UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

USTAR: Improved Compression of k-mer Sets with Counters Using De Bruijn Graphs

Enrico Rossignolo Matteo Comin

ISBRA 2023

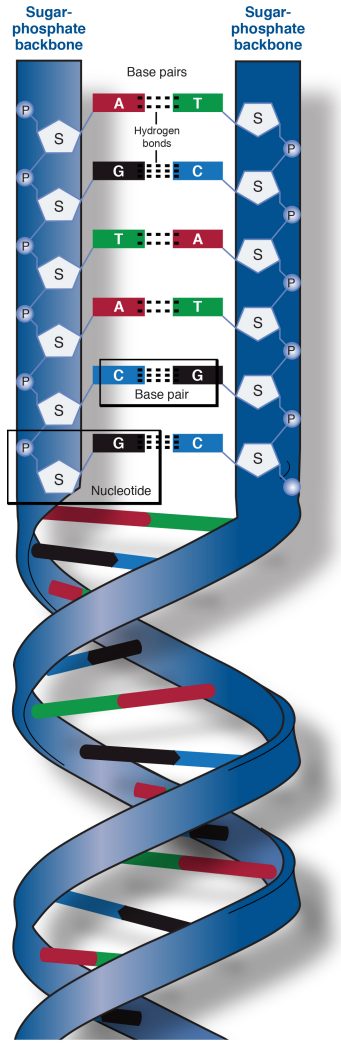
Outline

- Introduction
  - k-mers set and counters
  - applications
  - problems
- Methods
  - k-mers set compression: UST
  - Vertex-disjoint path cover problem
  - USTAR strategy
- Experimental setup
- Results
  - different k-mer sizes
  - graph density
  - time and memory
- Conclusions and future works

k-mers and counters

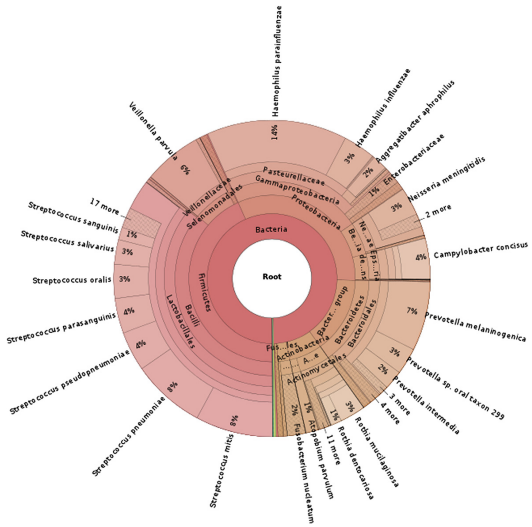
- k-mers: substrings of nucleobases (A,C,T,G) of length k
  - from longer DNA sequences
- counters: k-mers multiplicity
  - small values: read error
  - big values: repeat
- e.g. 7-mers list:

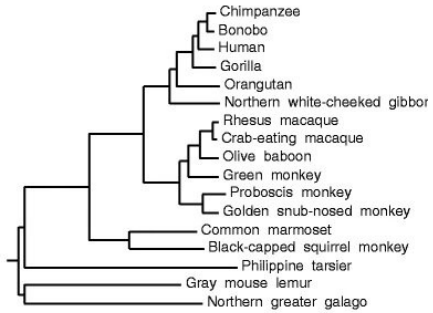
```
AAAAAAA 101
AAAACCC 42
AAAAGTA 1
...
```



