Succinct k-mer Sets Using Subset Rank Queries on the Spectral BWT

Jarno Alanko, Simon Puglisi, Jaakko Vuohtoniemi

DSB 2022

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

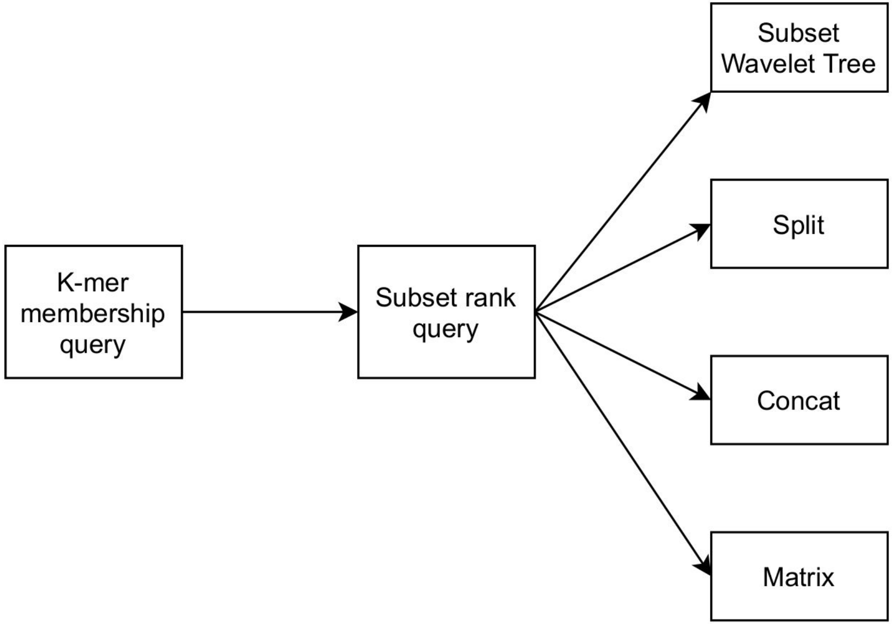
DALHOUSIE UNIVERSITY

Exact k-mer membership queries

- The BOSS data structure of Bowe et al. (WABI 2012)
  - Very compact
  - Complicated and slow
  - Used in: VARI, VG, Themisto
- Hashing
  - Not as small as BOSS
  - Fast
  - Used in: Bifrost, Pufferfish, Blight, FDBG, SSHash

This work

- The BOSS data structure can be seen as a particular implementation of a more general scheme.



Input: $$$TAGCAAGCACAGCATACAGA

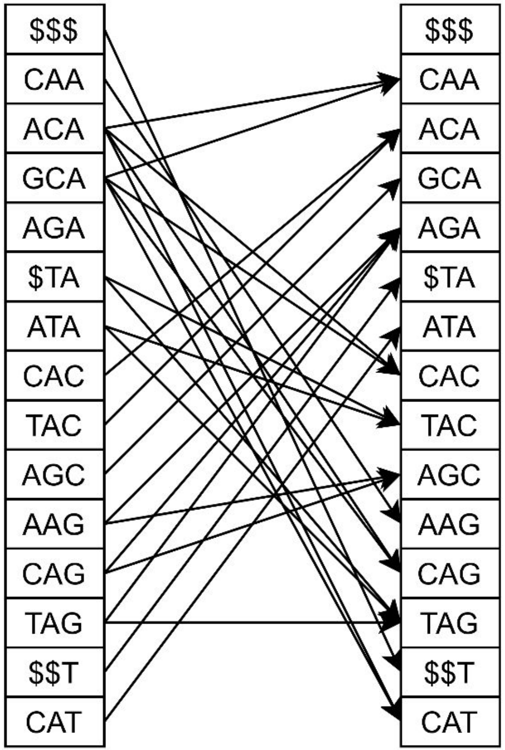
$$$
CAA
ACA
GCA
AGA
$TA
ATA
CAC
TAC
AGC
AAG
CAG
TAG
$$T
CAT

Input: $$$TAGCAAGCACAGCATACAGA

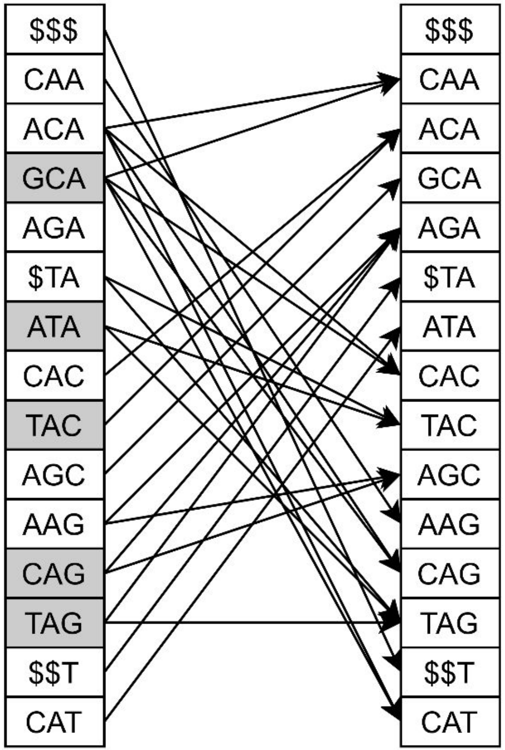
$$$
CAA
ACA
GCA
AGA
$TA
ATA
CAC
TAC
AGC
AAG
CAG
TAG
$$T
CAT

