Pangenomic FM-indexes

Travis Gagie
Dalhousie University

et al.

Data Compression Conference

March 23rd, 2023

DNA alignment

WGS

genome	G	A	T	A	C	A	T
read 1	G	A	T	A			
read 2		A	T	A	C		
read 3				A	C	A	T

WGS

reference	G	A	T	T	A	C	A	T
read 1	G	A	T	-	A			
read 2		A	-	T	A	C		
read 3					A	C	A	T
output	G	A	T	-	A	C	A	T

reference bias

reference	G	A	T	T	A	C	A	T
read 1	G	A	T	-	A			
read 2		A	-	T	A	G		
read 3					A	G	A	T
output	G	A	T	-	A	C	A	T

reference bias

To the scientists' puzzlement, however, the boy's sequence showed no sign of the mutation in the gene known to cause Baratela Scott, called XYLT1. Nor did the DNA of the next boy with the disorder, or the next. As they tried to compare the boys' DNA sequences to the reference genome, it was like trying to check a spelling in a Webster's from which a prankster had torn handfuls of pages. Many pieces of the boys' genomes, called short reads,"weren't in the reference genome at all," ... There was no way to check them for disease-causing misspellings.

— Sharon Begley, *Stat News*, March 11th, 2019

reference bias

This bias limits the kind of genetic variation that can be detected, leaving some patients without diagnoses and potentially without proper treatment. What is more, people who share less ancestry with the man from Buffalo will probably benefit less from the incoming era of precision medicine, which promises to tailor healthcare to individuals.

[O]ur understanding of diversity within populations of European descent is now so good that we can start to use it for precision medicine. But for other populations, "We do not have the same kind of data ... [This] is going to increase healthcare disparities above and beyond what they are today." ... [A] huge new project is offering a different solution with the aim to represent global diversity: a human pangenome.

— Ida Emilie Steinmark, *Guardian*, January 29th, 2023

— webpage of the *Human Pangenome Reference Consortium*

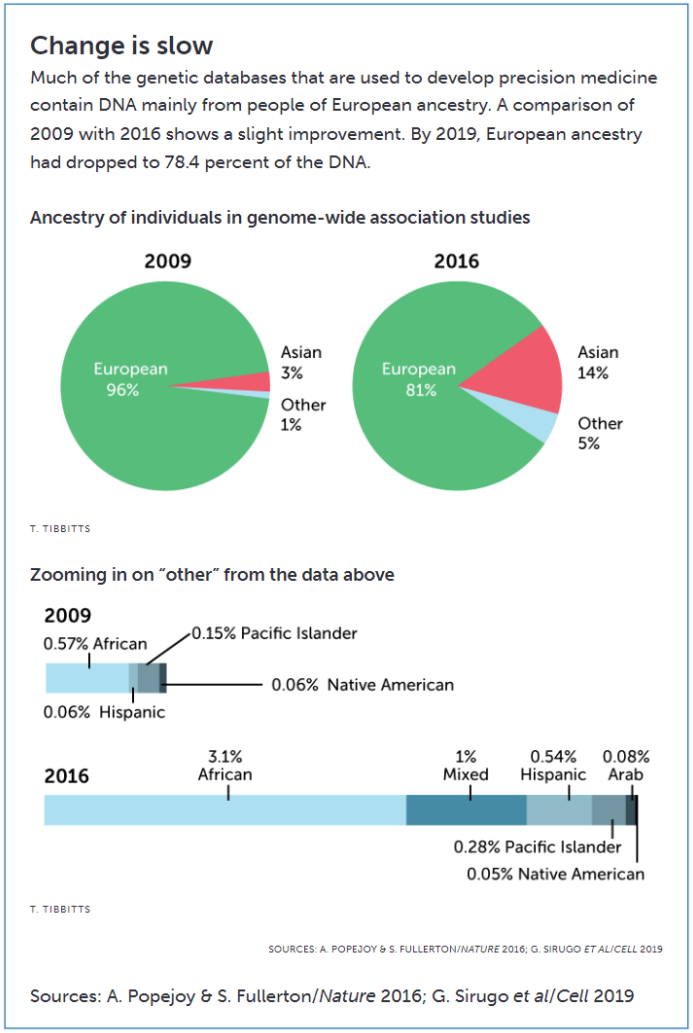
EDI

Around 38,000 cancer patients in England and approximately 154,000 patients in the US are initiated on fluoropyrimidine-based [chemotherapy] treatments every year [but] between 20% and 30% of the people who receive these drugs require lower doses, because their bodies struggle to process them. If given the standard dose, they experience reactions which can vary from severe to fatal.