Applications

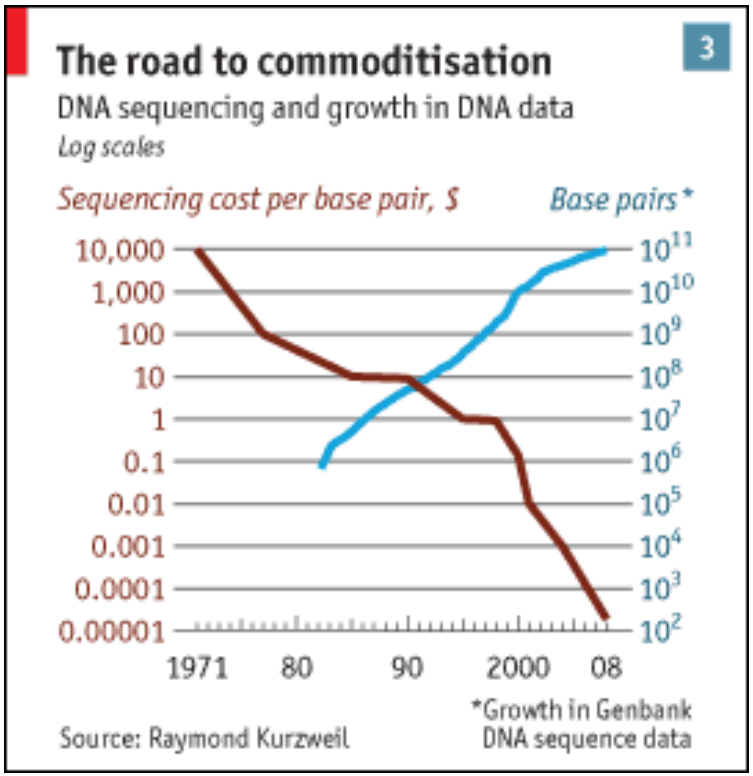
- Assembly: Eulerian walks in cdBGs that can be found efficiently (Spades)
- Metagenomics:
  Taxonomy labelling (Kraken[1])
  - 900 times faster than tools based on alignment
- Phylogenetics: Mash[2]
- Database searching (BIGSI)





[1] *Wood, D.E., Salzberg, S.L. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15, R46 (2014)*
[2] *Ondov, B.D., Treangen, T.J., Melsted, P. et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17, 132 (2016)*

- cost per genome drops $\Longrightarrow$ a lot of data (BIGSI: 13 TB)
- huge datasets need medium to long term storage for analysis
- need an efficient data structure that allows queries



The Economist

k-mers set compression

- minimum number of bits needed to compress n k-mers[1]:

$$log \binom{4^k}{n}$$

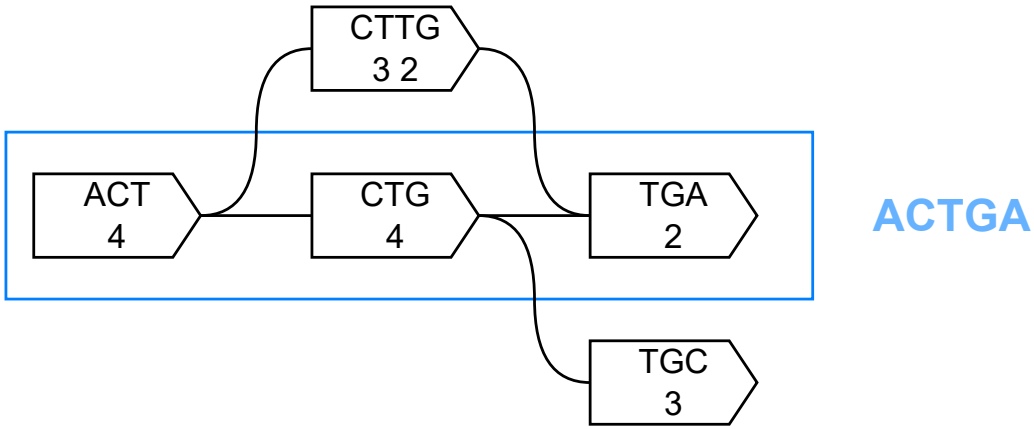
- k-mer counters
  - KMC: disk files
  - DSK: hash tables
  - Squeakr: Bloom filters

- leverage spectrum-like property
  - ProphAsm
  - UST

[1] *Conway, T.C., Bromage, A.J.: Succinct data structures for assembling large genomes. Bioinformatics 27(4), 479–486 (01 2011)*

k-mers set compression: UST

- UST[1] uses compacted de Bruijn Graphs (cdBGs)
  - nodes: k-mers
  - arcs: between k-mers that share a $k-1$ string
- UST glues k-mers: $k-1$ characters saved per gluing

