  - Risultato fortemente dipendente dall’ordinamento

GATTACA

L’approccio scelto

- Trasformata di Burrows e Wheeler (BWT)
  - Data una sequenza di caratteri, ne produce una permutazione reversibile che ne migliora la comprimibilità
  - Risultato fortemente dipendente dall’ordinamento

GATTACA	A < C < G < T	T < A < G < C

L’approccio scelto

- Trasformata di Burrows e Wheeler (BWT)
  - Data una sequenza di caratteri, ne produce una permutazione reversibile che ne migliora la comprimibilità
  - Risultato fortemente dipendente dall’ordinamento

GATTACA	A < C < G < T	T < A < G < C

```
G A T T A C A
A T T A C A G
T T A C A G A
T A C A G A T
A C A G A T T
C A G A T T A
A G A T T A C
```

L’approccio scelto

- Trasformata di Burrows e Wheeler (BWT)
  - Data una sequenza di caratteri, ne produce una permutazione reversibile che ne migliora la comprimibilità
  - Risultato fortemente dipendente dall’ordinamento

GATTACA	A < C < G < T	T < A < G < C

G A T T A C A	A C A G A T T
A T T A C A G	A G A T T A C
T T A C A G A	A T T A C A G
T A C A G A T	C A G A T T A
A C A G A T T	G A T T A C A
C A G A T T A	T A C A G A T
A G A T T A C	T T A C A G A

L’approccio scelto

- Trasformata di Burrows e Wheeler (BWT)
  - Data una sequenza di caratteri, ne produce una permutazione reversibile che ne migliora la comprimibilità
  - Risultato fortemente dipendente dall’ordinamento

GATTACA	A < C < G < T	T < A < G < C

```
G A T T A C A        A C A G A T T
A T T A C A G        A G A T T A C
T T A C A G A        A T T A C A G
T A C A G A T        C A G A T T A
A C A G A T T        G A T T A C A
C A G A T T A        T A C A G A T
A G A T T A C        T T A C A G A
```

L’approccio scelto

- Trasformata di Burrows e Wheeler (BWT)
  - Data una sequenza di caratteri, ne produce una permutazione reversibile che ne migliora la comprimibilità
  - Risultato fortemente dipendente dall’ordinamento

GATTACA	A < C < G < T	T < A < G < C
G A T T A C A	A C A G A T T	T T A C A G A
A T T A C A G	A G A T T A C	T A C A G A T
T T A C A G A	A T T A C A G	A T T A C A G
T A C A G A T	C A G A T T A	A G A T T A C
A C A G A T T	G A T T A C A	A C A G A T T
C A G A T T A	T A C A G A T	G A T T A C A
A G A T T A C	T T A C A G A	C A G A T T A

L’approccio scelto

- Trasformata di Burrows e Wheeler (BWT)
  - Data una sequenza di caratteri, ne produce una permutazione reversibile che ne migliora la comprimibilità
  - Risultato fortemente dipendente dall’ordinamento

GATTACA	A < C < G < T	T < A < G < C
G A T T A C A	A C A G A T T	T T A C A G A
A T T A C A G	A G A T T A C	T A C A G A T
T T A C A G A	A T T A C A G	A T T A C A G
T A C A G A T	C A G A T T A	A G A T T A C
A C A G A T T	G A T T A C A	A C A G A T T
C A G A T T A	T A C A G A T	G A T T A C A
A G A T T A C	T T A C A G A	C A G A T T A

Scrambled BWT

Külekci, On scrambling the Burrows-Wheeler transform to provide privacy in lossless compression, Computer Security, 2012

Scrambled BWT

Külekci, On scrambling the Burrows-Wheeler transform to provide privacy in lossless compression, Computer Security, 2012

- Idea
  - Costruire un crittosistema simmetrico la cui chiave di cifratura corrisponda ad uno specifico ordinamento (permutazione) dei simboli