$$$
CAA
ACA
GCA
AGA
$TA
ATA
CAC
TAC
AGC
AAG
CAG
TAG
$$T
CAT

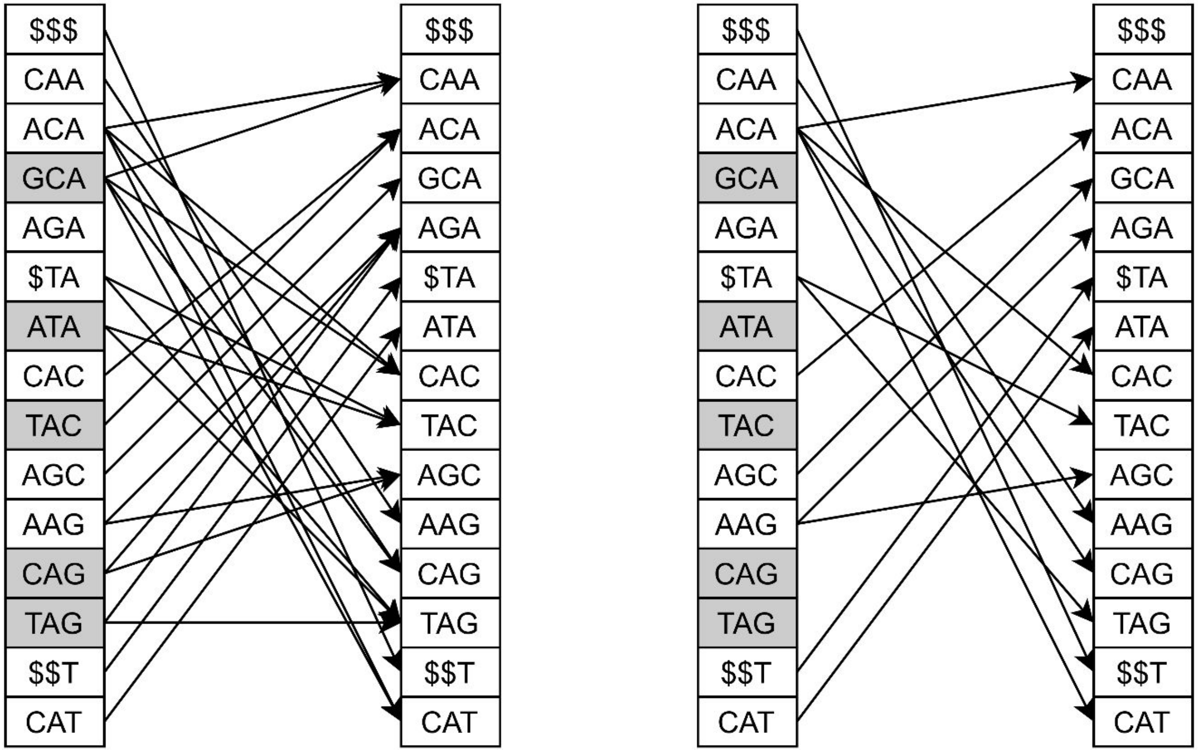
Input: $$$TAGCAAGCACAGCATACAGA



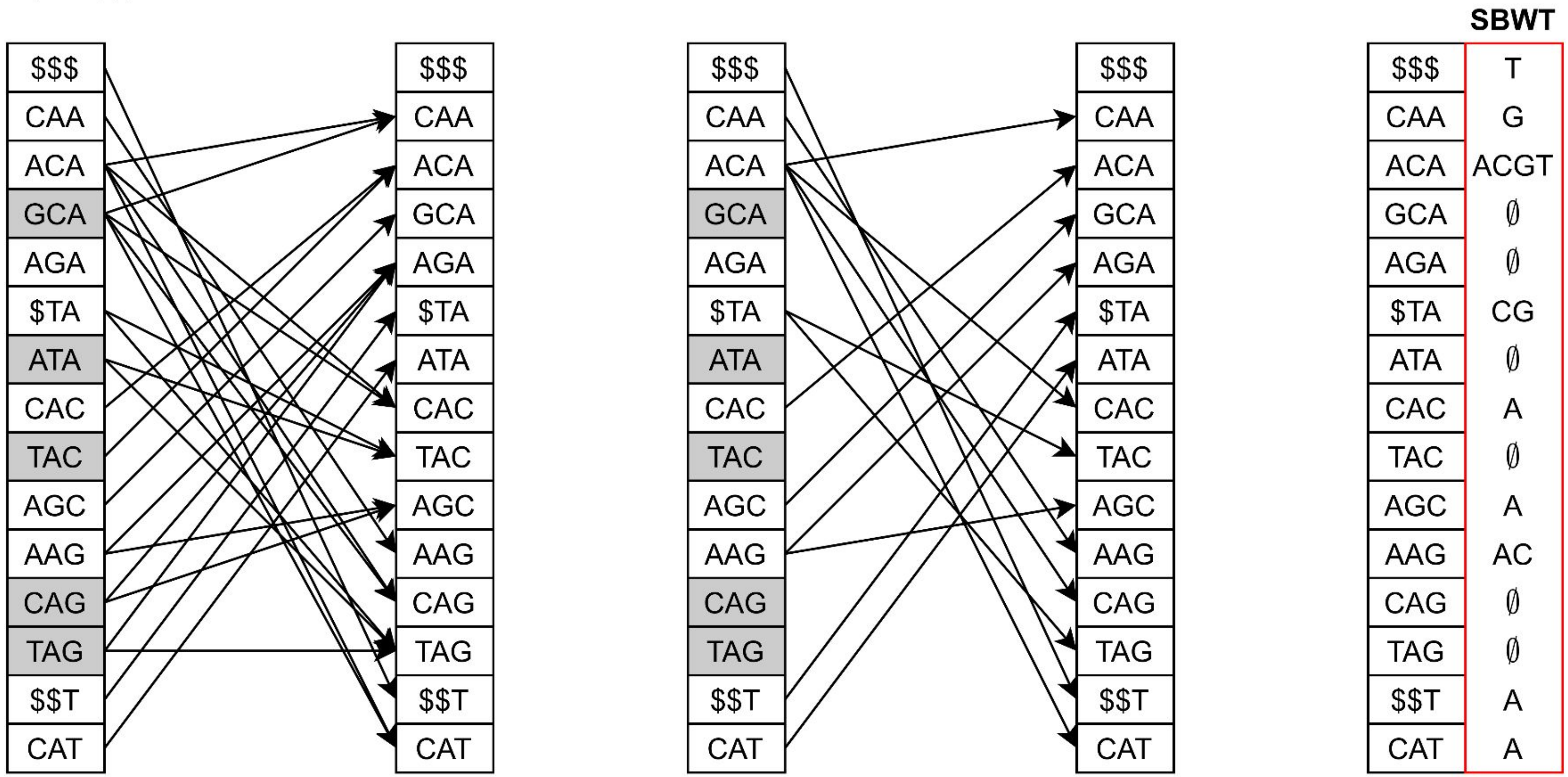
Input: $$$TAGCAAGCACAGCATACAGA



Input: $$$TAGCAAGCACAGCATACAGA



Input: $$$TAGCAAGCACAGCATACAGA



Algorithm 1 SBWT k-mer search query.
Input: k-mer S.
Output: The colexicographic rank of k-mer S in the underlying spectrum of the SBWT, or 0 if S is not in the spectrum.

```
function SEARCH(S):
    [ℓ, r] ← [1, n]
    for i = 1, ..., k do
        c ← S[i]
        ℓ ← 1 + C[c] + subsetrank_c(ℓ − 1) + 1
        r ← 1 + C[c] + subsetrank_c(r)
        if ℓ > r then
            return 0
    return ℓ.
```

Subset rank query

{T}, {G}, {ACGT}, ∅, ∅, {CG}, ∅, {A}, ∅, {A}, {AC}, ∅, ∅, {A}, {A}

$\text{SubsetRank}_A(9) = ?$

Subset rank query

{T}, {G}, {ACGT}, ∅, ∅, {CG}, ∅, {A}, ∅, {A}, {AC}, ∅, ∅, {A}, {A}

$\text{SubsetRank}_A(9) = ?$

Subset rank query

{T}, {G}, {ACGT}, ∅, ∅, {CG}, ∅, {A}, ∅, {A}, {AC}, ∅, ∅, {A}, {A}

$\text{SubsetRank}_A(9) = ?$

Subset rank query

{T}, {G}, {ACGT}, ∅, ∅, {CG}, ∅, {A}, ∅, {A}, {AC}, ∅, ∅, {A}, {A}

$\text{SubsetRank}_A(9) = 2$

Subset rank query

{T}, {G}, {ACGT}, ∅, ∅, {CG}, ∅, {A}, ∅, {A}, {AC}, ∅, ∅, {A}, {A}

$\text{SubsetRank}_A(9) = 2$

Lemma 1. *The SBWT supports k-mer membership queries in $O(kt)$ time, where t is the time for a subset rank query.*

	$$$	CAA	ACA	GCA	AGA	$TA	ATA	CAC	TAC	AGC	AAG	CAG	TAG	$$T	CAT
A	0	0	1	0	0	0	0	1	0	1	1	0	0	1	1
C	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0
G	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0
T	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0

	$$$	CAA	ACA	GCA	AGA	$TA	ATA	CAC	TAC	AGC	AAG	CAG	TAG	$$T	CAT
A	0	0	1	0	0	0	0	1	0	1	1	0	0	1	1
C	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0
G	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0
T	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Analysis

- The $\sigma \times n$ matrix has $n-1$ one-bits, so $Pr(1) = \frac{n-1}{\sigma n} \leq \frac{1}{\sigma}$
- We get $H_0(\text{Matrix}) \leq n(\log \sigma + 1/\ln 2) = O(n \log \sigma)$
- For $\sigma = 4$ we have $\left(-\frac{1}{4}\log(\frac{1}{4}) - \frac{3}{4}\log(\frac{3}{4})\right) \cdot 4 \approx 3.245$

