UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

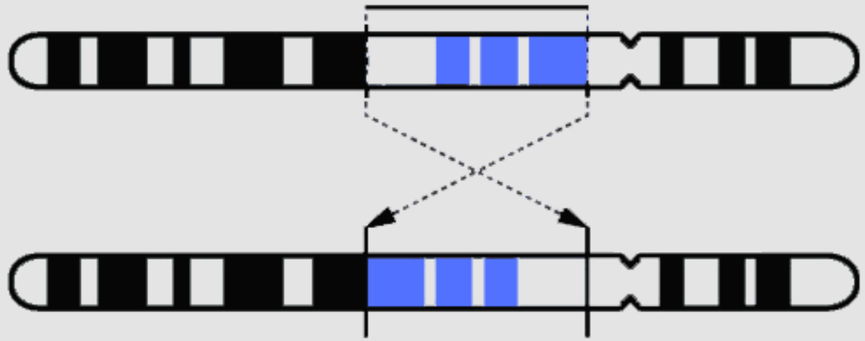
Detection of Genomic Inversions Using Sample-Specific Strings

Silvia Cambiago
Matr. 879382

Relatore: Prof.ssa Paola Bonizzoni
Correlatore: Dott. Luca Denti

Inversioni

- Variazioni strutturali
- Inversione e complemento basi
  - A ⇔ T, C ⇔ G
- Generate da Non-Allelic Homologous Recombination



R: AACTTTGGCTAAGACCCGAAACCA

T: AACTTTCGGGTCTTAGCCAAACCA

Motivazione

PROBLEMA

Identificare le inversioni

OBIETTIVO

Proporre un algoritmo di pattern matching che le rilevi con precisione

IMPORTANZA

Sono associate a diverse patologie e coinvolte nell’evoluzione

Idea di base

Article | Published: 22 December 2022

SVDSS: structural variation discovery in hard-to-call genomic regions using sample-specific strings from accurate long reads

Luca Denti, Parsoa Khorsand, Paola Bonizzoni ✉, Fereydoun Hormozdiari ✉ & Rayan Chikhi ✉

Nature Methods **20**, 550–558 (2023) | Cite this article

Problema

IDENTIFICAZIONE DELLE INVERSIONI

INPUT

Genoma R
Target T
Collezione di stringhe specifiche (SFS) per T

OUTPUT

Collezione di segmenti T[i, j] invertiti rispetto al genoma di riferimento R

Approccio

R: CCATGATTCGCATGCACTTG

T: CCATTGCATGCGAATCCTTG

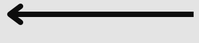
SFSs: [CATT, CCT]

GCATGCGAAT

ATTCGCATGC

ATTCGCATGC

R: CCATGATTCGCATGCACTTG



R: CCATGATTCGCATGCACTTG

T: CCATTGCATGCGAATCCTTG

Utilizzando KMP (Knuth-Morris-Pratt)