Scrambled BWT

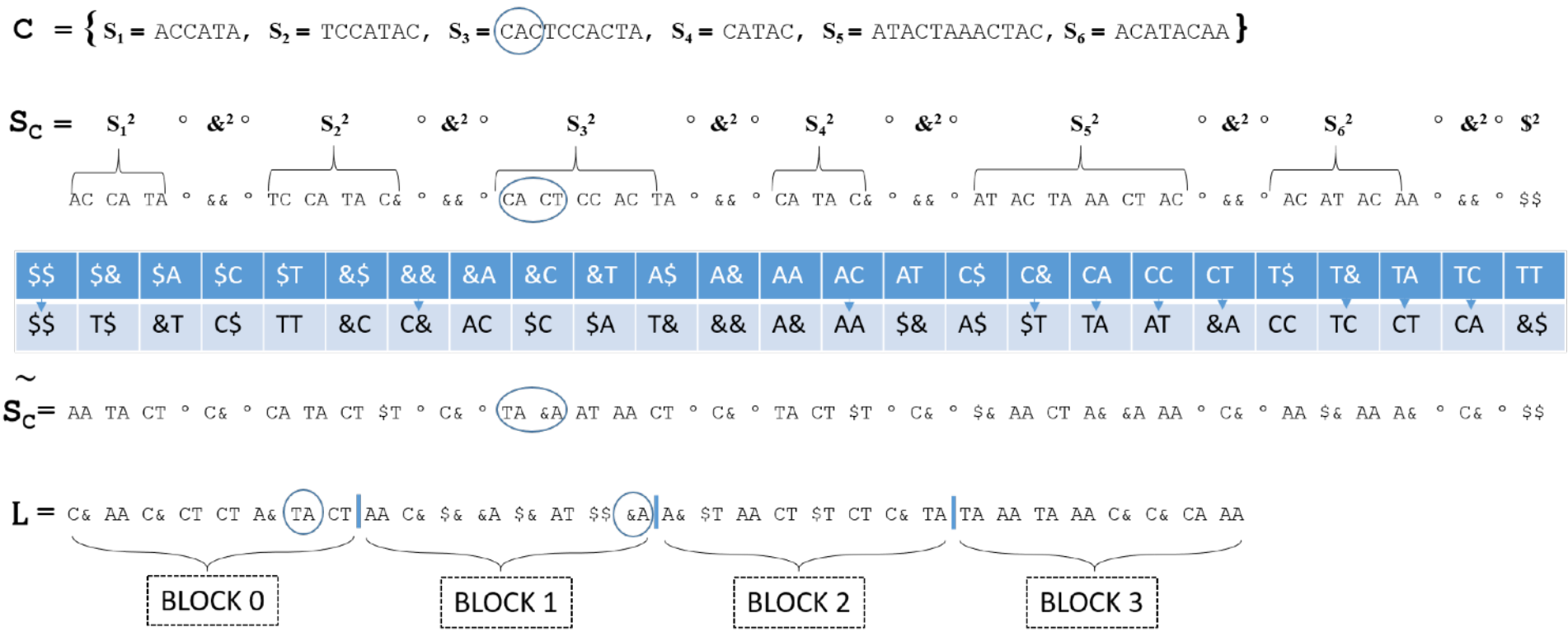
Külekci, On scrambling the Burrows-Wheeler transform to provide privacy in lossless compression, Computer Security, 2012

- Idea
  - Costruire un crittosistema simmetrico la cui chiave di cifratura corrisponda ad uno specifico ordinamento (permutazione) dei simboli
- Per le sequenze nucleotidiche:
  - alfabeto di 4 caratteri
  - 4! possibili ordinamenti (un ordinamento per ogni possibile permutazione dei 4 simboli)
  - Problema: crittosistema facilmente attaccabile
    - La chiave avrebbe solo 4!=24 possibili valori
    - per decrittare indebitamente una sequenza mediante un **attacco di forza bruta** basterebbe fare al più 24 tentativi
  - Soluzione: **estensione dell'alfabeto**
    - raggruppamento dei caratteri *k a k*

