UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Compression of k-mers set with counters

Leonardo Gemin, Enrico Rossignolo

November 10, 2022

Outline

- Introduction
- k-mer set representations
- Methods
- Results
- Conclusions

Introduction

- k-mer $\longrightarrow$ substring of length k
- Huge list $(4^k) \Longrightarrow$ bottleneck
- Wide range of application:
  - genome assembly
  - metagenomics
  - database searching
  - ...

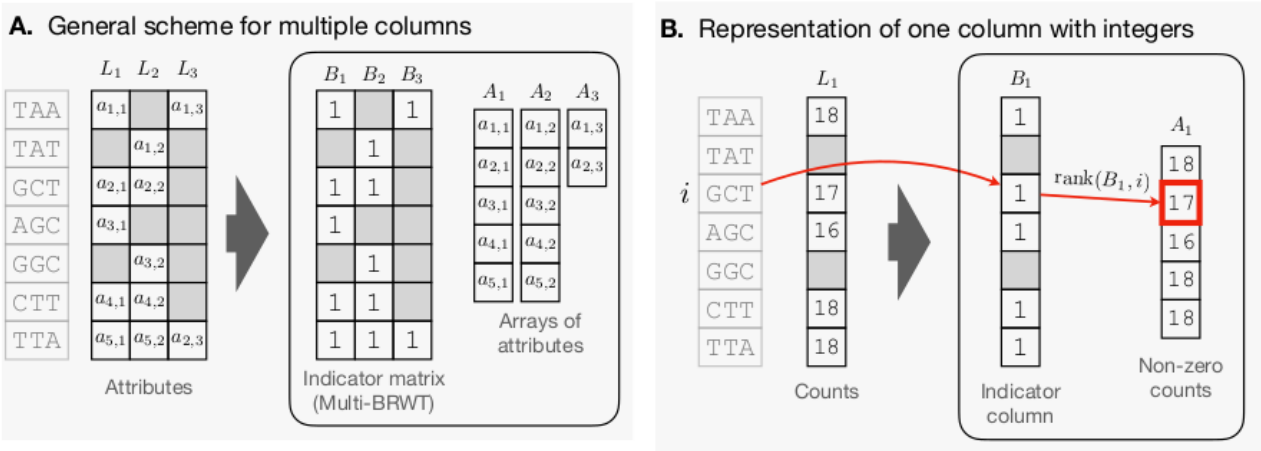
k-mer set representations

- (counting) de Bruijn graphs $\longrightarrow$ Metagraph
- contigs $\longrightarrow$ Metagraph
- unitigs $\longrightarrow$ BCALM2
- simplitigs $\longrightarrow$ prophAsm
- SPSSs $\longrightarrow$ UST

k-mer set representations: dBGs (1/2)

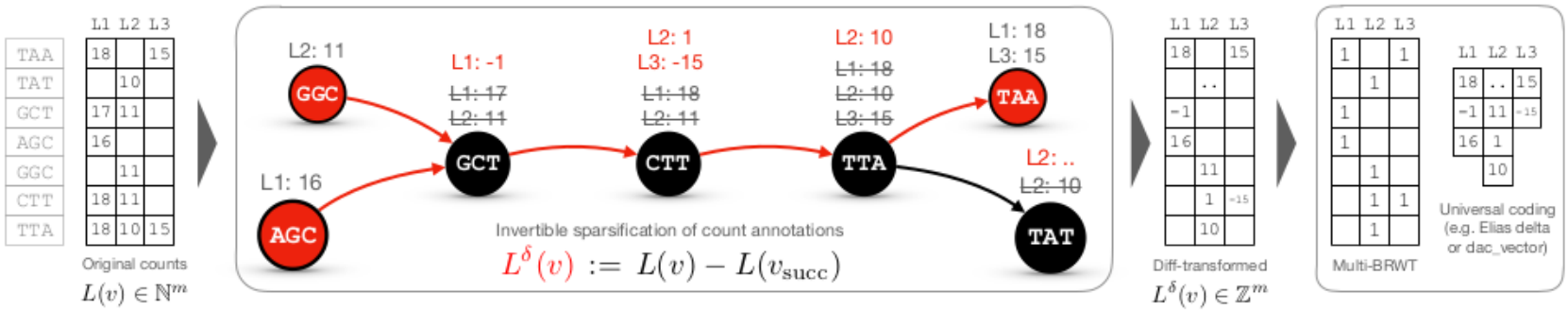
- **dBG(K)**: $G = (V, A)$ where
  - $V = K$
  - $A = \left\{(u, v) \in K^2 \middle| suf_{k-1}u = pref_{k-1}v\right\}$

- counting: external data structures



k-mer set representations: dBGs (2/2)