In recent years, genetic-sequencing studies have started to get to the bottom of why some people react so badly ... The only problem is that these studies were done entirely on white people ... "Ethnic minority patients will usually be given conventional doses of the drugs ... Some of these patients will carry other ethnic-specific variants which also affect their ability to metabolise these drugs, but we do not currently genotype for those, largely because we do not know."

— David Cox, *BBC*, February 28th, 2023

EDI



— Tina Hesman Saey, *Science News*, March 21st, 2021

pangenome graphs

[T]he project is not just about sequencing more diverse data. "We need to come up with a better data structure to encode that information," ... That data structure is called a genome graph. In contrast to the current reference, which is just a long string of letters, the genome graph shows variation between genomes as detours on an otherwise shared path. That will enable researchers and doctors to map short reads to the version of the path that best fits their sample.

The natural question is: how does one choose who gets to represent the world? ... 350 people might do a better job of representing the world than one person, but "[the consortium] have made some choices about groups ... Who did they sample? Who did they not sample?" As long as the reference contains only a subset, arguably someone will not make the cut.

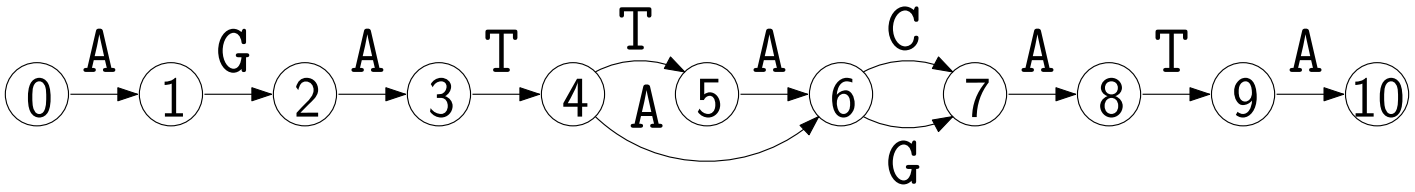
— Ida Emilie Steinmark, *Guardian*, January 29th, 2023

pangenome graphs

dataset:

- GATTACAT
- AGATACAT
- GATACAT
- GATTAGAT
- GATTAGATA

pangenome graph:

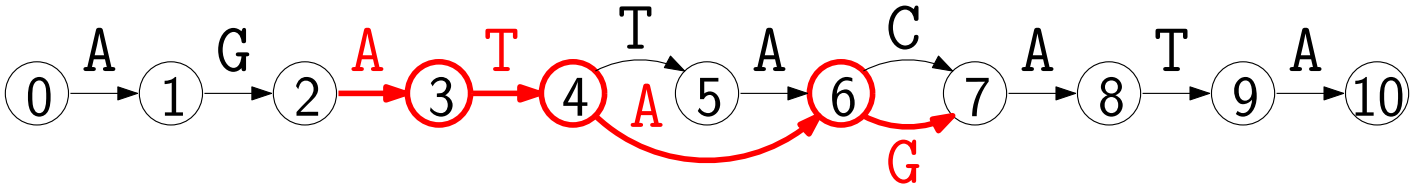


pangenome graphs

dataset:

- GATTACAT
- AGATACAT
- GATACAT
- GATTAGAT
- GATTAGATA

pangenome graph:

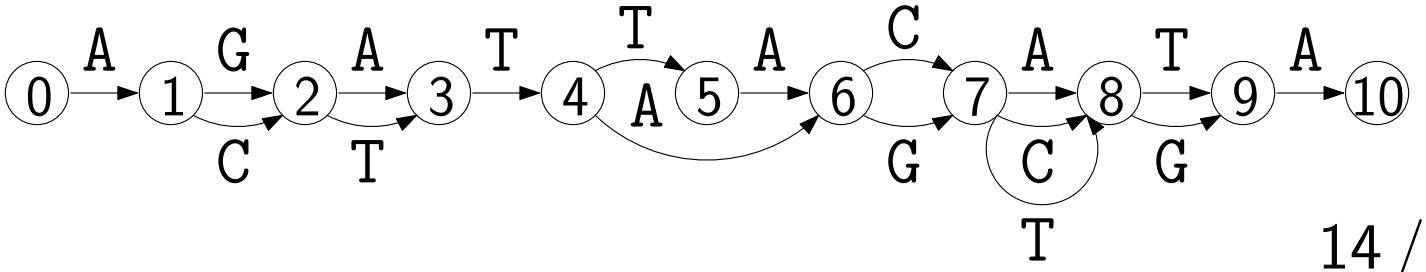


pangenome graphs

dataset:

- GATTACAT
- AGATACAT
- GATACAT
- GATTAGAT
- GATTAGATA
- CATTACAT
- GTTAGAT
- GATTCCATA
- GATTACAGA

pangenome graph:

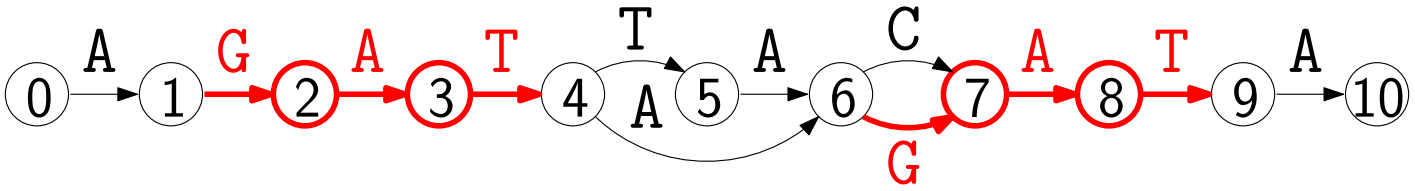


pangenome graphs

dataset:

- GATTACAT
- AGATACAT
- GATACAT
- GATTAGAT
- GATTAGATA

pangenome graph:

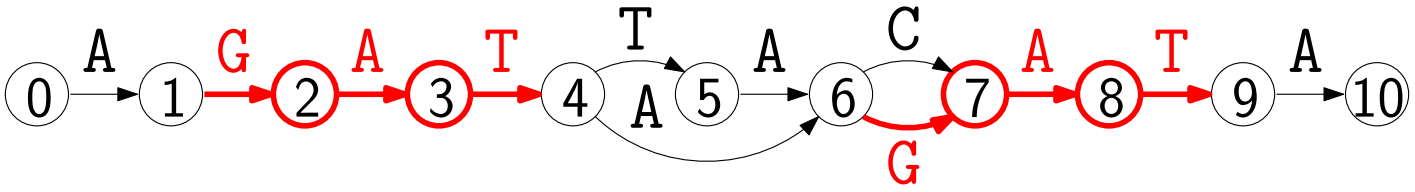


pangenome graphs

dataset:

- GATTACAT
- AGATACAT
- GATACAT
- GATTAGAT
- GATTAGATA

pangenome graph:



Theorem (sketch)

Given a pangenomic dataset of total length n and a pangenome graph for it, we can store the dataset in reasonable space — that is, much *less than n — such that when given a read of length m, we can*

1. *find good seeds in the read with respect to the dataset in $O(m \log n)$ time*
2. *for each seed, list all the distinct vertices in the graph where we start processing occurrences in the dataset of that seed, in constant time per vertex listed.*