Subset wavelet tree

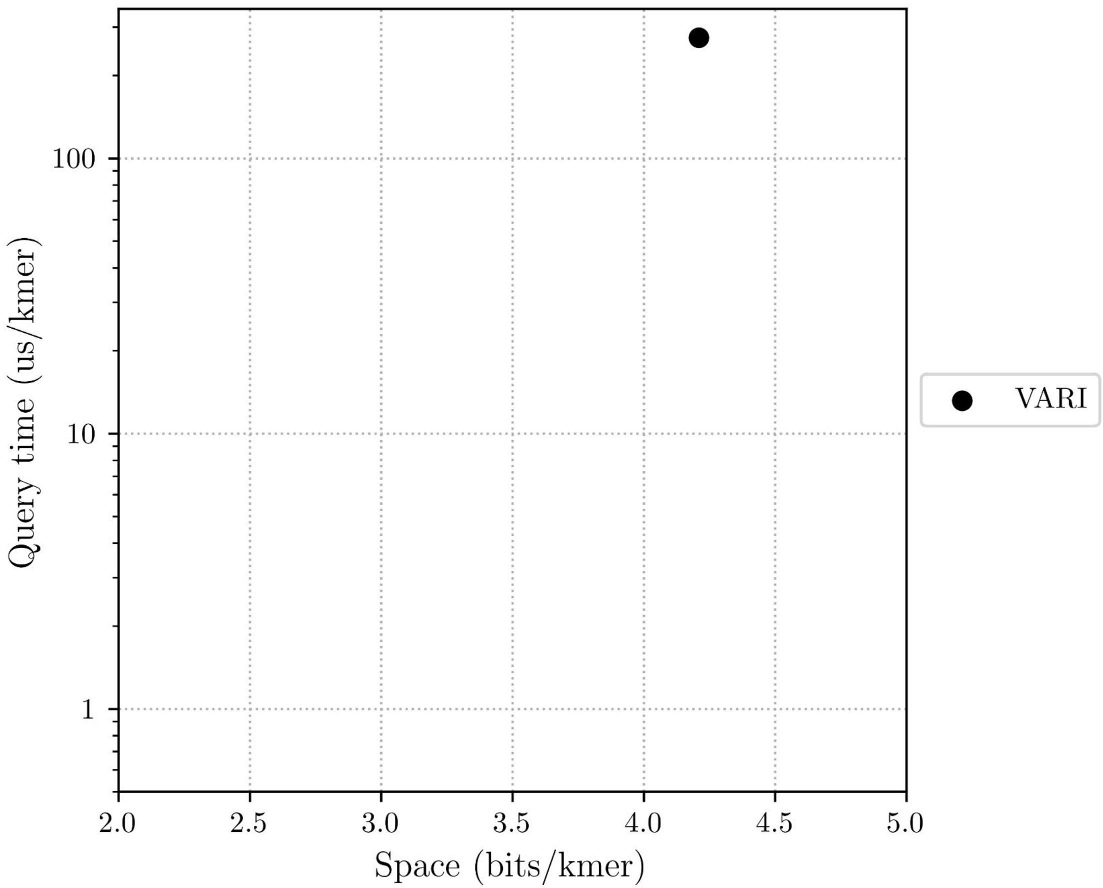
	T	G	ACGT			CG		A		A	AC			A	A
AC	0	0	1	0	0	1	0	1	0	1	1	0	0	1	1
GT	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0