Sicurezza *In memory*

- Criticità per le sequenze di DNA
  - Forti regolarità
  - Disponibilità di testo in chiaro
- Grado di omofonia cresce in funzione di *k*
  - *Numero dei possibili tentativi che l'attaccante deve fare nel caso peggiore per un CPA*
  - *Dell'ordine di 10^{22} per k=4*
  - *Dell'ordine di 10^{100} per k=5,6,7,8*
- Scrambling globale
  - distribuito sull'intera **collezione di sequenze**
  - percentuale di crittogramma in memoria molto bassa

Esempio


C = { S1 = ACCATA, S2 = TCCATAC, S3 = CACTCCACTA, S4 = CATAC, S5 = ATACTAAACTAC, S6 = ACATACAA }
SC = S1² ° &² ° S2² ° &² ° S3² ° &² ° S4² ° &² ° S5² ° &² ° S6² ° &² ° $²
AC CA TA ° && ° TC CA TA C& ° && ° CA CT CC AC TA ° && ° CA TA C& ° && ° AT AC TA AA CT AC ° && ° AC AT AC AA ° && ° $$
$$ $& $A $C $T &$ && &A &C &T A$ A& AA AC AT C$ C& CA CC CT T$ T& TA TC TT
$$ T$ &T C$ TT &C C& AC $C $A T& && A& AA $& A$ $T TA AT &A CC TC CT CA &$
~
SC= AA TA CT ° C& ° CA TA CT $T ° C& ° TA &A AT AA CT ° C& ° TA CT $T ° C& ° $& AA CT A& &A AA ° C& ° AA $& AA A& ° C& ° $$
L = C& AA C& CT CT A& TA CT | AA C& $& &A $& AT $$ &A | A& $T AA CT $T CT C& TA | TA AA TA AA C& C& CA AA
BLOCK 0
BLOCK 1
BLOCK 2
BLOCK 3

Cifratura dei blocchi

- Metodo di cifratura dei blocchi
  - Integra la Scrambled BWT
  - Sfrutta un cifrario a flusso ottimizzato
    - Basato su Salsa20
      - Pubblicato nel 2005
      - Non esiste ad oggi un attacco significativo noto
      - Estremamente efficiente anche su PC
    - PRNG usato anche per la Scrambled BWT
      - Permutare i codici dell'alfabeto dei k-meri richiede un generatore pseudo-casuale