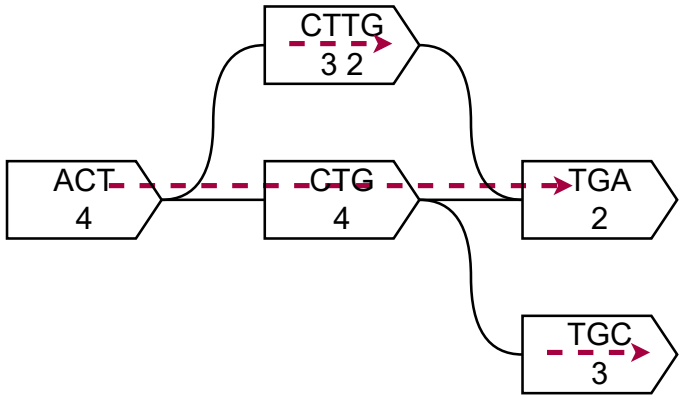


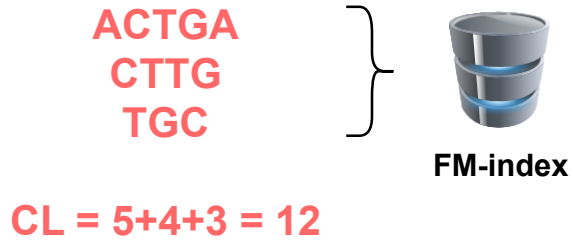
[1] *Rahman A, Medevedev P. Representation of k-Mer Sets Using Spectrum-Preserving String Sets. J Comput Biol. 2021*

Vertex-disjoint path cover

- The minimum **vertex-disjoint path cover problem**: find the minimum number of vertex-disjoint paths that cover the graph.

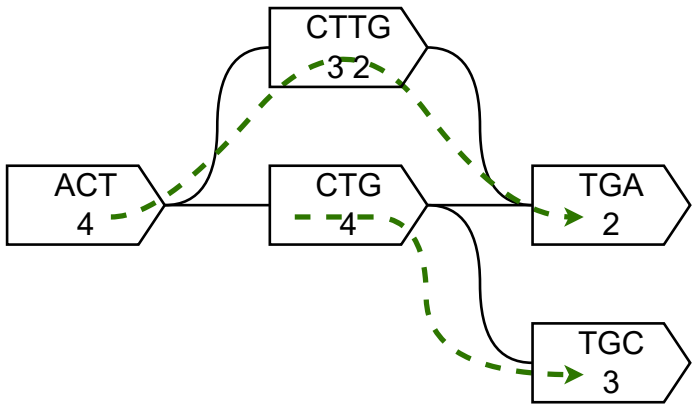
- UST and ProphAsm search for an approximate solution using a greedy approach: always take the first node available
  - choose a node as seed; extend it as much as possible; repeat
  - this can lead to isolated nodes!





USTAR: our strategy

- USTAR selects the less connected nodes
  $\Longrightarrow$ avoid congested nodes and reduce isolated nodes
- compress counts that generally have a skewed distribution: select as seed the node with greater average count
  $\Longrightarrow$ bring equal counts close together