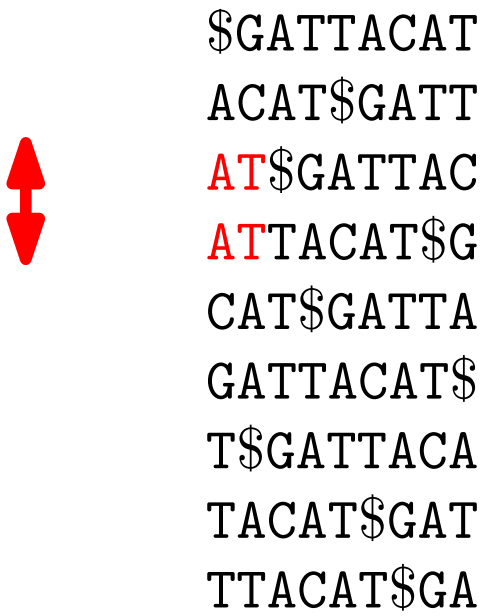
FM-indexes

data structure

$GATTACAT
ACAT$GATT
AT$GATTAC
ATTACAT$G
CAT$GATTA
GATTACAT$
T$GATTACA
TACAT$GAT
TTACAT$GA

0	1	2	3	4	5	6	7	8
G	A	T	T	A	C	A	T	$

counting



0	1	2	3	4	5	6	7	8
G	A	T	T	A	C	A	T	$

locating

8	$GATTACAT
4	ACAT$GATT
6	AT$GATTAC
1	ATTACAT$G
5	CAT$GATTA
0	GATTACAT$
7	T$GATTACA
3	TACAT$GAT
2	TTACAT$GA

0	1	2	3	4	5	6	7	8
G	A	T	T	A	C	A	T	$

MEM-finding

A substring $P[i..j]$ of P is a *maximal exact match* (MEM) between P and T if

- $P[i..j]$ occurs in T
- neither $P[i-1..j]$ nor $P[i..j+1]$ occurs in T.

The MEMs of `GATAGAT` with respect to `GATTACAT` are `GAT` (twice) and `TA`, but with respect to

`GATTACAT$AGATACAT$GATACAT$GATTAGAT$GATTAGATA$`

they are `GATA` and `TAGAT`.

MEM-finding

Theorem

We can store a text $T[1..n]$ over $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ in $2n + o(n)$ bits such that when given a pattern $P[1..m]$, we can

1. *find the MEMs of P with respect to T in $O(m \log^{1+\epsilon} n)$ time*
2. *list their occurrences' positions in T in $O(\log^{1+\epsilon} n)$ time per occurrence.*

Why not just scale this up for pangenomic alignment?

1. $2n + o(n)$ is way too big when n is huge (even just $o(n)$ may be too big!)
2. the MEMs could occur too often in the dataset

scaling up

	fast counting	fast locating	fast MEM-finding	scalable
FM-index (*JACM*, 2005)	yes	yes	yes	no
RLCSA (*JCB*, 2010)	yes	no	no	yes
Toehold Lemma (*Algorithmica*, 2018)	yes	1	no	yes
r-index (*JACM*, 2020)	yes	yes	no	yes
MONI (*JCB*, 2022)	no	yes	yes	yes

	fast counting	fast locating	fast MEM-finding	scalable
FM-index (*JACM*, 2005)	yes	yes	yes	no
RLCSA (*JCB*, 2010)	yes	no	no	yes
Toehold Lemma (*Algorithmica*, 2018)	yes	1	no	yes
r-index (*JACM*, 2020)	yes	yes	no	yes
MONI (*JCB*, 2022)	no	yes	yes	yes

RLCSA

G A T T A C A T $ A G A T A C A T $ G A T A C A T $ G A T T A G A T $ G A T T A G A T A $

BWT	context
T	$AGATACA
T	$GATACA
T	$GATTACA
T	$GATTAGA
A	$GATTAGAT
T	A$GATTAGA
T	ACAT$AGA
T	ACAT$GA
T	ACAT$GAT
T	AGAT$GAT
T	AGATA$GAT
$	AGATACAT
C	AT$AGATA
C	AT$GATA
C	AT$GATTA

BWT	context
G	AT$GATTA
G	ATA$GATTA
G	ATACAT$A
G	ATACAT$
G	ATTACAT$
G	ATTAGAT$
G	ATTAGATA$
A	CAT$AGAT
A	CAT$GAT
A	CAT$GATT
A	GAT$GATT
A	GATA$GATT
A	GATACAT$
$	GATACAT
$	GATTACAT

BWT	context
$	GATTAGAT
$	GATTAGATA
A	T$AGATAC
A	T$GATAC
A	T$GATTAC
A	T$GATTAG
A	TA$GATTAG
A	TACAT$AG
A	TACAT$G
T	TACAT$GA
T	TAGAT$GA
T	TAGATA$GA
A	TTACAT$G
A	TTAGAT$G
A	TTAGATA$G