	ACGT	CG	A	A	AC	A	A
A	1	0	1	1	1	1	1
C	1	1	0	0	1	0	0

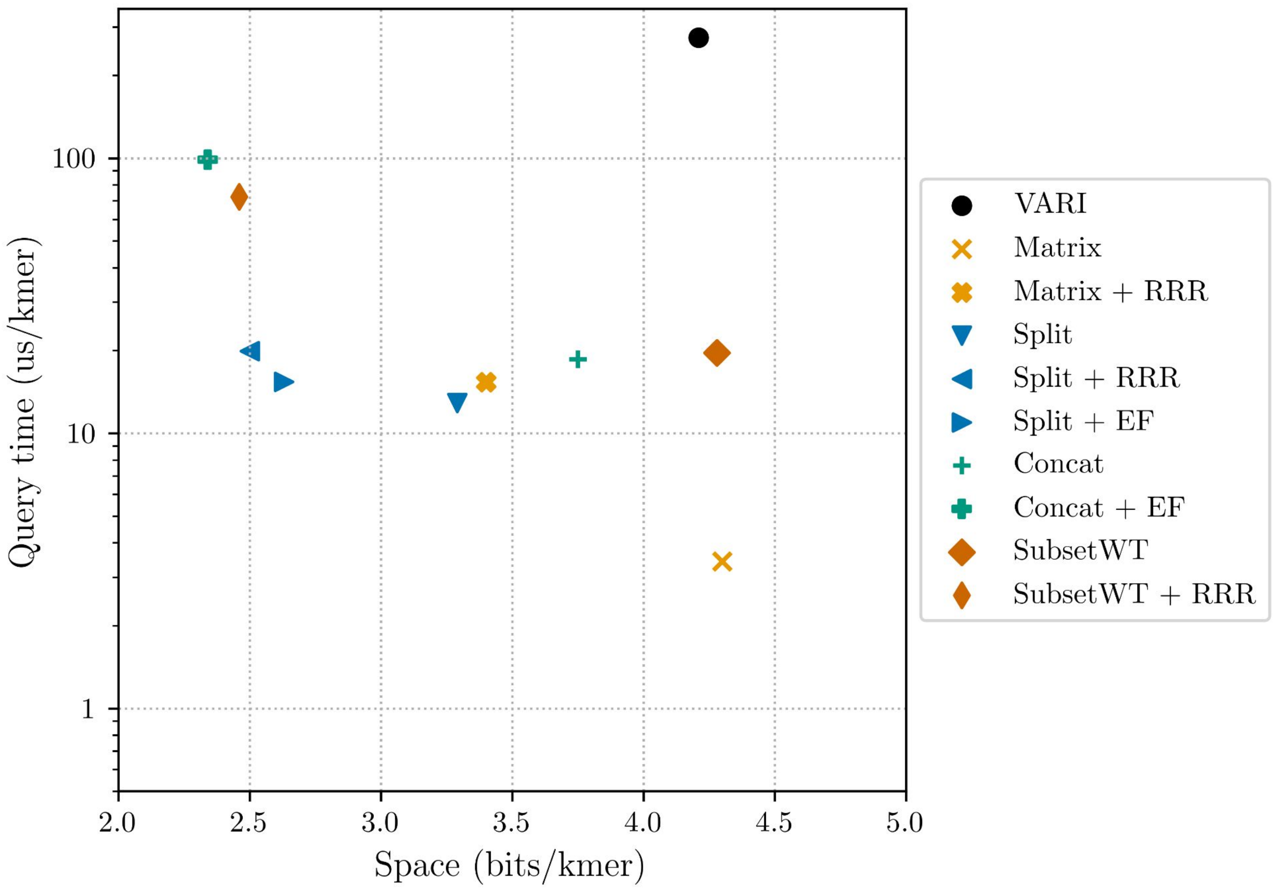
	T	G	ACGT	CG
G	0	1	1	1
T	1	0	1	0

Experiments

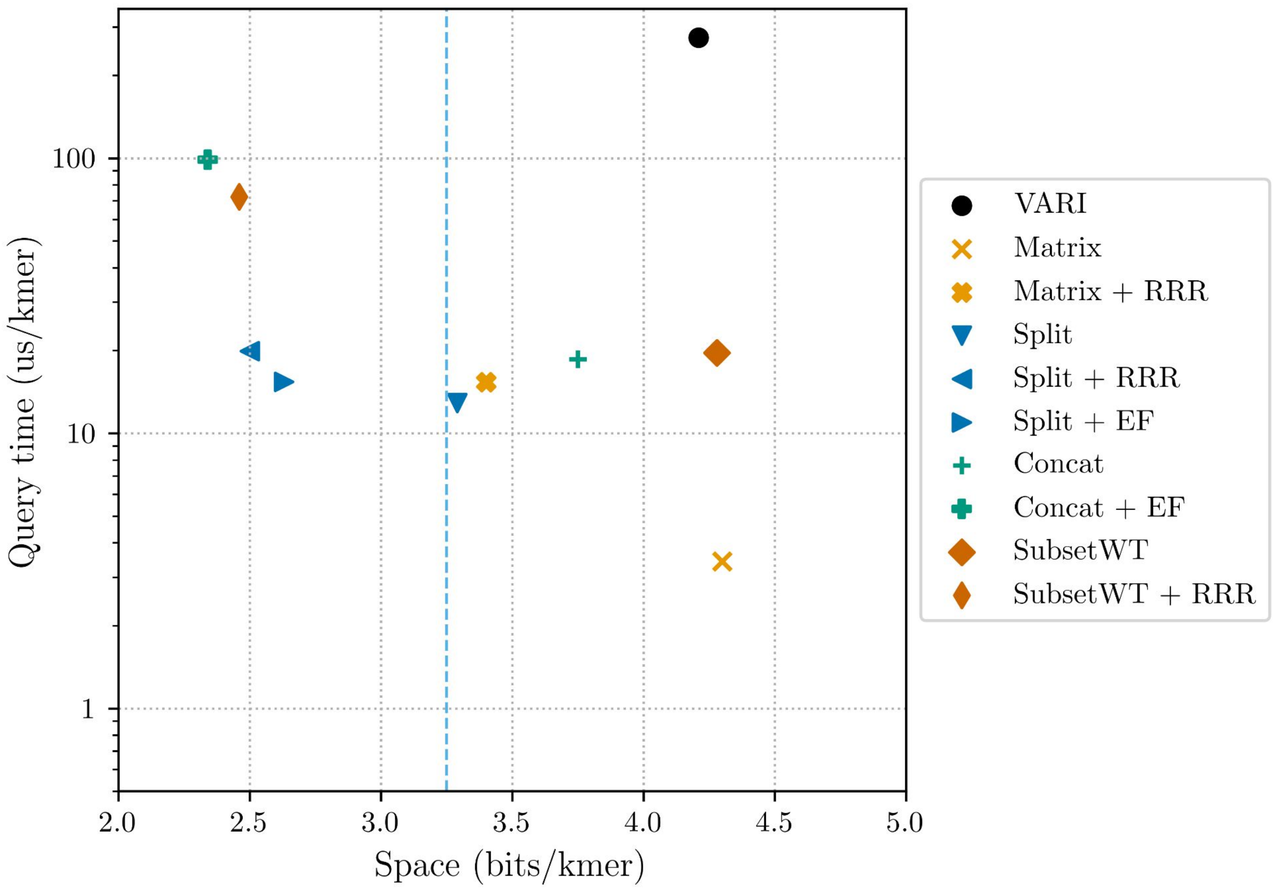
- Competitors
  - SBWT
  - VARI
  - SSHash
  - Bifrost
- Data