ACTTGA
CTGC

CL = 6+4 = 10
CL = 5+4+3 = 12

FM-index

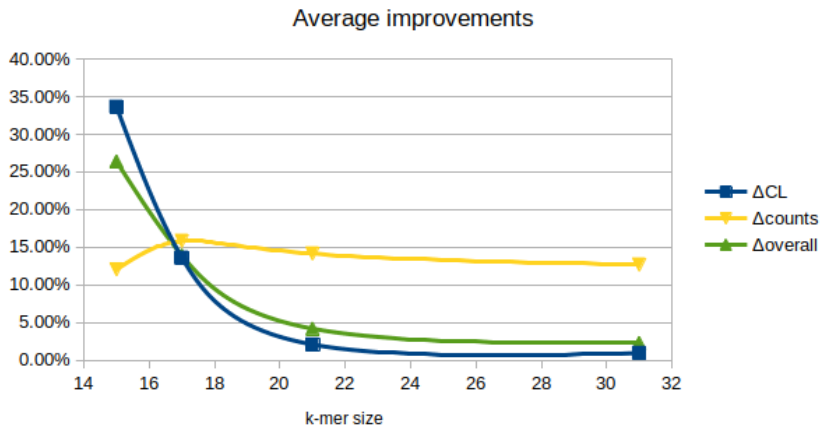
Experimental setup

- 20 dataset taken from major papers on k-mer compression

Dataset	DSK	KMC	Squeakr	BCALM	UST	USTAR
SRR001665_1	76,729,965	63,102,295	62,769,135	43,100,358	12,641,658	12,332,551
SRR001665_2	89,356,517	73,618,021	70,364,023	54,549,879	15,492,263	15,109,673
SRR061958_1	1,853,355,280	1,512,526,861	1,214,304,214	792,616,145	194,173,905	185,905,825
SRR061958_2	2,269,394,440	1,850,606,165	1,445,986,432	940,752,737	235,657,588	225,975,765
SRR062379_1	771,892,475	633,615,665	559,244,946	334,386,186	82,713,766	79,283,723
SRR062379_2	766,644,876	629,023,699	556,418,241	327,042,925	80,164,746	76,708,406
SRR10260779_1	594,043,132	489,620,438	459,620,233	272,605,742	64,644,700	61,724,139
SRR10260779_2	661,730,544	545,447,915	501,793,581	311,074,932	72,772,294	69,375,320
SRR11458718_1	660,336,575	547,192,385	515,587,875	278,247,157	64,694,925	61,236,404
SRR11458718_2	699,675,661	580,313,686	542,321,885	304,467,895	68,982,466	65,438,050
SRR13605073_1	286,147,403	236,056,529	244,522,615	110,324,321	25,833,347	24,546,244
SRR14005143_1	72,421,457	59,702,423	75,386,963	26,222,881	6,419,520	6,220,215
SRR14005143_2	148,413,200	121,547,826	126,063,532	51,976,493	13,117,896	12,655,430
SRR332538_1	61,647,503	50,466,675	65,343,140	21,192,576	5,737,778	5,599,034
SRR332538_2	125,336,255	100,440,228	116,698,603	77,057,667	14,410,775	13,528,977
SRR341725_1	972,617,730	799,134,833	700,565,933	262,398,076	80,436,678	78,193,253
SRR341725_2	1,005,087,513	825,643,578	719,993,731	277,159,709	84,250,689	81,877,574
SRR5853087_1	1,494,920,206	1,234,975,195	1,084,532,779	1,073,165,506	191,108,921	177,278,725
SRR957915_1	1,016,375,644	837,315,550	732,334,056	590,259,971	122,748,678	116,872,195
SRR957915_2	1,589,786,146	1,301,318,582	1,062,835,997	829,172,816	182,073,051	172,757,385
Average	760,795,626	624,583,427	542,834,396	348,888,699	**80,903,782**	**77,130,944**

- Given the k-mer set S, we used the following metrics:
  - $CL = \sum_{s \in S} |s|$
  - *counts*: size of compressed counts file
  - *overall* = *fasta* + *counts*