RLCSA

GATTACAT$ AGATACAT$ GATACAT$ GATTAGAT$ GATTAGATA$

BWT	context
T	$AGATACA
T	$GATACA
T	$GATTACA
T	$GATTAGA
A	$GATTAGAT
T	A$GATTAGA
T	ACAT$AGA
T	ACAT$GA
T	ACAT$GAT
T	AGAT$GAT
T	AGATA$GAT
$	AGATACAT
C	AT$AGATA
C	AT$GATA
C	AT$GATTA

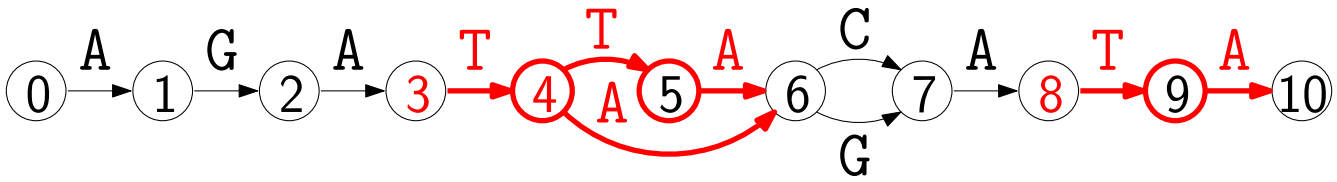
BWT	context
G	AT$GATTA
G	ATA$GATTA
G	ATACAT$A
G	ATACAT$
G	ATTACAT$
G	ATTAGAT$
G	ATTAGATA$
A	CAT$AGAT
A	CAT$GAT
A	CAT$GATT
A	GAT$GATT
A	GATA$GATT
A	GATACAT$
$	GATACAT
$	GATTACAT

BWT	context
$	GATTAGAT
$	GATTAGATA
A	T$AGATAC
A	T$GATAC
A	T$GATTAC
A	T$GATTAG
A	TA$GATTAG
A	TACAT$AG
A	TACAT$G
T	TACAT$GA
T	TAGAT$GA
T	TAGATA$GA
A	TTACAT$G
A	TTAGAT$G
A	TTAGATA$G

RLCSA

G A T T A C A T $ A G A T A C A T $ G A T A C A T $ G A T T A G A T $ G A T T A G A T A $

tag	context	tag	context	tag	context
9	$AGATACA	7	AT$GATTA	1	GATTAGAT
9	$GATACA	7	ATA$GATTA	1	GATTAGATA
9	$GATTACA	2	ATACAT$A	8	T$AGATAC
9	$GATTAGA	2	ATACAT$	8	T$GATAC
10	$GATTAGAT	2	ATTACAT$	8	T$GATTAC
9	A$GATTAGA	2	ATTAGAT$	8	T$GATTAG
4	ACAT$AGA	2	ATTAGATA$	8	TA$GATTAG
4	ACAT$GA	6	CAT$AGAT	3	TACAT$AG
5	ACAT$GAT	6	CAT$GAT	3	TACAT$G
5	AGAT$GAT	6	CAT$GATT	4	TACAT$GA
5	AGATA$GAT	6	GAT$GATT	4	TAGAT$GA
0	AGATACAT	6	GATA$GATT	4	TAGATA$GA
7	AT$AGATA	1	GATACAT$	3	TTACAT$G
7	AT$GATA	1	GATACAT	3	TTAGAT$G
7	AT$GATTA	1	GATTACAT	3	TTAGATA$G



RLCSA

Theorem

We can store a text $T[1..n]$ over $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}, \$\}$ in $O(r)$ space, where r is the number of runs in BWT*, such that when given a pattern $P[1..m]$ we can count the occurrences of P in T in $O(m)$ time.*

Corollary

We can store T in $O(r + t)$ space, where t is the numbers of runs in tag*, such that when given P we can list all the distinct vertices where we start processing occurrences of P in T, in $O(m)$ time plus constant time per vertex listed.*