  - Viral pangenome (SARS-CoV-2)
  - Bacterial pangenome (E. coli)
  - Metagenome reads.
- k = 31



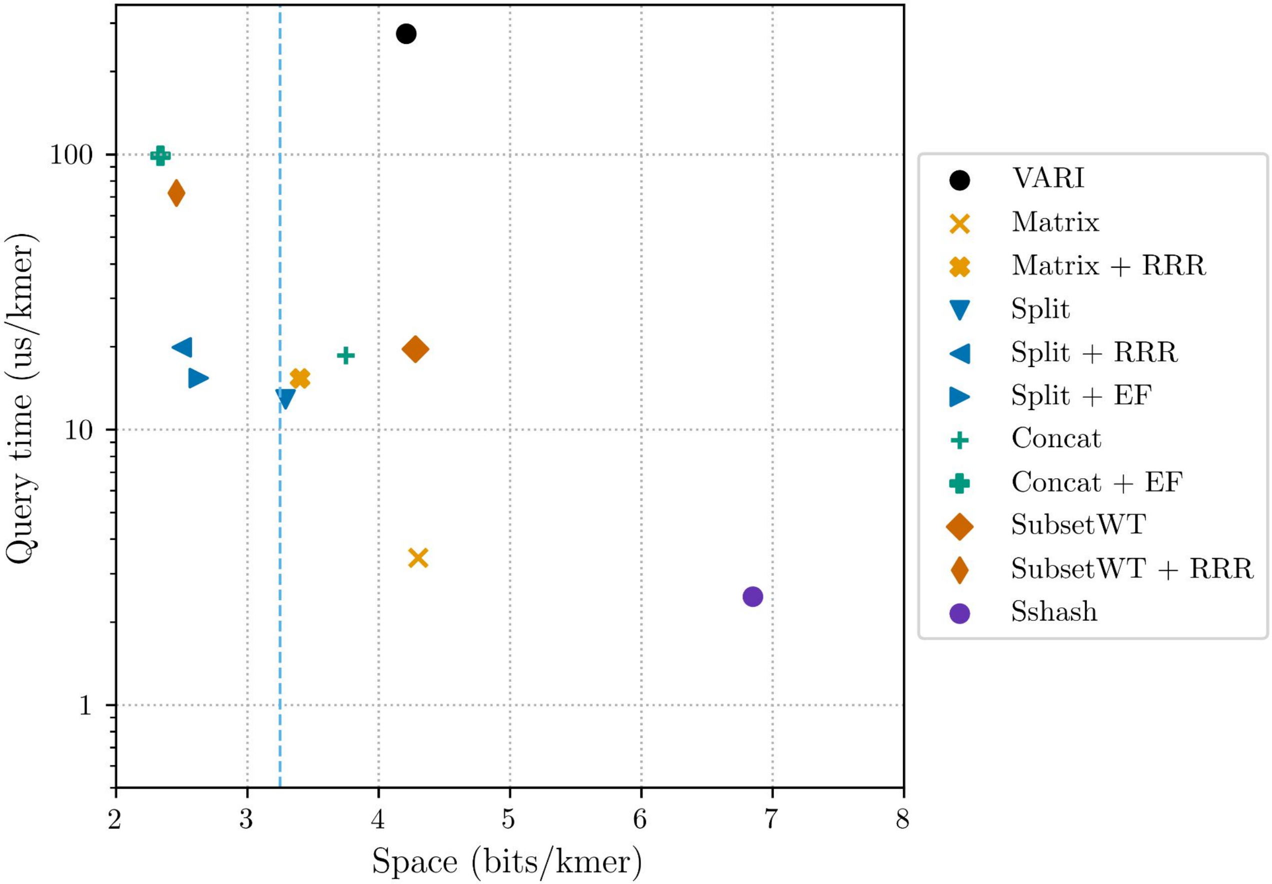
Query time (us/kmer)
100
10
1
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Space (bits/kmer)
VARI



