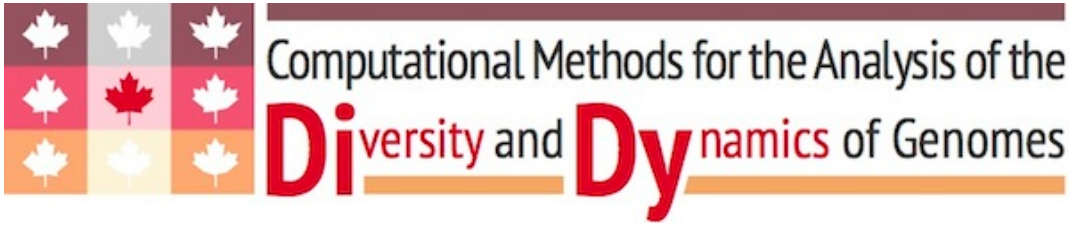


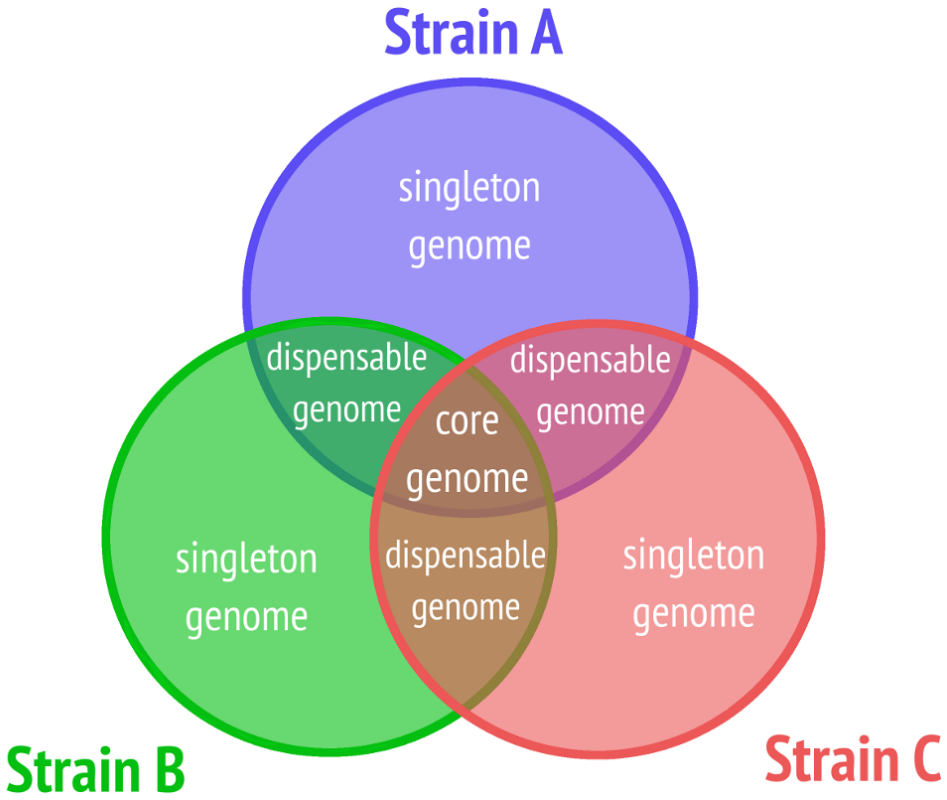
BLOOM FILTER TRIE

AN EFFICIENT DATA STRUCTURE FOR COLORED DE BRUIJN GRAPHS

Guillaume Holley
Roland Wittler
Jens Stoye