Il software

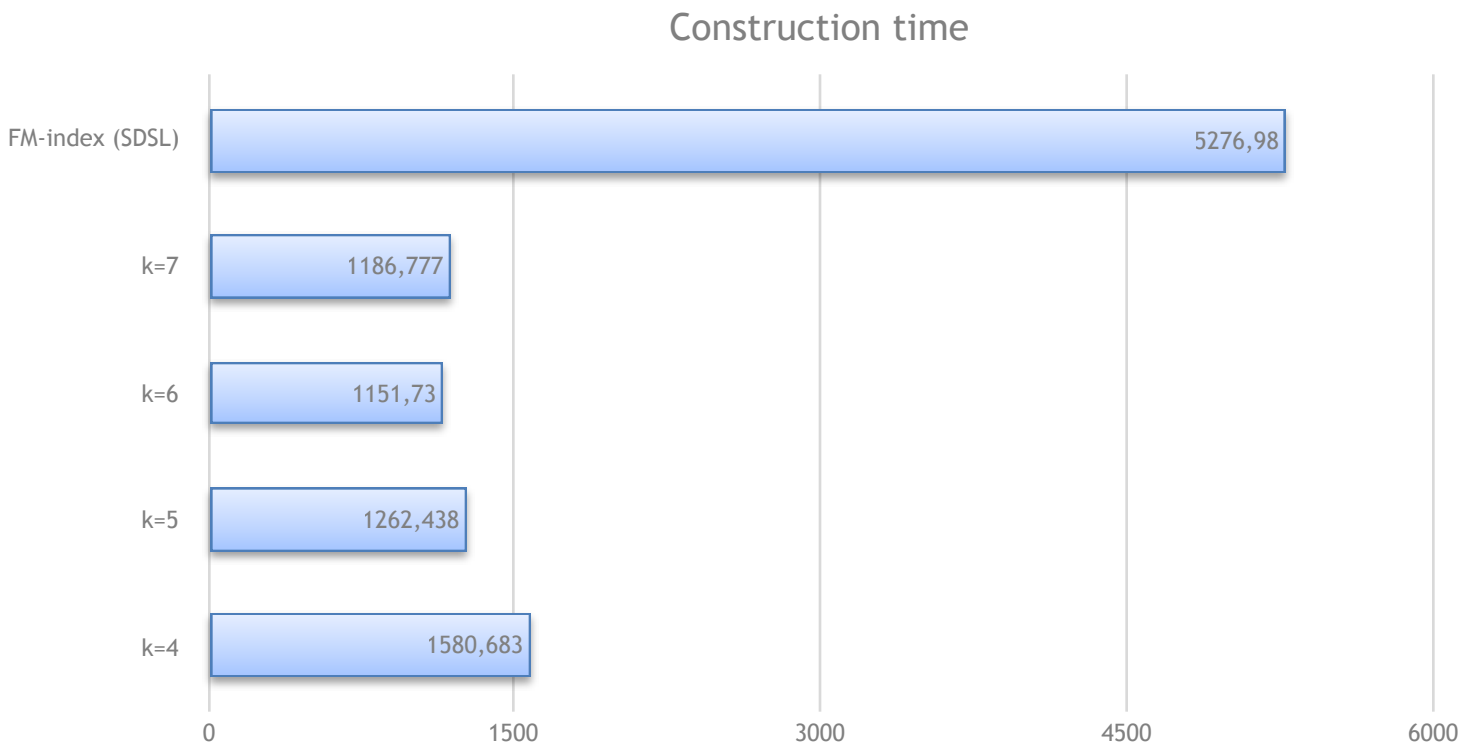
- Codice sorgente disponibile su GitHub
  - https://github.com/montecuollo/E2FM
- Sviluppato in C++
- Input: Collezioni di sequenze in formato FASTA
- Operazioni supportate
  - Costruzioni degli indici
  - Ricerca esatta di pattern
  - Estrazione di sequenze/sottosequenze

