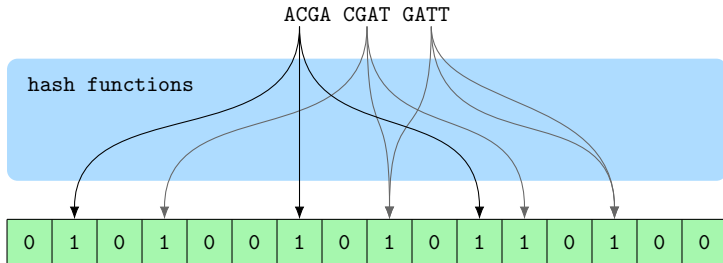


COBS: A Compact Bit-Sliced Signature Index

Timo Bingmann, Phelim Bradley, Florian Gauger, and Zamin Iqbal · 2019-10-08 @ SPIRE'19

INSTITUTE OF THEORETICAL INFORMATICS – ALGORITHMICS



Abstract

We present COBS, a COmpact Bit-sliced Signature index, which is a cross-over between an inverted index and Bloom filters. Our target application is to index k -mers of DNA samples or q -grams from text documents and process approximate pattern matching queries on the corpus with a user-chosen coverage threshold. Query results may contain a number of false positives which decreases exponentially with the query length. We compare COBS to seven other index software packages on 100 000 microbial DNA samples. COBS' compact but simple data structure outperforms the other indexes in construction time and query performance with Mantis by Pandey et al. in second place. However, unlike Mantis and other previous work, COBS does not need the complete index in RAM and is thus designed to scale to larger document sets.



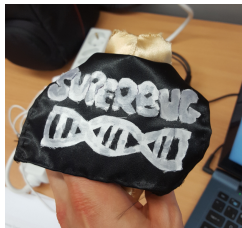
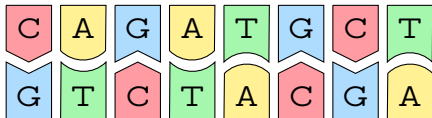
This document is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

Motivation

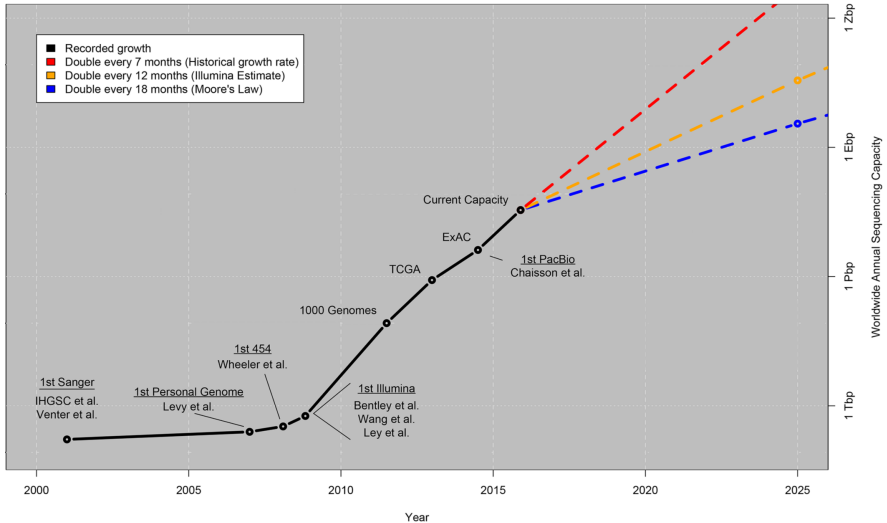
Need **approximate search** in
petabytes of DNA data.