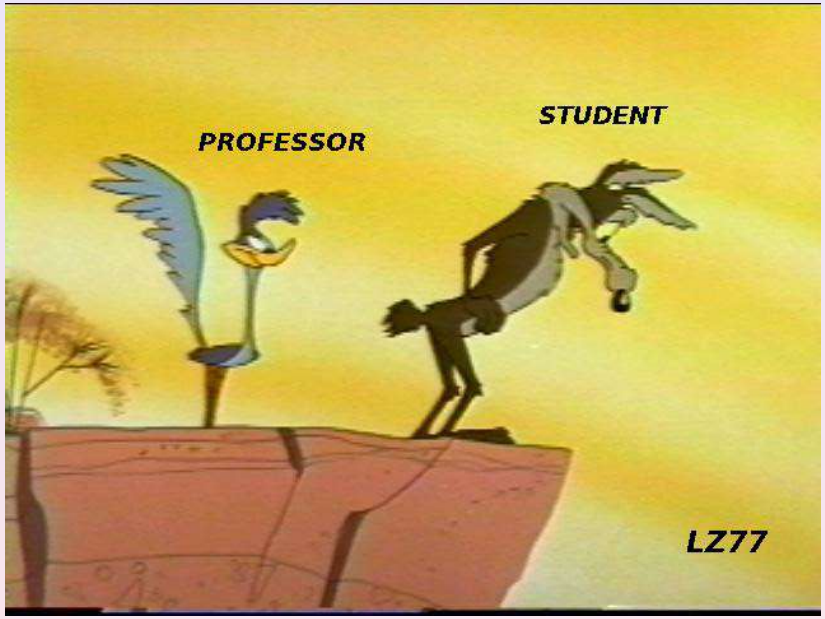
...but no MEMs!

A New Challenge: Full Compression

Application Scenario: Computational Biology

- Sequencing genomes is becoming cheap and fast.
- We are not far from the day where we will have databases of thousands or millions of genomes.
- The applications of such a database are unimaginable, BUT...
- 1 million uncompressed genomes $\Longrightarrow$ about 3 petabytes
- a classical suffix tree $\Longrightarrow$ 30 petabytes
- compressed with current self-indexes $\Longrightarrow$ 750 terabytes
- just the sublinear part we mentioned $\Longrightarrow$ 200 terabytes
- Overall, the best we can do requires close to 1 petabyte.

Indexing LZ77: The Next Step in Self-Indexing



Gonzalo Navarro
Department of Computer Science, University of Chile
gnavarro@dcc.uchile.cl

	fast counting	fast locating	fast MEM-finding	scalable
FM-index (*JACM*, 2005)	yes	yes	yes	no
RLCSA (*JCB*, 2010)	yes	no	no	yes
Toehold Lemma (*Algorithmica*, 2018)	yes	1	no	yes
r-index (*JACM*, 2020)	yes	yes	no	yes
MONI (*JCB*, 2022)	no	yes	yes	yes

Toehold Lemma

GATTACAT$ AGATACAT$ GATACAT$ GATTAGAT$ GATTAGAT$

BWT	context
T	$AGATACA
T	$GATACA
T	$GATTACA
T	$GATTAGA
A	$GATTAGAT
T	A$GATTAGA
T	ACAT$AGA
T	ACAT$GA
T	ACAT$GAT
T	AGAT$GAT
T	AGATA$GAT
$	AGATACAT
C	AT$AGATA
C	AT$GATA
C	AT$GATTA

BWT	context
G	AT$GATTA
G	ATA$GATTA
G	ATACAT$A
G	ATACAT$
G	ATTACAT$
G	ATTAGAT$
G	ATTAGATA$
A	CAT$AGAT
A	CAT$GAT
A	CAT$GATT
A	GAT$GATT
A	GATA$GATT
A	GATACAT$
$	GATACAT
$	GATTACAT

BWT	context
$	GATTAGAT
$	GATTAGATA
A	T$AGATAC
A	T$GATAC
A	T$GATTAC
A	T$GATTAG
A	TA$GATTAG
A	TACAT$AG
A	TACAT$G
T	TACAT$GA
T	TAGAT$GA
T	TAGATA$GA
A	TTACAT$G
A	TTAGAT$G
A	TTAGATA$G