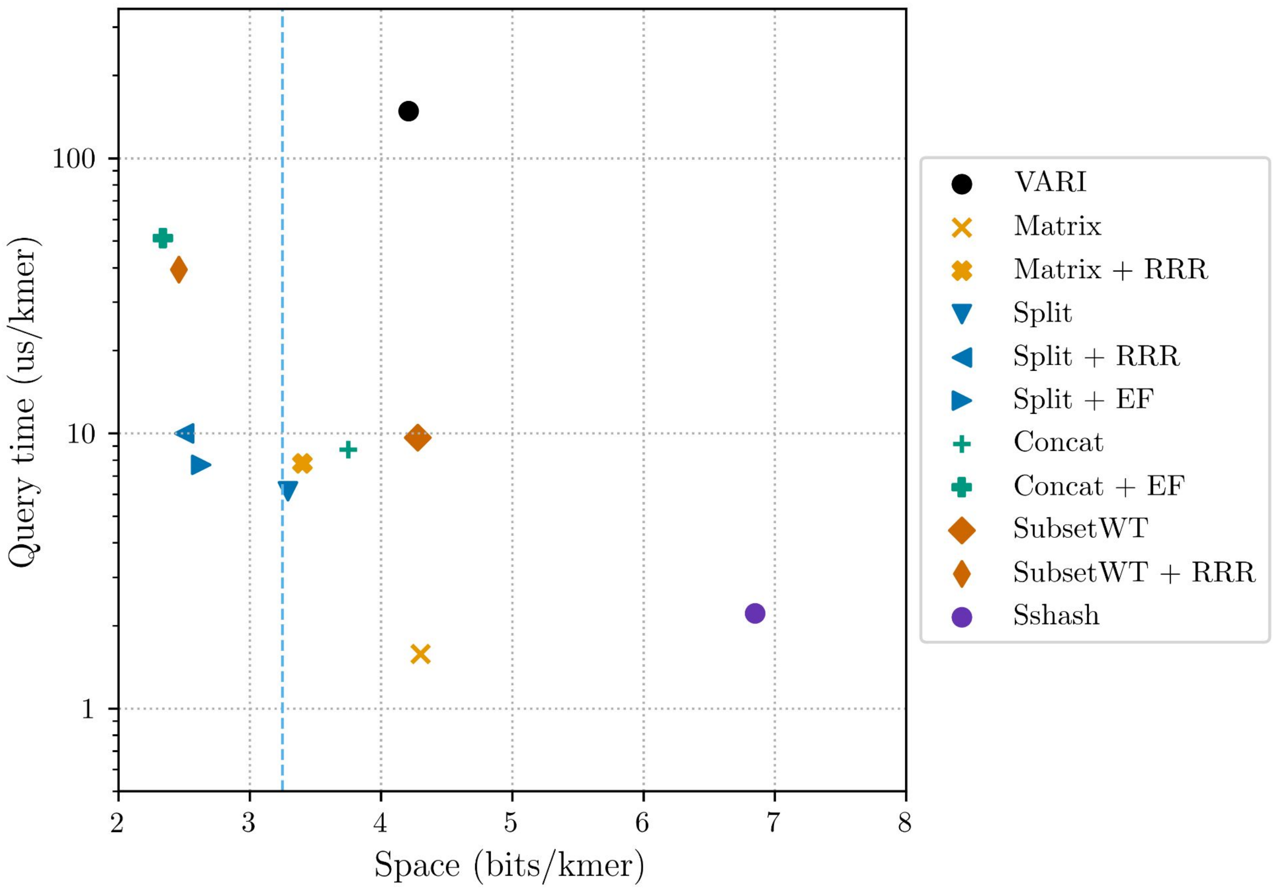
Query time (us/kmer)
100
10
1
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Space (bits/kmer)
VARI
Matrix
Matrix + RRR
Split
Split + RRR
Split + EF
Concat
Concat + EF
SubsetWT
SubsetWT + RRR