Applications:

- study global threats to public health
- epidemiology
- basic science of disease



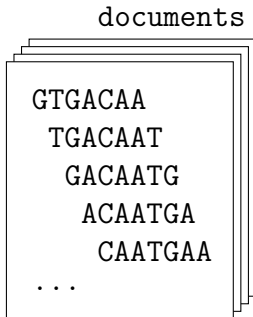
Motivation



from Stephens et al. "Big data: astronomical or genomical?" (2015)

Approximate Pattern Matching

query
ATGACAATGACG
←————→
100-1000



k-mers/*q*-grams

Approximate Pattern Matching

query

ATGACAATGACG

ATGACAA

TGACAAT

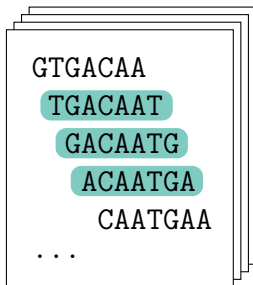
GACAATG

ACAATGA

CAATGAC

AATGACG

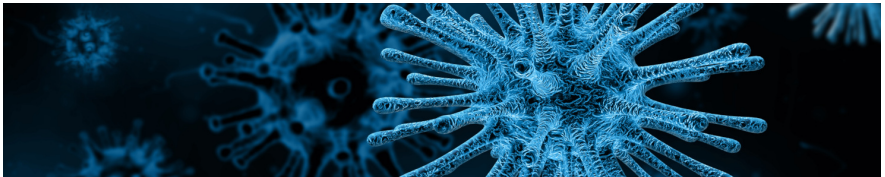
documents



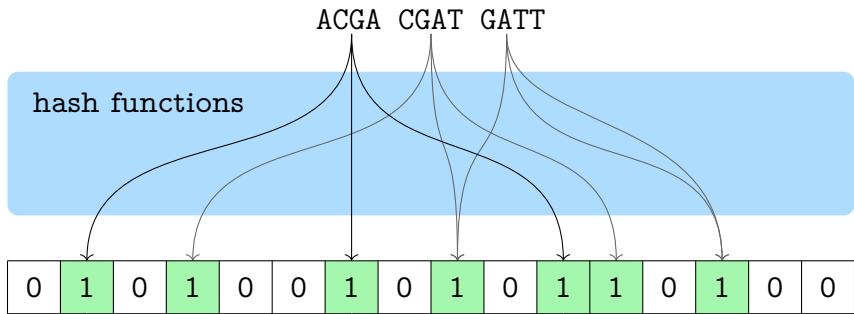
k-mers/*q*-grams

Related Work

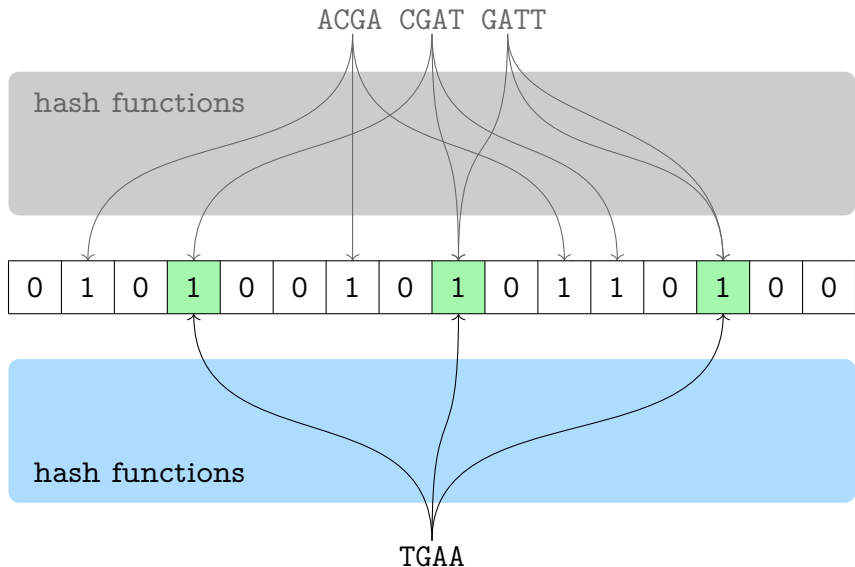
- Sequence Bloom Tree [SK16]
- Split Sequence Bloom Tree [SK18]
- AllSome Sequence Bloom Tree [Sun+18]
- HowDe Sequence Bloom Tree [HM18]
- SeqOthello [Yu+18]
- MANTIS [Pan+18]
- Bitsliced Genomic Signature Index [Bra+19]