Toehold Lemma

G A T T A C A T $ A G A T A C A T $ G A T A C A T $ G A T T A G A T $ G A T T A G A T A $

BWT	context
T	$AGATACA
T	$GATACA
T	$GATTACA
T	$GATTAGA
A	$GATTAGAT
T	A$GATTAGA
T	ACAT$AGA
T	ACAT$GA
T	ACAT$GAT
T	AGAT$GAT
T	AGATA$GAT
$	AGATACAT
C	AT$AGATA
C	AT$GATA
C	AT$GATTA

BWT	context
G	AT$GATTA
G	ATA$GATTA
G	ATACAT$A
G	ATACAT$
G	ATTACAT$
G	ATTAGAT$
G	ATTAGATA$
A	CAT$AGAT
A	CAT$GAT
A	CAT$GATT
A	GAT$GATT
A	GATA$GATT
A	GATACAT$
$	GATACAT
$	GATTACAT

BWT	context
$	GATTAGAT
$	GATTAGATA
A	T$AGATAC
A	T$GATAC
A	T$GATTAC
A	T$GATTAG
A	TA$GATTAG
A	TACAT$AG
A	TACAT$G
T	TACAT$GA
T	TAGAT$GA
T	TAGATA$GA
A	TTACAT$G
A	TTAGAT$G
A	TTAGATA$G

T A G A T

Toehold Lemma

G A T T A C A T $ A G A T A C A T $ G A T A C A T $ G A T T A G A T $ G A T T A G A T A $

BWT	context
T	$AGATACA
T	$GATACA
T	$GATTACA
T	$GATTAGA
A	$GATTAGAT
T	A$GATTAGA
T	ACAT$AGA
T	ACAT$GA
T	ACAT$GAT
T	AGAT$GAT
T	AGATA$GAT
$	AGATACAT
C	AT$AGATA
C	AT$GATA
C	AT$GATTA

BWT	context
G	AT$GATTA
G	ATA$GATTA
G	ATACAT$A
G	ATACAT$
G	ATTACAT$
G	ATTAGAT$
G	ATTAGATA$
A	CAT$AGAT
A	CAT$GAT
A	CAT$GATT
A	GAT$GATT
A	GATA$GATT
A	GATACAT$
$	GATACAT
$	GATTACAT

BWT	context
$	GATTAGAT
$	GATTAGATA
A	T$AGATAC
A	T$GATAC
A	T$GATTAC
A	T$GATTAG
A	TA$GATTAG
A	TACAT$AG
A	TACAT$G
T	TACAT$GA
T	TAGAT$GA
T	TAGATA$GA
A	TTACAT$G
A	TTAGAT$G
A	TTAGATA$G

T A G A T

Toehold Lemma

GATTACAT$ AGATAGATA$ GATACAT$ GATTAGAT$ GATTAGAT$

BWT	context
T	$AGATACA
T	$GATACA
T	$GATTACA
T	$GATTAGA
A	$GATTAGAT
T	A$GATTAGA
T	ACAT$AGA
T	ACAT$GA
T	ACAT$GAT
T	AGAT$GAT
T	AGATA$GAT
$	AGATACAT
C	AT$AGATA
C	AT$GATA
C	AT$GATTA

BWT	context
G	AT$GATTA
G	ATA$GATTA
G	ATACAT$A
G	ATACAT$
G	ATTACAT$
G	ATTAGAT$
G	ATTAGATA$
A	CAT$AGAT
A	CAT$GAT
A	CAT$GATT
A	GAT$GATT
A	GATA$GATT
A	GATACAT$
$	GATACAT
$	GATTACAT

BWT	context
$	GATTAGAT
$	GATTAGATA
A	T$AGATAC
A	T$GATAC
A	T$GATTAC
A	T$GATTAG
A	TA$GATTAG
A	TACAT$AG
A	TACAT$G
T	TACAT$GA
T	TAGAT$GA
T	TAGATA$GA
A	TTACAT$G
A	TTAGAT$G
A	TTAGATA$G