Genome Informatics group
Faculty of Technology
Bielefeld University

PAN-GENOME



Set of strain genomes from the same species

REQUIREMENTS

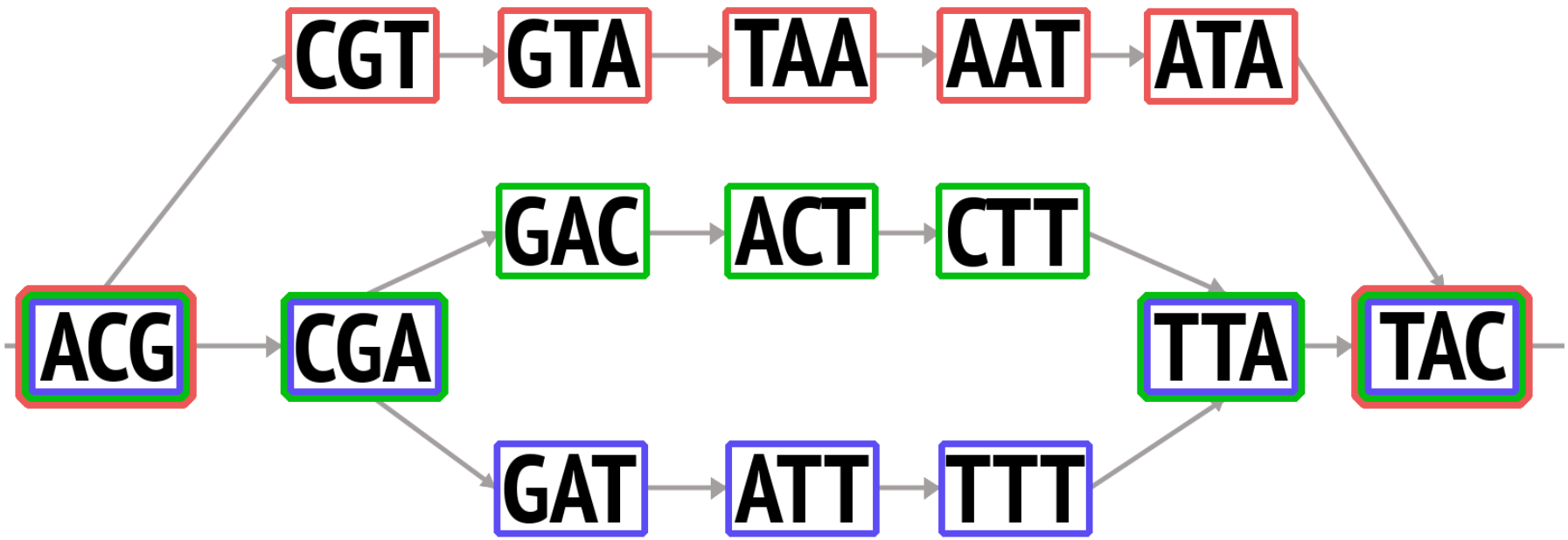
→ assembled or unassembled genomes as input