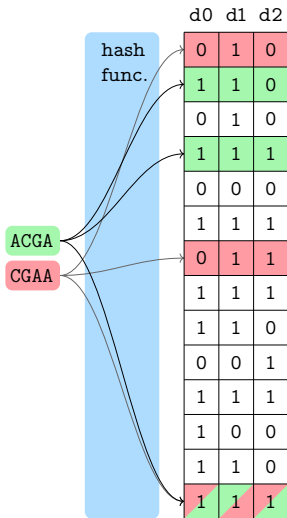
Bloom Filter



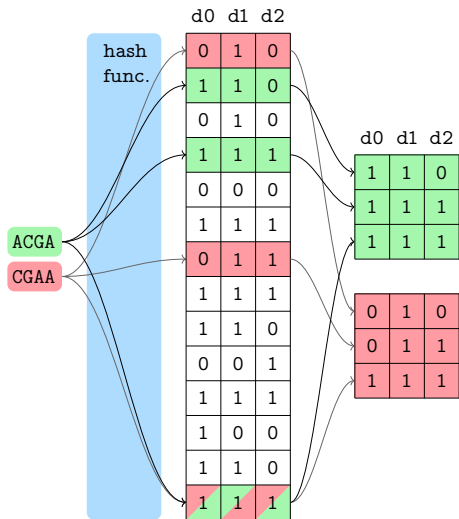
Bloom Filter



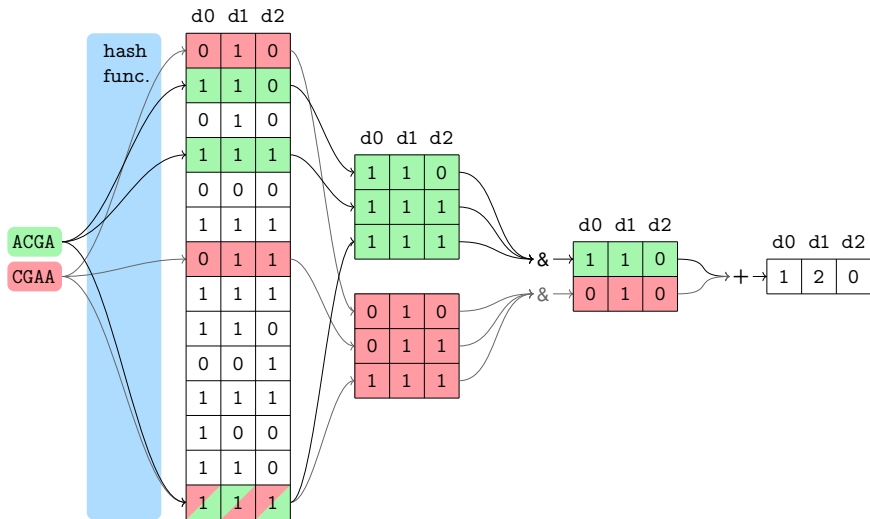
Bit-Sliced Signature Index



Bit-Sliced Signature Index



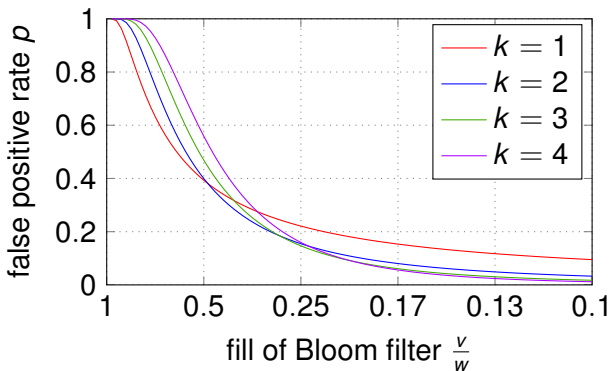
Bit-Sliced Signature Index



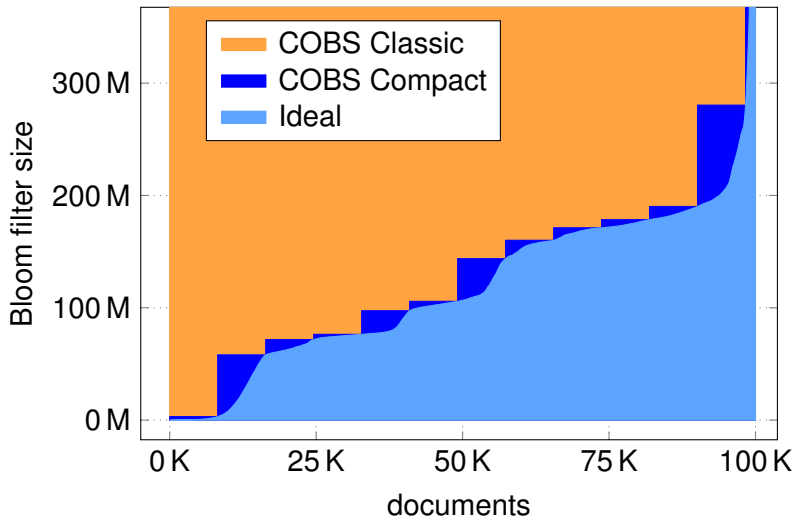
Bloom Filter Parameters

Theorem: False Positive Rate of a Query, Thm 2 in [SK16]

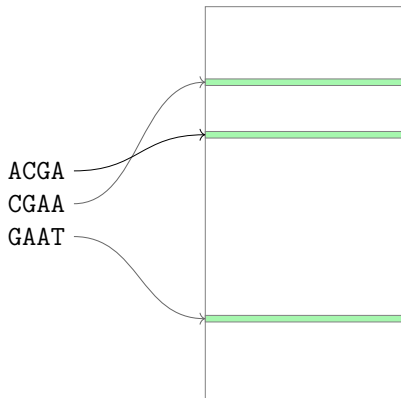
Let P be a query containing ℓ distinct q -grams and K a threshold. If we consider the terms as being independent, the probability that more than $\lfloor K\ell \rfloor$ false-positive terms occur in a filter f with false positive rate p is $1 - \sum_{i=0}^{\lfloor K\ell \rfloor} \binom{\ell}{i} p^i (1-p)^{\ell-i}$.



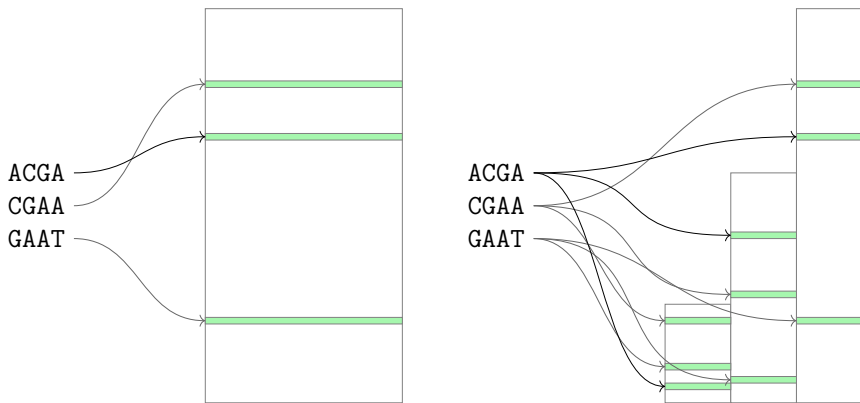
Compact Bit-Sliced Signature Index



COBS: Disk Access Pattern



COBS: Disk Access Pattern



more about disk, SSD, and NVMe access pattern speeds:

<https://panthema.net/2019/0322-nvme-batched-block-access-speed/>