Prestazioni

Prestazioni

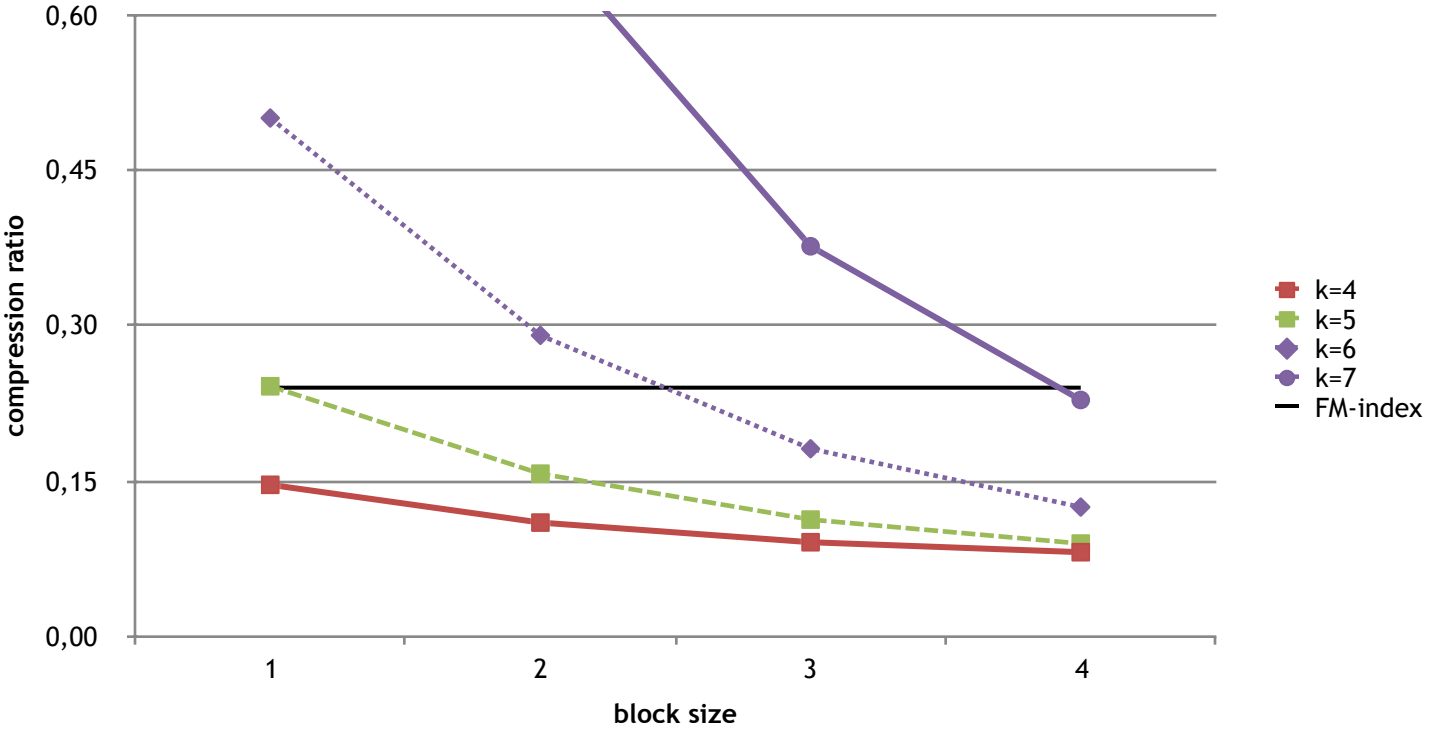
- Diversi tipi di test:
  - Collezioni di cromosomi di individui del progetto 1000 genomi
  - Collezioni di sequenze cromosomiche generate in modo pseudo-casuale a partire dal genoma di riferimento
    - tasso di mutazione: 0.1%
    - tasso di inserzioni e delezioni: 0.013%
    - lunghezza delle inserzioni e delle delezioni: 1÷16
- Risultati
  - Costruzione degli indici molto rapida
  - Ottimi rapporti di compressione (risparmio di spazio fino al 95%)
  - Tempo per la ricerca esatta di pattern: [*msec*]÷[100*msec*]
- Tool di riferimento: *Sdsl library*