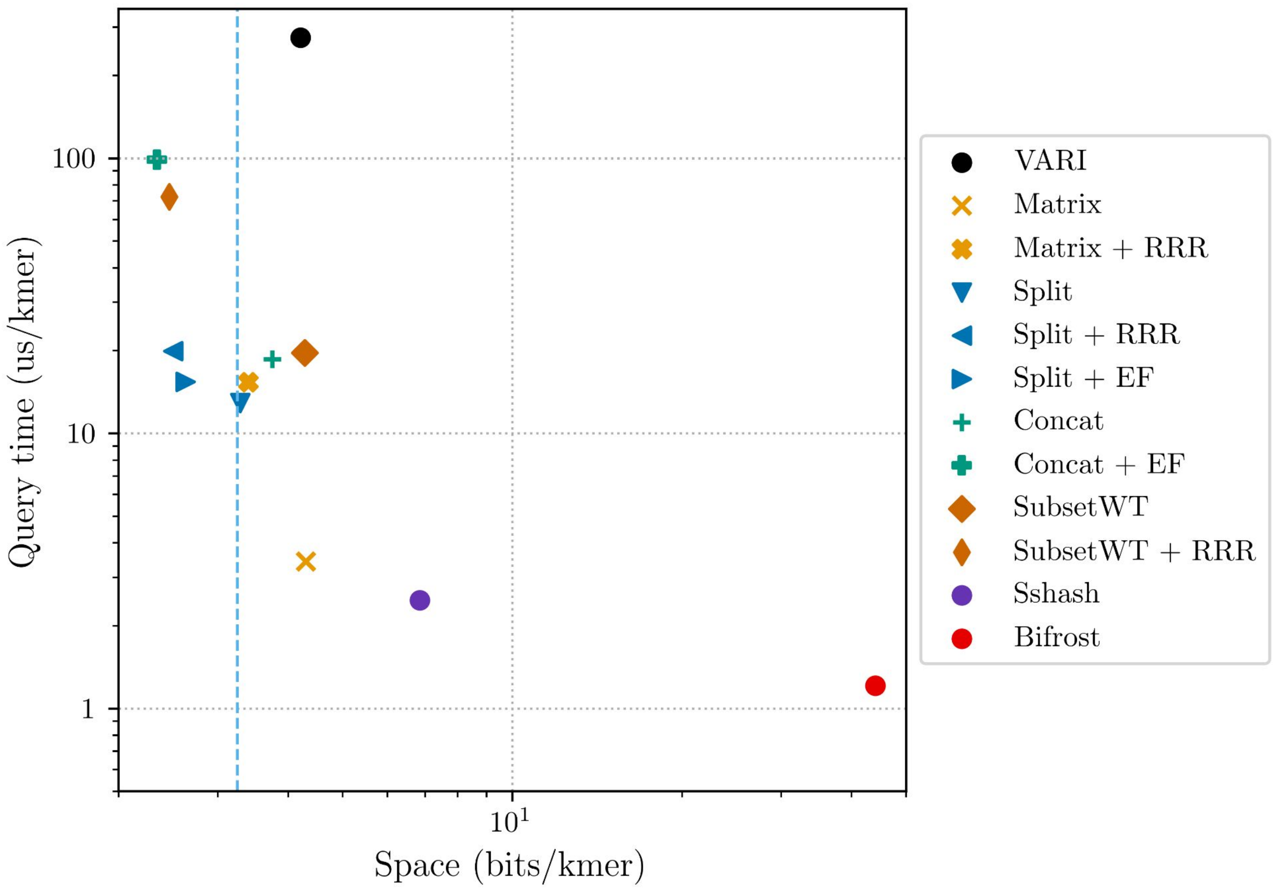
Query time (us/kmer)
100
10
1
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Space (bits/kmer)
VARI
Matrix
Matrix + RRR
Split
Split + RRR
Split + EF
Concat
Concat + EF
SubsetWT
SubsetWT + RRR



Query time (us/kmer)
100
10
1
2
3
4
5
6
7
8
Space (bits/kmer)
VARI
Matrix
Matrix + RRR
Split
Split + RRR
Split + EF
Concat
Concat + EF
SubsetWT
SubsetWT + RRR
Sshash