COBS: Summary

COBS Index Design: (values used in practice)

- use $k = 1$ hash functions with $f = 0.3$ false positive rate
- compact $\Theta(B) = 4 \text{ Ki}$ documents into subindices

COBS Software:

- C++ implementation started by Florian Gauger
- can read Text, Fasta, Fastq, and McCortex files
- parallelized and multi-level if needed construction
- SIMD instructions in query processing



Experiments – Software and Machine

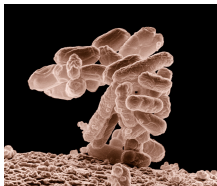
Eight Software Packages:

- Sequence Bloom Tree (SBT) [SK16]
- Split Sequence Bloom Tree (SSBT) [SK18]
- AllSome Sequence Bloom Tree (AllSome-SBT) [Sun+18]
- HowDe Sequence Bloom Tree (HowDe-SBT) [HM18]
- MANTIS [Pan+18]
- SeqOthello [Yu+18]
- Bitsliced Genomic Signature Index (BIGSI) [Bra+19]
- our Classic Bit-Sliced Index (Classic BSI) [this]
- and COBS [this]

Machine:

- Intel Gold 6138 2.0 GHz 4 × 20 cores with 768 GiB RAM.
- 4 × 2 TB NVMe Samsung 970 EVO SSD as software RAID 0.