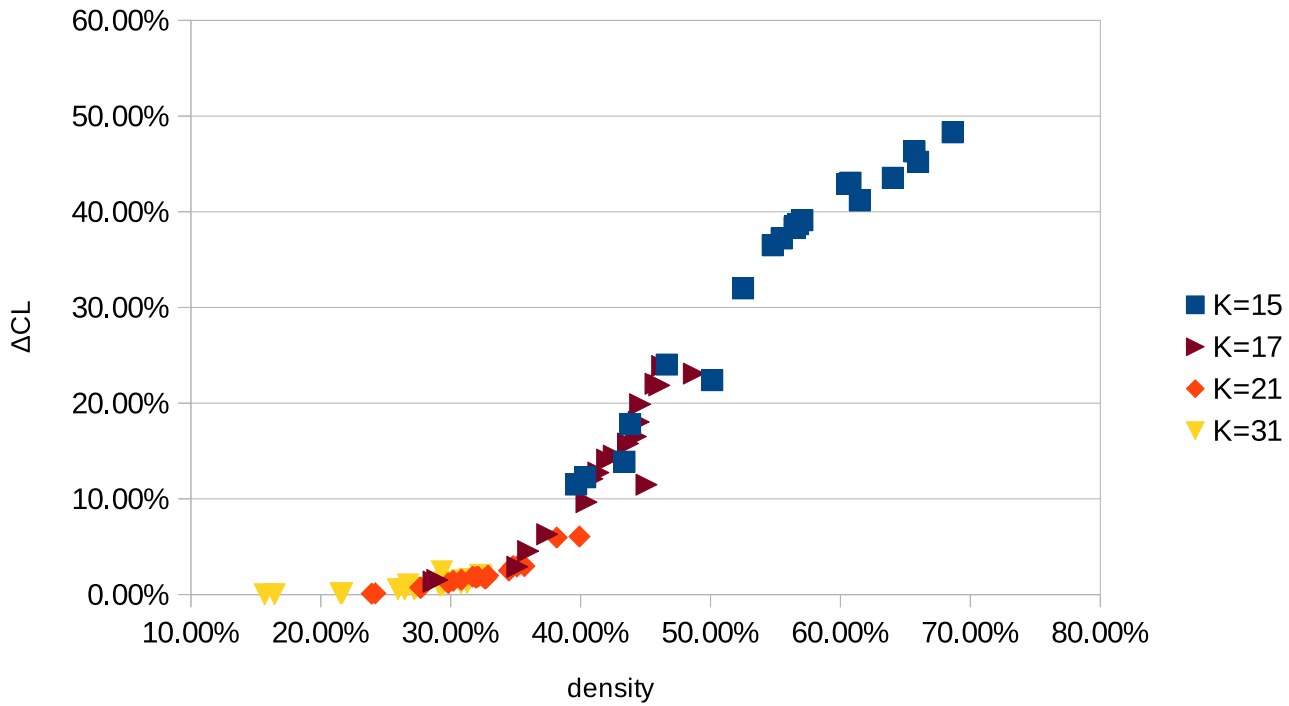
Results: varying k



k-mer size	ΔCL	Δcounts	Δoverall
15	33.64%	12.07%	26.40%
17	13.61%	15.85%	13.92%
21	2.10%	14.17%	4.20%
31	0.97%	12.70%	2.30%

- improvement w.r.t. UST
- for every value of k we have better compression
- counts compression is almost constant
- overall compression increases for lower values of k

Results: graph density



- each point represents a k-mer set
- $density = \frac{\#arcs}{maximum\ \#arcs}$
- denser graph $\Longrightarrow$ more arcs $\Longrightarrow$ more paths
- better results: USTAR can choose among many paths

Results: time and memory

Average	Time [s]		Memory [GB]	
k-mer size	UST	USTAR	UST	USTAR
15	1,647	**228**	26	**12**
17	681	**98**	10	**6**
21	270	**49**	3	**3**
31	191	**45**	3	**3**

- Time and memory increase with lower k due to the higher number of arcs (density)
- On average, USTAR uses less time and memory

Conclusions

- USTAR consistently outperforms UST obtaining smaller k-mers and counts files
- Taking the less connected node is a good strategy
- Denser graphs lead to better compression
- In the future larger datasets will produce denser graphs
- Future works:
  - choose unbalanced nodes as seeds
  - reuse the same node if needed

Thanks for your attention!

Future works

name	UST	USTAR	USTAR2
SRR001665_1	36,357,928	36,324,848	33,638,588
SRR001665_2	45,751,142	45,694,102	41,864,643
SRR061958_1	623,862,618	191,039,506	178,434,526
SRR061958_2	767,654,838	211,459,109	198,026,097
SRR062379_1	252,418,995	248,519,235	226,998,129
SRR062379_2	246,073,774	241,478,754	220,352,087
SRR10260779_1	188,012,488	184,854,088	170,629,253
SRR10260779_2	214,245,523	210,202,663	192,382,255
SRR11458718_1	189,827,141	185,070,581	170,218,475
SRR11458718_2	202,891,014	196,865,834	179,815,285
SRR13605073_1	86,006,020	84,974,720	81,822,046
SRR14005143_1	19,355,339	19,020,479	17,477,215
SRR14005143_2	42,328,593	41,492,693	37,213,243
SRR332538_1	18,649,027	18,382,747	17,615,333
SRR332538_2	49,648,910	46,689,430	41,053,913
SRR341725_1	245,548,134	243,816,714	236,221,721
SRR341725_2	258,344,641	256,477,401	247,741,588
SRR5853087_1	587,246,289	551,618,109	484,650,727
SRR957915_1	377,292,074	366,210,794	325,707,476
SRR957915_2	579,294,390	562,058,930	501,809,029
average	251,540,444	197,112,537	**180,183,581**

- USTAR2 chooses as seed unbalanced node (in-out arcs difference)
- already used node can be selected by USTAR2 during path extension

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

USTAR: Improved Compression of k-mer Sets with Counters Using De Bruijn Graphs

Enrico Rossignolo Matteo Comin

ISBRA 2023