→ alignment-free

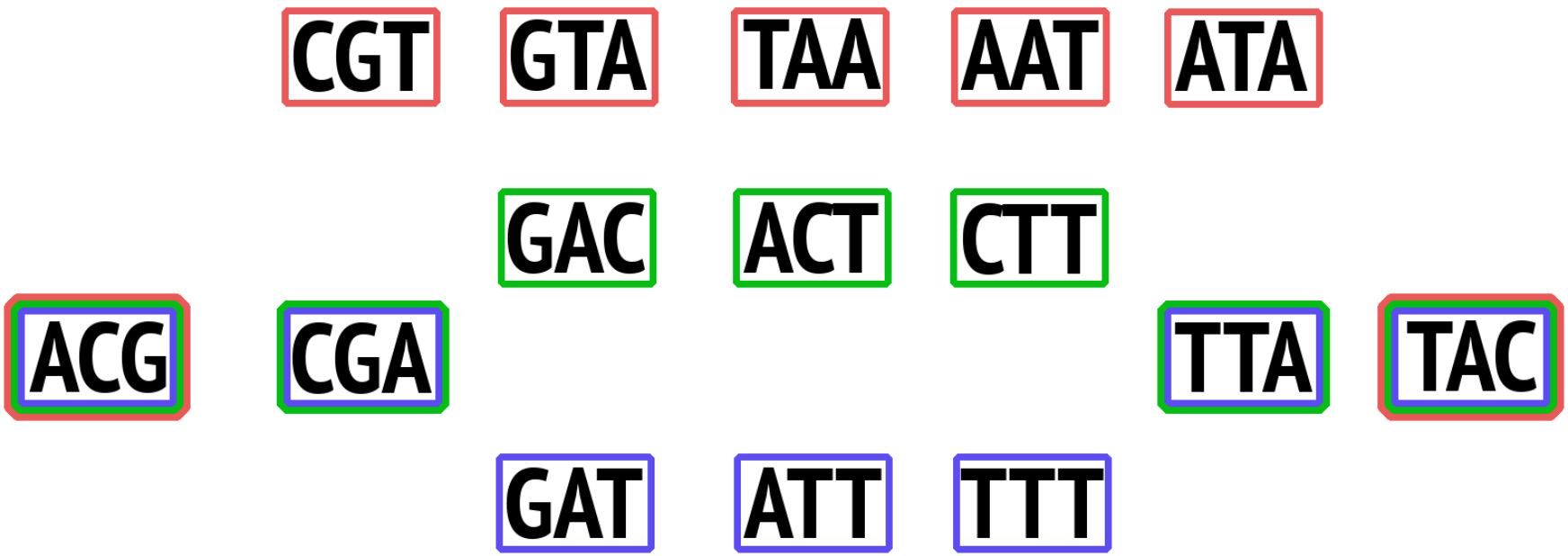
→ reference-free

→ incremental

COLORED DE BRUIJN GRAPH (IQBAL ET AL. 2012)



COLORED DE BRUIJN GRAPH



TRIE (FREDKING. 1960)

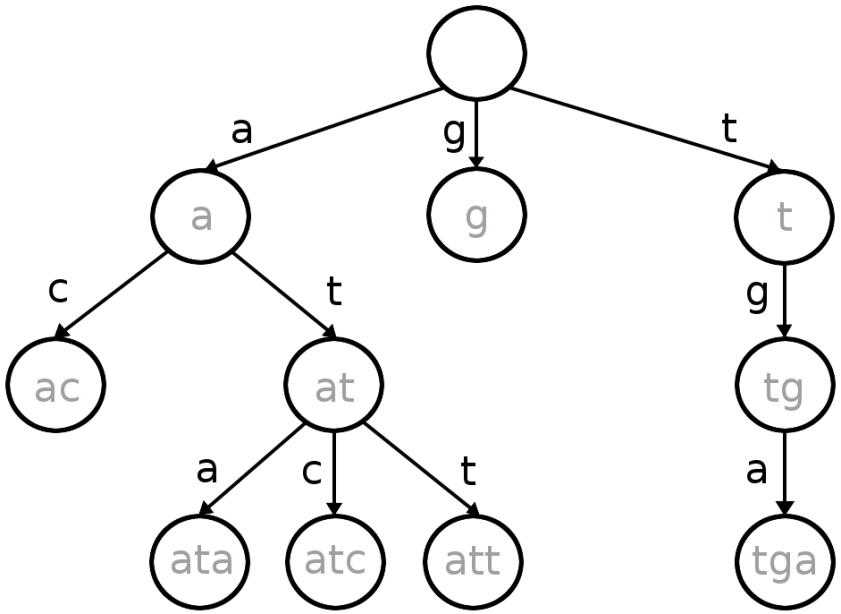


Figure: Trie for “ac”, “ata”, “atc”, “att”, “g” and “tga”.