Costruzione dell'indice



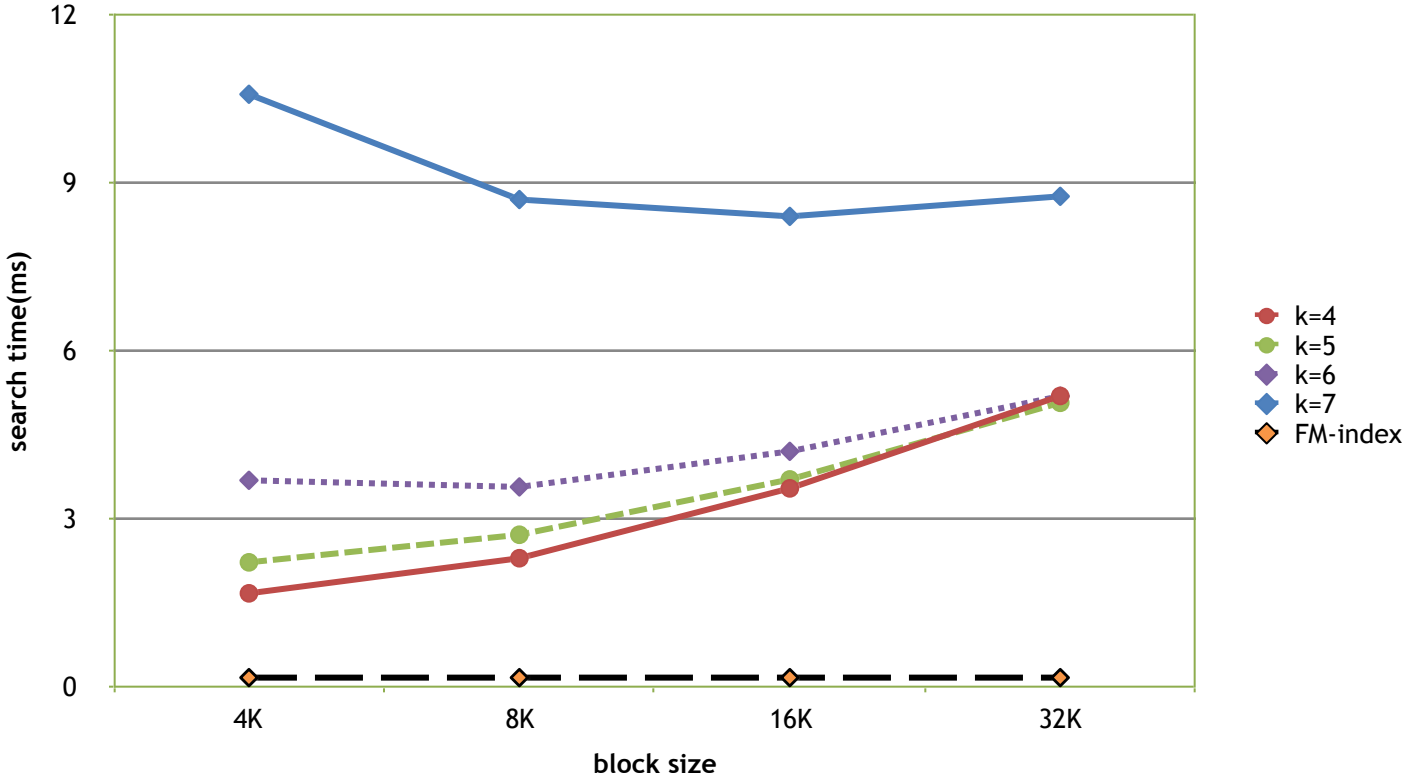
Confronto tra il tempo di costruzione (in secondi) di un indice E2FM e di un FM-index implementato con la libreria Sdsl per una ***collezione di 50 cromosomi 11 del progetto 1000 genomi****.*

Rapporto di compressione



Confronto tra il rapporto di compressione di un indice E2FM e di un FM-index implementato con la libreria Sdsl

Tempo di ricerca



Tempi di ricerca medi (in millisecondi) di pattern di lunghezza variabile tra 20 e 500 basi

Conclusioni e prospettive

Conclusioni e prospettive

- Risultati in linea con le aspettative
  - **La cifratura non inficia le performance**
  - Ottimi rapporti di compressione
  - Tempi variabili tra qualche millisecondo e qualche centinaio di millisecondi per eseguire ricerche di pattern in memoria di massa

Conclusioni e prospettive

- Risultati in linea con le aspettative
  - **La cifratura non inficia le performance**
  - Ottimi rapporti di compressione
  - Tempi variabili tra qualche millisecondo e qualche centinaio di millisecondi per eseguire ricerche di pattern in memoria di massa
- Potenziali futuri sviluppi
  - Realizzazione di software per la gestione di DBMS commerciali
  - Implementazione di algoritmi per la compressione cifrata di dati genomici ad alto grado di similarità con un modello *role-based* di controllo degli accessi
  - Implementazione di algoritmi di allineamento e ricerca inesatta
  - Studi e sperimentazioni per la resilienza agli attacchi con computer quantistici