Experiments – Data



Microbial Data:

- 100 000 microbial (viri and bacteria) documents from European Nucleotide Archive (ENA)
- Split into 100, 250, 500, 1 000, 2 500, . . . , 100 000 subsets.
- Average document size ≈ 42.77 MiB, ≈ 4 TiB in total.
- ENA contained $1.5 \cdot 10^9$ documents in 2018.

Queries: four batches, with

- length $\ell \in \{31, 100, 1\,000, 10\,000\}$,
containing $q \in \{100\,000, 100\,000, 10\,000, 1\,000\}$
random true positives and q true negatives.
- Check each index software's results.

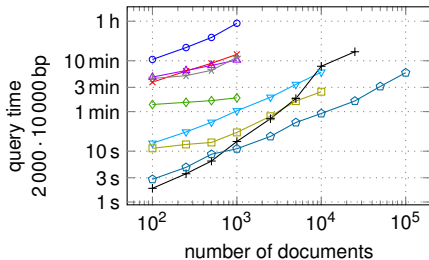
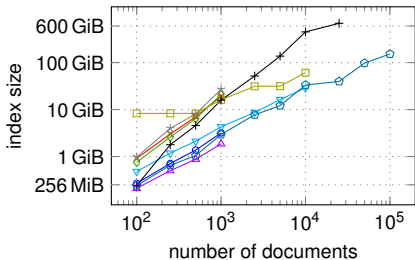
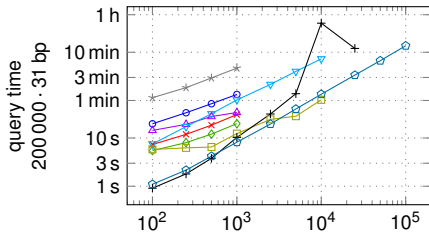
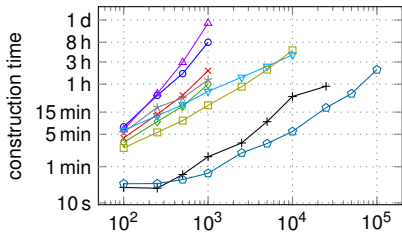
Results for 1000 Microbial Documents

phase	SBT	SSBT	AllSome-SBT	HowDe-SBT	Seq-Othello	Mantis	BIGSI	Classic BSI	COBS Compact
Construction Wall-Clock Time in Seconds									
count	2018	1974	1954	1959					
bloom	114	117	140	144	295	232	1881		
build	3097	21378	1401	68034	2225	987	2574	99	43
compress	1768	5187	80	3802		45			
total	6996	28657	3576	73939	2520	1264	4455	99	43
Construction CPU (User) Time in Seconds									
count	4574	4511	4475	4488					
bloom	11133	10967	10234	10278	28123	19162	169345		
build	855	5178	449	66872	2198	943	1767	1604	1430
compress	1569	4832	1663	2857		3423			
total	18131	25489	16821	84495	30320	23527	171113	1604	1430
Construction Maximum RSS Memory Usage in MiB									
count	518	518	518	518					
bloom	641	640	640	640	634	1756	4244		
build	11028	1523	7140	108147	12137	88357	246806	16245	2616
compress	10953	992	560	963		16613			
maximum	11028	1523	7140	108147	12137	88357	246806	16245	2616
Index Size in MiB									
size	19844	3254	21335	1911	4410	16486	27794	16236	3022