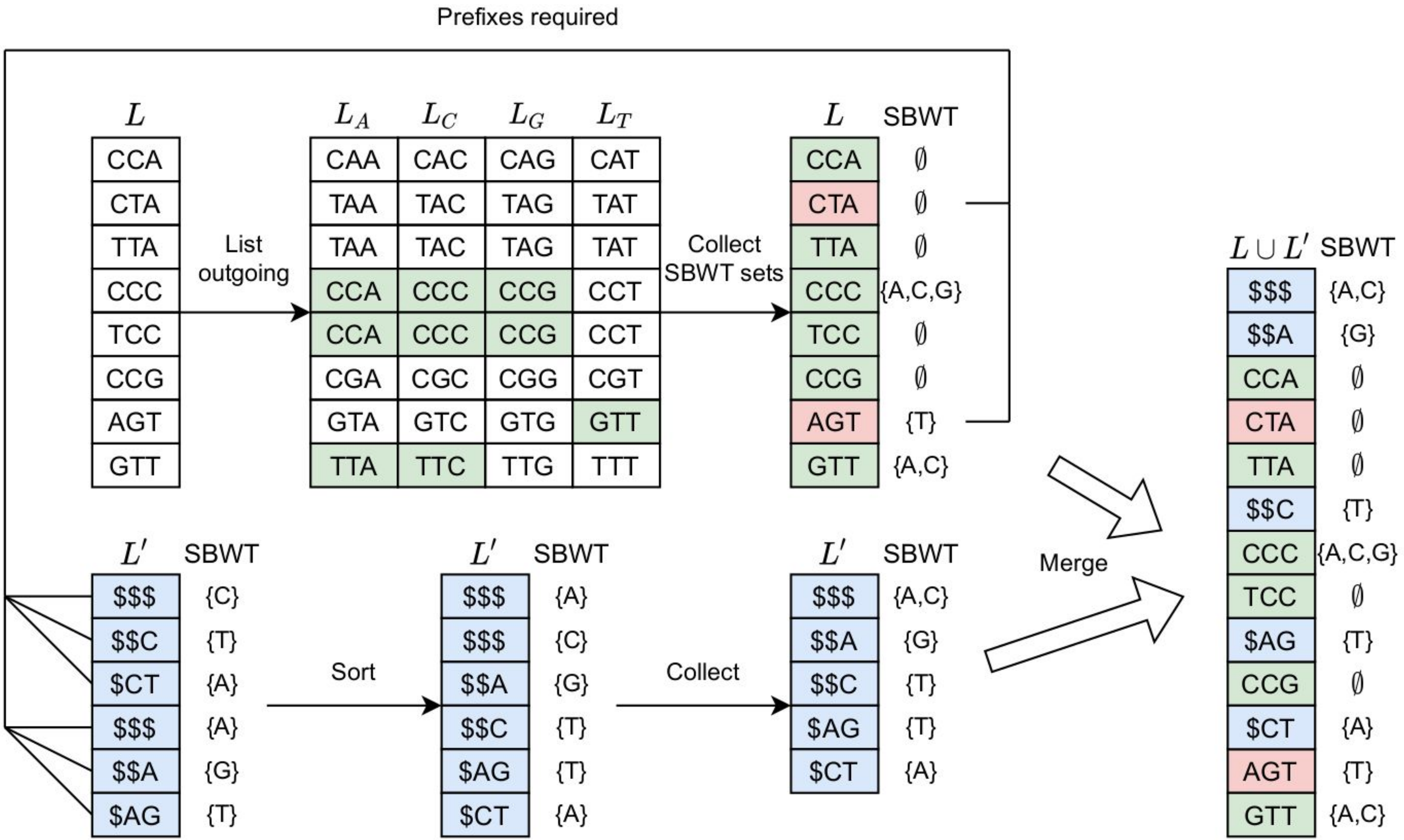
Query time (us/kmer)
100
10
1
Space (bits/kmer)
2
3
4
5
6
7
8
VARI
Matrix
Matrix + RRR
Split
Split + RRR
Split + EF
Concat
Concat + EF
SubsetWT
SubsetWT + RRR
Sshash


Query time (us/kmer)
100
10
1
Space (bits/kmer)
10^1
VARI
Matrix
Matrix + RRR
Split
Split + RRR
Split + EF
Concat
Concat + EF
SubsetWT
SubsetWT + RRR
Sshash
Bifrost

Conclusion

- BWT-based methods are competitive again
- Things omitted from this talk:
  - Construction
  - Streaming queries
  - Entropy optimization

Thank you for your attention


Prefixes required
L
CCA
CTA
TTA
CCC
TCC
CCG
AGT
GTT
List outgoing
L_A L_C L_G L_T
CAA CAC CAG CAT
TAA TAC TAG TAT
TAA TAC TAG TAT
CCA CCC CCG CCT
CCA CCC CCG CCT
CGA CGC CGG CGT
GTA GTC GTG GTT
TTA TTC TTG TTT
Collect SBWT sets
L SBWT
CCA ∅
CTA ∅
TTA ∅
CCC {A,C,G}
TCC ∅
CCG ∅
AGT {T}
GTT {A,C}
L' SBWT
$$$ {C}
$$C {T}
$CT {A}
$$$ {A}
$$A {G}
$AG {T}
Sort
L' SBWT
$$$ {A}
$$$ {C}
$$A {G}
$$C {T}
$AG {T}
$CT {A}
Collect
L' SBWT
$$$ {A,C}
$$A {G}
$$C {T}
$AG {T}
$CT {A}
Merge
$L \cup L'$ SBWT
$$$ {A,C}
$$A {G}
CCA ∅
CTA ∅
TTA ∅
$$C {T}
CCC {A,C,G}
TCC ∅
$AG {T}
CCG ∅
$CT {A}
AGT {T}
GTT {A,C}

Split representation

	$$$	CAA	ACA	GCA	AGA	$TA	ATA	CAC	TAC	AGC	AAG	CAG	TAG	$$T	CAT
M	0	0	1	0	0	0	0	1	0	1	1	0	0	1	1
	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0
	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0
	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
B	0	0	1	0	1	0	0	0	1	0	0	0	1	0	0

	ACA	AGA	TAC	TAG
$\mathbf{M^+}$	1	0	0	0
	1	0	0	0
	1	0	0	0
	0	0	0	0

	$$$	CAA	GCA	$TA	ATA	CAC	AGC	AAG	CAG	$$T	CAT
$\mathbf{M^-}$	0	0	0	0	0	1	1	1	0	1	1
	0	0	0	1	0	0	0	0	1	0	0
	0	1	0	0	1	0	0	0	0	0	0
	1	0	1	0	0	0	0	0	0	0	0
W	T	G	T	C	G	A	A	A	C	A	A

Concatenated representation

	$$$	CAA	ACA	GCA	AGA	$TA	ATA	CAC	TAC	AGC	AAG	CAG	TAG	$$T	CAT
A	0	0	1	0	0	0	0	1	0	1	1	0	0	1	1
C	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0
G	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0
T	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0

T	C	A	C	G	T	$	$	C	G	$	A	$	A	A	C	$	$	A	A
1	1	1	0	0	0	1	1	1	0	1	1	1	1	1	0	1	1	1	1