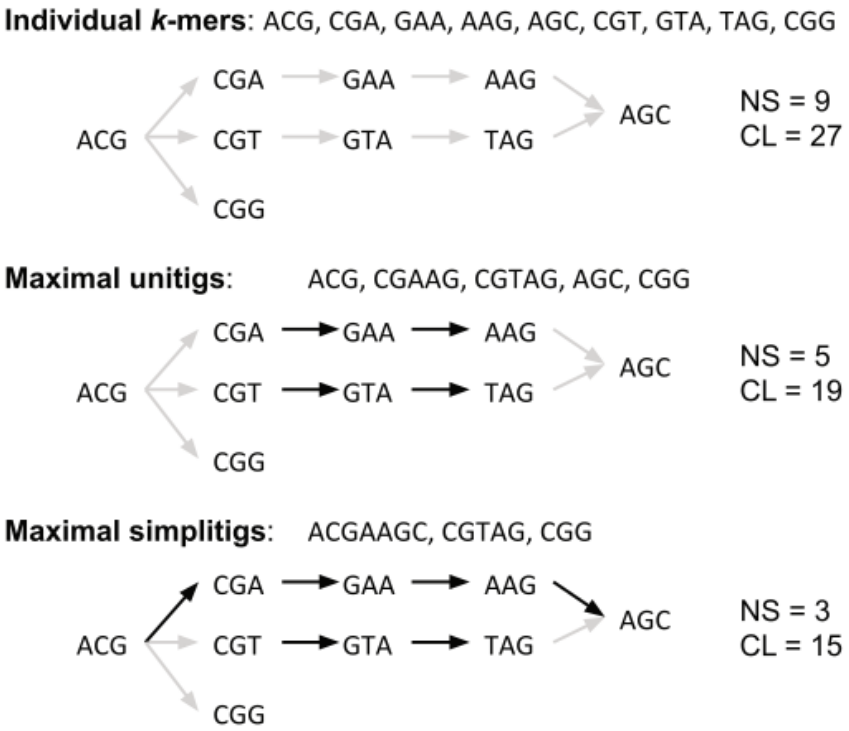
- RRR and Multi-BRWT for binary matrix
- row diff transform for annotations arrays
  - successors assigned arbitrarily
  - anchors are unchanged and break the recursion
  - $\Longrightarrow$ labels represented with less bits!
  - $\Longrightarrow$ binary matrix sparsified!



k-mer set representations: contigs and unitigs

- **contigs**: any path in a dBG
- **unitigs**: any path with non-branching internal nodes
- **simplitigs**: ?

- **simplitigs**: any path on a disjoint path cover in a dBG
- $CL = n_k + (k-1) \cdot NS \implies CL \geq n_k + k - 1$
- not tight: $CL^\star = 9 + (3-1) = 11 < 15$
- ProphAsm is sub-optimal with a greedy approach



k-mer set representations: SPSS (1/2)

- **k-spectrum** of a string s:

$$sp^k(s) \triangleq \{\{\text{canonized k-mers of } s\}\}$$

- for a set S:

$$sp^k(S) \triangleq \bigcup_{s \in S} sp^k(s)$$

- S is a **Spectrum Preserving String Set** of a k-mer set K iff
  1. $sp^k(S) = sp^k(K)$
  2. $|s| \geq k \;\; \forall s \in S$

k-mer set representations: SPSS (2/2)

- Let S^{opt} be a minimum SPSS representing K and W^{opt} a minimum path cover of cdBG(K). Then

$$weight(S^{opt}) = |K| + |W^{opt}| \cdot (k-1)$$

- W is an SPSS of $K \implies$ equivalence with simplitigs
- Lower bound (not tight)

$$weight(S^{opt}) \geq |K| + \left(\left\lceil \frac{n_{dead} + n_{sp}}{2} \right\rceil + n_{iso} \right) \cdot (k-1)$$

where $n_{sp} = \sum_{(u,su) \in (V,\{0,1\})} max\{0, |B_{u,su}| - 1\}$

- UST is sub-optimal with a greedy approach

Methods (1/2)

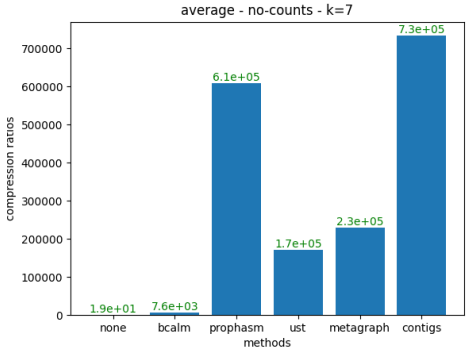
Accession	Description
SRR000001	Human haplotype map
SRR21394969	E. Coli
SRR21394970	Coronavirus 2
SRR21284212	Salmonella
SRR21073883	N. Gonorrhea
RND1664714109	Random sequence

Methods (2/2)

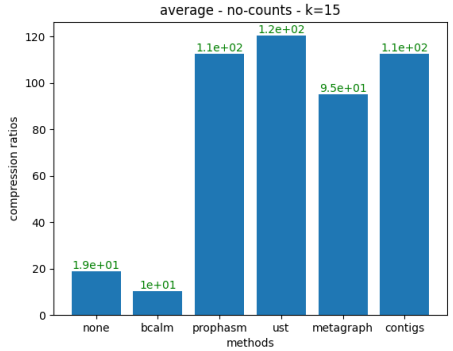
Workflow:

- from FASTA to k-mer
  - help: bash script
  - k-mer length: 7, 15, 23, 31
- with vs without counts
- compression
  - lzma
  - MFCompress
  - gZip

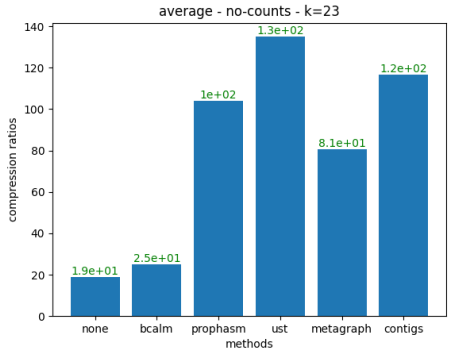
Results (without counts)



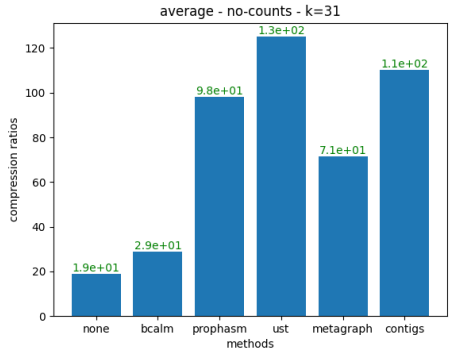
(a) $k = 7$



(b) $k = 15$

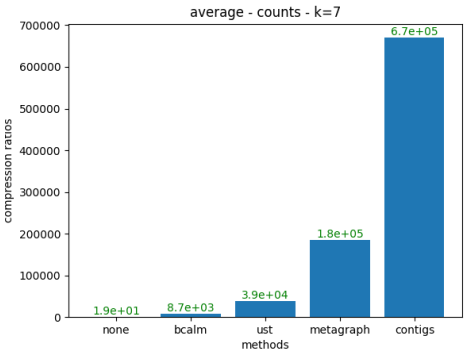


(c) $k = 23$

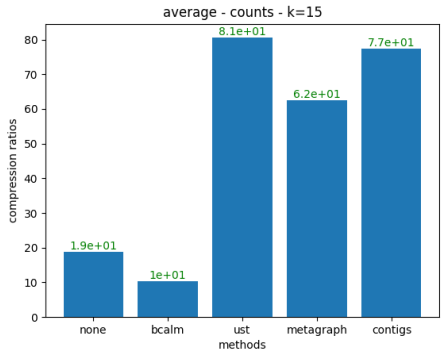


(d) $k = 31$

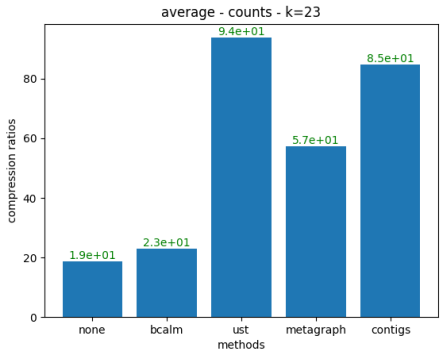
Results (with counts)



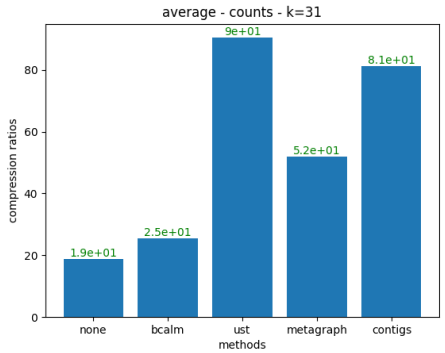
(a) $k = 7$



(b) $k = 15$



(c) $k = 23$



(d) $k = 31$

A particular case

SRR21073883 (N. **Gonorrhea**): really high compression ratios!

k-mer lenght	bcalm	prophasm	ust	metagraph	contigs
15	34.37	352.40	393.78	308.85	366.26
23	80.034	323.00	442.52	256.55	369.77
31	90.89	304.78	409.91	224.03	346.93

(a) without counts

k-mer length	bcalm	ust	metagraph	contigs
15	34.12	260.98	201.39	252.74
23	75.41	308.95	181.29	274.31
31	83.90	300.16	162.52	262.04

(b) with counts

Conclusions

- First observations:
  - tools return compression ratios higher than with a direct compression
  - low k-mer length $\longrightarrow$ high compression ratio
- Bcalm $\longrightarrow$ no advantage, depends on repetitiveness
- ProphAsm (only without counts) $\longrightarrow$ better with $k = 7$
- UST $\longrightarrow$ the best among all
- Metagraph $\longrightarrow$ not particularly efficient
  - contigs $\longrightarrow$ rivals UST

The End

Thanks for your attention!

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Compression of k-mers set with counters

Leonardo Gemin, Enrico Rossignolo

November 10, 2022