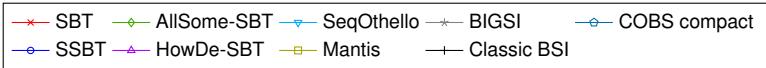
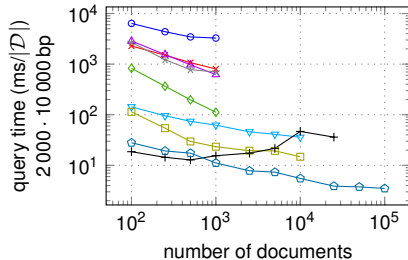
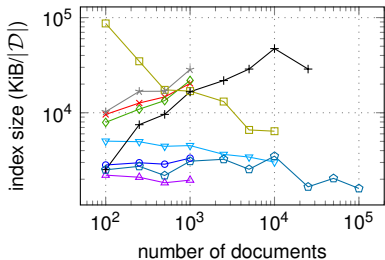
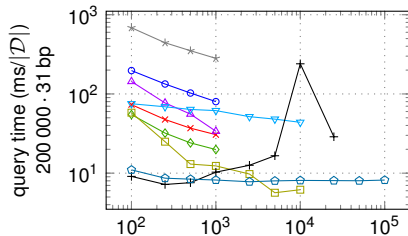
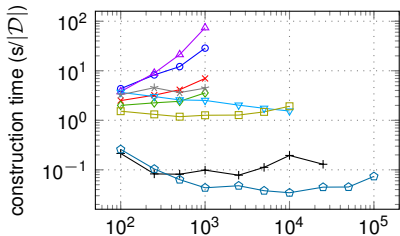
Results for 1000 Microbial Documents

phase	SBT	AllSome- SSBT	HowDe- SBT	Seq- Othello	Mantis	BIGSI	Classic BSI	COBS Compact	
ℓ	Query Wall-Clock Time in Seconds								
31 bp r0	31	80	20	34	62	12	281	10	8
31 bp r2	26	76	19	33	62	13	289	9	8
100 bp r0	663	3 183	100	600	73	22	783	14	9
100 bp r2	649	3 153	95	588	73	23	455	14	9
1000 bp r0	794	3 466	112	670	63	21	660	15	10
1000 bp r2	781	3 435	108	659	64	27	310	13	10
10000 bp r0	802	3 273	112	622	62	23	699	16	11
10000 bp r2	790	3 243	111	613	62	22	316	15	11
total r0-r2	6 775	29 833	1 007	5 710	783	252	5 177	154	114
	Document False Positive Rate for 31 bp Queries								
rate	0.004	0.004	0.004	0.004	0.001	0.000	0.027	0.024	0.227

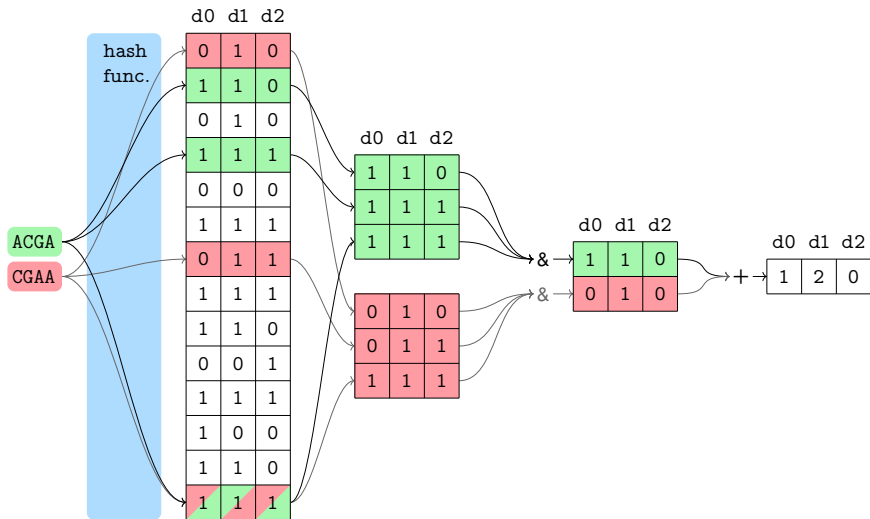
Scaling Results Microbial Documents



Scaling Results Microbial Documents



Bit-Sliced Signature Index



Conclusion

Software:

- COBS is available as **open source**:
<https://panthema.net/cobs/>
- soon: more documentation and **Python** front-end module

Future Work:

- Daniel Ferizovic tried **clustering** of documents
- also working on dealing with **insertions** and **deletions**
- **batched query** processing e.g. for whole genomes
- **distributed** COBS query processing for ENA-scale index
- adapt **completely different filter** for use as an index

Questions?