Complessità

n: numero SFSs

r: lunghezza genoma

$\overline{m}$: lunghezza media segmenti invertiti

TEMPO

$\Theta(r + n \cdot \overline{m})$

SPAZIO

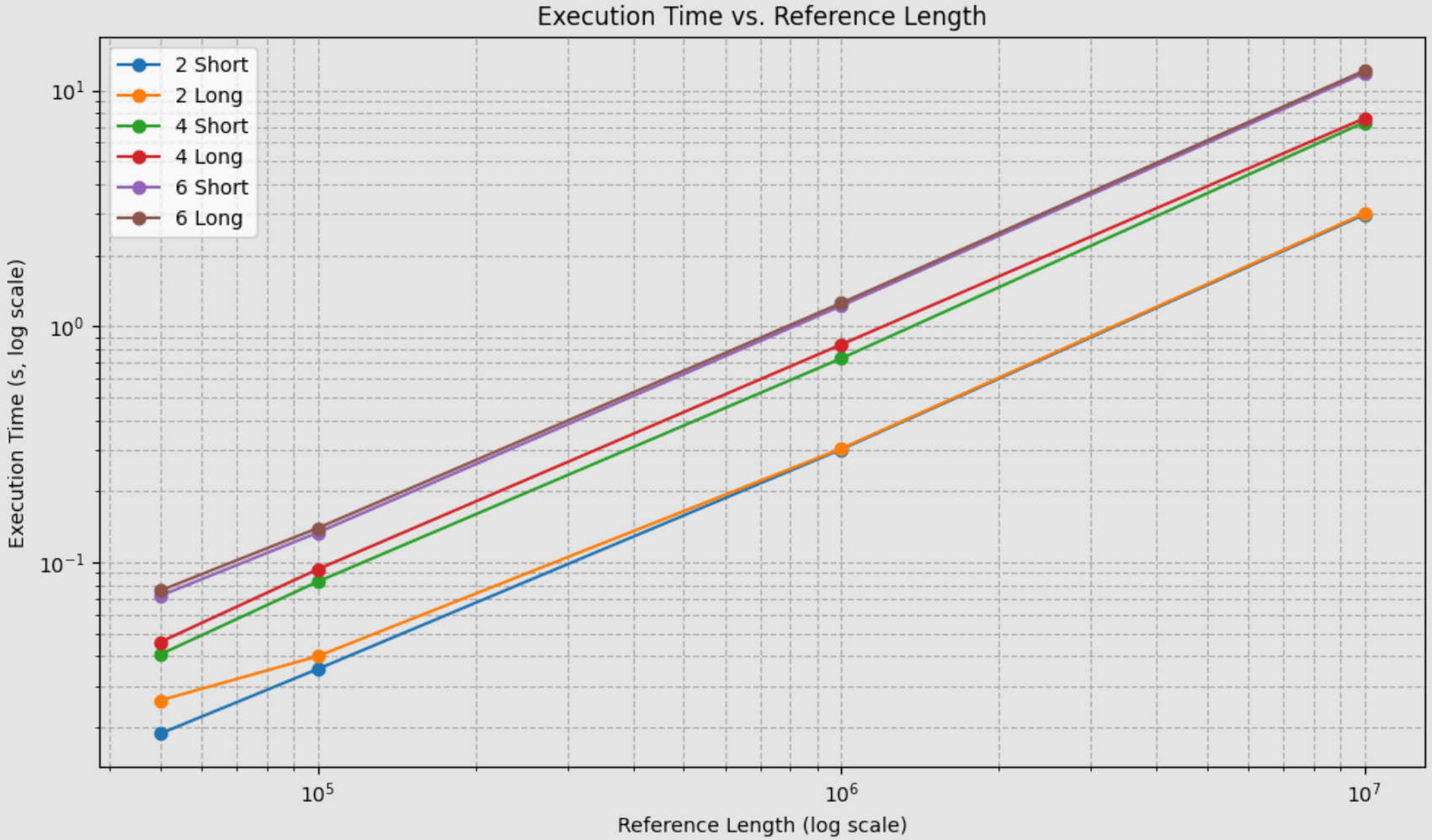
$\Theta(n \cdot \overline{m})$

Risultati

- 2 lunghezze
  - Corte: ~ 250 bp
  - Lunghe: ~ 6500 bp

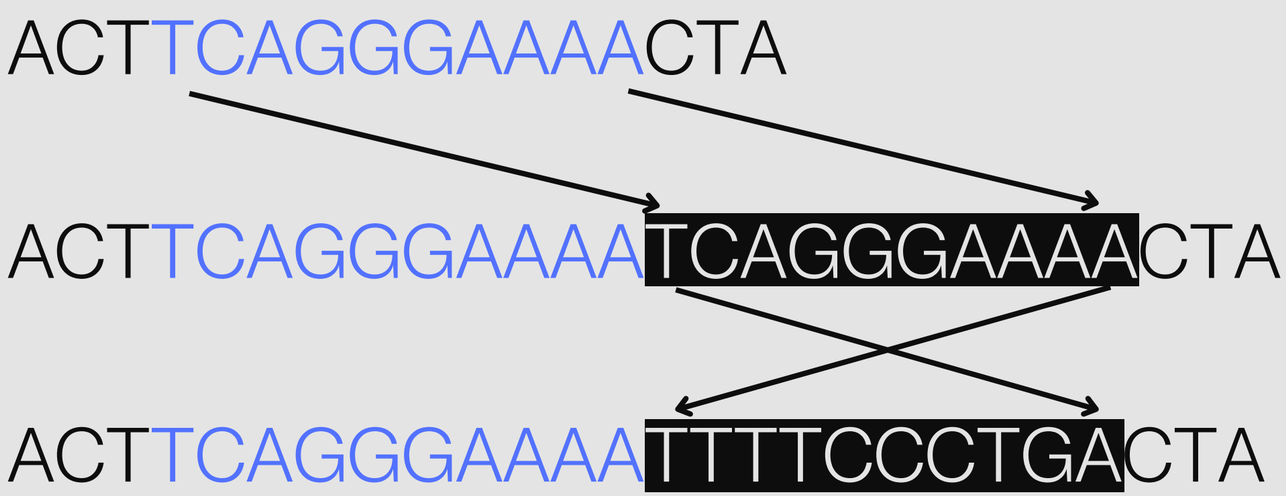
REFERENCE	TARGET	2 INVERSIONI		4 INVERSIONI		6 INVERSIONI	
		CORTE	LUNGHE	CORTE	LUNGHE	CORTE	LUNGHE
50000	50000	0.0188	0.0260	0.0407	0.0459	0.0721	0.0758
100000	50000	0.0353	0.0400	0.0828	0.0934	0.1336	0.1396
1000000	50000	0.03009	0.03029	0.7305	0.8364	1.2190	1.2535
10000000	50000	2.9765	3.0082	7.3140	7.6067	11.8296	12.1238

Risultati


Execution Time vs. Reference Length
2 Short
2 Long
4 Short
4 Long
6 Short
6 Long
Execution Time (s, log scale)
10^{1}
10^{0}
10^{-1}
10^{5}
10^{6}
10^{7}
Reference Length (log scale)

Duplicazioni Invertite

- Riconosce l’inversione ma non la presenza di una duplicazione invertita



R: ACTTCAGGGAAAACTA

T: ACTTCAGGGAAAATTTTCCCTGACTA

Conclusioni

FATTORE DOMINANTE

La lunghezza del genoma di riferimento r

BUONA EFFICIENZA

Per genomi di media lunghezza

SCALABILITÀ LIMITATA

Per l’influenza di r

Prospettive Future

OTTIMIZZAZIONE

Migliorare la gestione di sequenze di riferimento lunghe

GESTIONE ERRORI

Per poter gestire dati reali con mutazioni