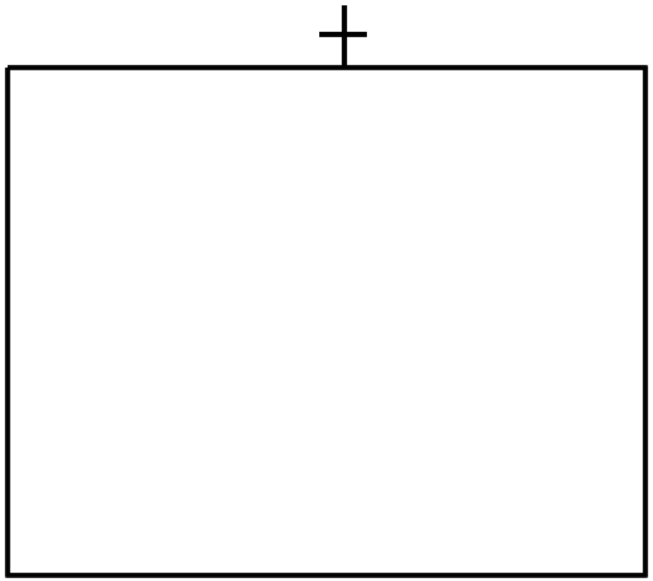
→ Stores shared prefix once

→ No need to know how many strings to insert

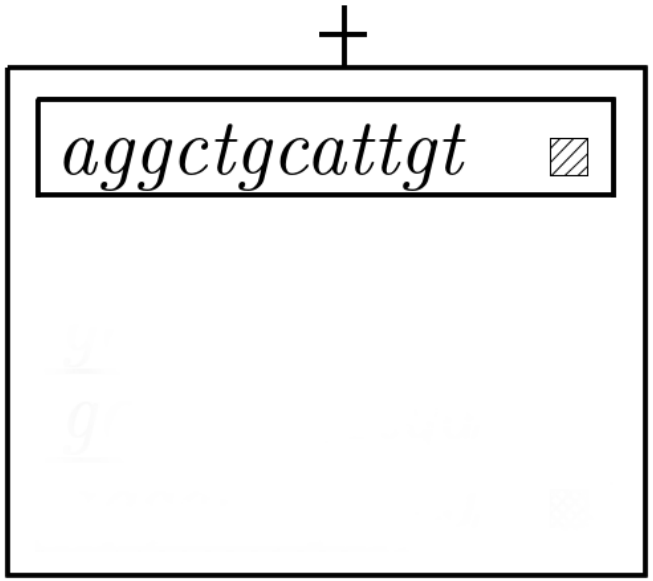
→ Optimal use of the memory

Bloom Filter Trie

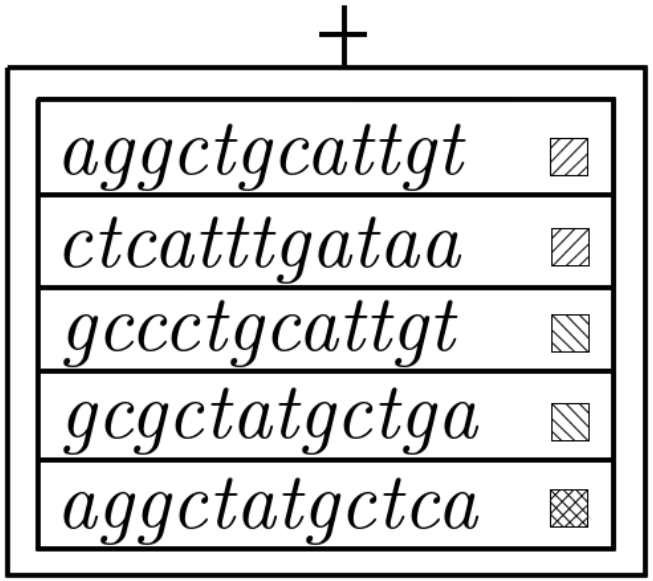
EMPTY NODE



NODE WITH UNCOMPRESSED CONTAINER



NODE WITH UNCOMPRESSED CONTAINER



BURSTING OF AN UNCOMPRESSED CONTAINER

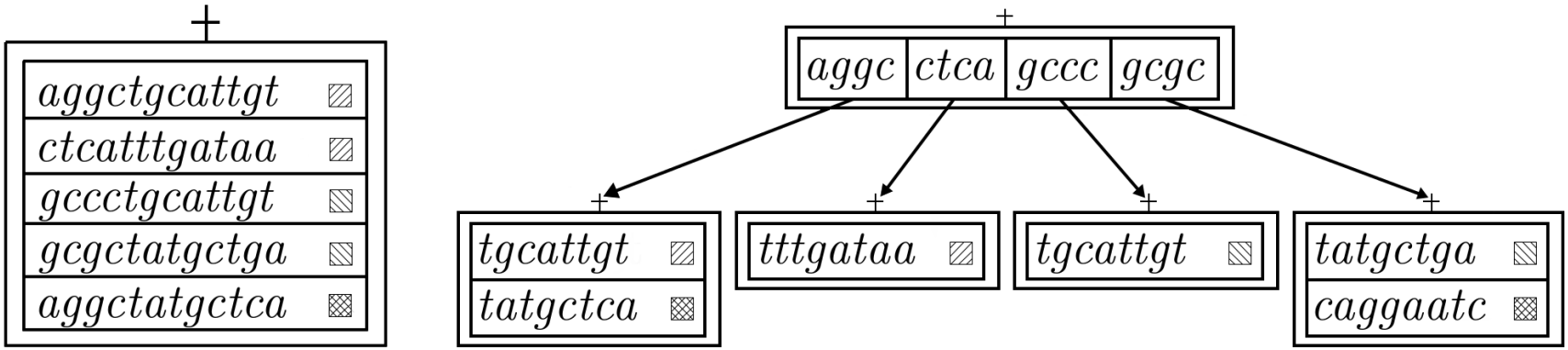
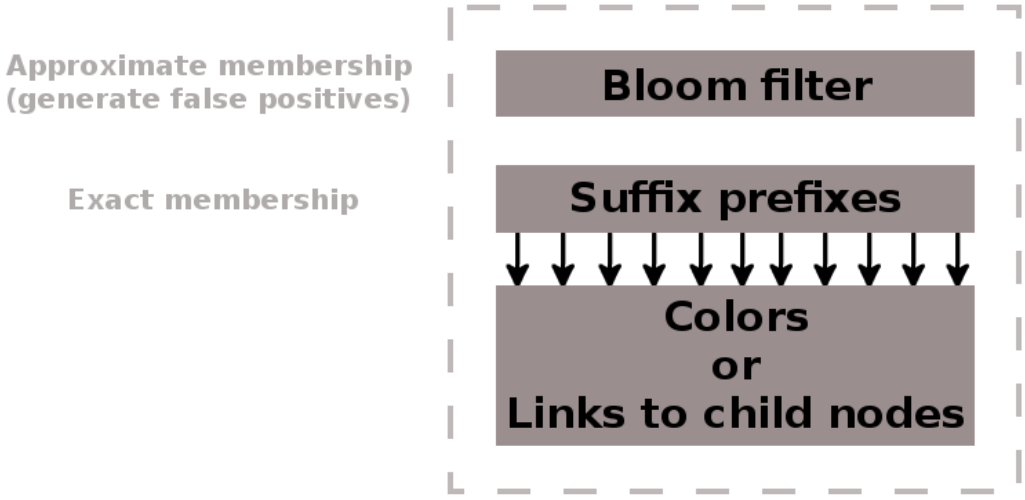


Figure: (a) BFT before bursting (b) “gcgccaggaatc” is inserted, uncompressed container capacity is exceeded, a burst is triggered. Prefix length is 4.

COMPRESSED CONTAINER



Bloom filter for:

→ Approximate membership of suffix prefixes in CCs

→ Inserting suffix prefixes in CCs

→ Accelerate graph traversal

Results

EXPERIMENT

Dataset: Bioproject PRJEB5438

473 isolates of *Pseudomonas aeruginosa* → 339 Gbp

Data structures:

→ Bloom Filter Trie

→ Sequence Bloom Tree (Solomon et al. 2016)
approximate membership only

Parameters:

→ $k = 36$, $cutoff = 3$

→ SBT with a false positive rate of 1 %

K-mer counting: 3.27 billion 36-mers extracted

EXPERIMENT

Table: Building time, RAM usage and disk size.

	BFT	SBT
Time	141 min	**76 min**
Peak of memory	4027 MB	**517 MB**
Disk size	**696 MB**	7969 MB
Compressed disk size	**439 MB**	7858 MB

EXPERIMENT

Table: Query time and RAM usage.

	BFT		SBT	
	Time	RAM	Time	RAM
Query 1	**0.82 min**	**786 MB**	304 min	11749 MB
Query 2	**0.56 min**	**786 MB**	46 min	4973 MB

Query 1: 10 million 36-mers sampled from the dataset
Query 2: ~7 million 36-mers sampled from one isolate

EXPERIMENT

Table: Branching query time and RAM usage.

	BFT	
	Time	RAM
Query 1	47 s	786 MB
Query 2	27 s	786 MB