T A G A T

Toehold Lemma

GATTACAT$ AGATACAT$ GATACAT$ GATTAGAT$ GATTAGAT$

BWT	context
T	$AGATACA
T	$GATACA
T	$GATTACA
T	$GATTAGA
A	$GATTAGAT
T	A$GATTAGA
T	ACAT$AGA
T	ACAT$GA
T	ACAT$GAT
T	AGAT$GAT
T	AGATA$GAT
$	AGATACAT
C	AT$AGATA
C	AT$GATA
C	AT$GATTA

BWT	context
G	AT$GATTA
G	ATA$GATTA
G	ATACAT$A
G	ATACAT$
G	ATTACAT$
G	ATTAGAT$
G	ATTAGATA$
A	CAT$AGAT
A	CAT$GAT
A	CAT$GATT
A	GAT$GATT
A	GATA$GATT
A	GATACAT$
$	GATACAT
$	GATTACAT

BWT	context
$	GATTAGAT
$	GATTAGATA
A	T$AGATAC
A	T$GATAC
A	T$GATTAC
A	T$GATTAG
A	TA$GATTAG
A	TACAT$AG
A	TACAT$G
T	TACAT$GA
T	TAGAT$GA
T	TAGATA$GA
A	TTACAT$G
A	TTAGAT$G
A	TTAGATA$G

T A G A T

Toehold Lemma

GATTACAT$ AGATACAT$ GATACAT$ GATTAGAT$ GATTAGATA$

BWT	context
T	$AGATACA
T	$GATACA
T	$GATTACA
T	$GATTAGA
A	$GATTAGAT
T	A$GATTAGA
T	ACAT$AGA
T	ACAT$GA
T	ACAT$GAT
T	AGAT$GAT
T	AGATA$GAT
$	AGATACAT
C	AT$AGATA
C	AT$GATA
C	AT$GATTA

BWT	context
G	AT$GATTA
G	ATA$GATTA
G	ATACAT$A
G	ATACAT$
G	ATTACAT$
G	ATTAGAT$
G	ATTAGATA$
A	CAT$AGAT
A	CAT$GAT
A	CAT$GATT
A	GAT$GATT
A	GATA$GATT
A	GATACAT$
$	GATACAT
$	GATTACAT

BWT	context
$	GATTAGAT
$	GATTAGATA
A	T$AGATAC
A	T$GATAC
A	T$GATTAC
A	T$GATTAG
A	TA$GATTAG
A	TACAT$AG
A	TACAT$G
T	TACAT$GA
T	TAGAT$GA
T	TAGATA$GA
A	TTACAT$G
A	TTAGAT$G
A	TTAGATA$G

T A G A T

Toehold Lemma

GATTACAT$ AGATACAT$ GATACAT$ GATTACAT$ GATTAGAT$

BWT	context
T	$AGATACA
T	$GATACA
T	$GATTACA
T	$GATTAGA
A	$GATTAGAT
T	A$GATTAGA
T	ACAT$AGA
T	ACAT$GA
T	ACAT$GAT
T	AGAT$GAT
T	AGATA$GAT
$	AGATACAT
C	AT$AGATA
C	AT$GATA
C	AT$GATTA

BWT	context
G	AT$GATTA
G	ATA$GATTA
G	ATACAT$A
G	ATACAT$
G	ATTACAT$
G	ATTAGAT$
G	ATTAGATA$
A	CAT$AGAT
A	CAT$GAT
A	CAT$GATT
A	GAT$GATT
A	GATA$GATT
A	GATACAT$
$	GATACAT
$	GATTACAT

BWT	context
$	GATTAGAT
$	GATTAGATA
A	T$AGATAC
A	T$GATAC
A	T$GATTAC
A	T$GATTAG
A	TA$GATTAG
A	TACAT$AG
A	TACAT$G
T	TACAT$GA
T	TAGAT$GA
T	TAGATA$GA
A	TTACAT$G
A	TTAGAT$G
A	TTAGATA$G

T A G A T

Toehold Lemma

GATTACAT$ AGATACAT$ GATACAT$ GATTAGAT$ GATTAGATA$

BWT	context
T	$AGATACA
T	$GATACA
T	$GATTACA
T	$GATTAGA
A	$GATTAGAT
T	A$GATTAGA
T	ACAT$AGA
T	ACAT$GA
T	ACAT$GAT
T	AGAT$GAT
T	AGATA$GAT
$	AGATACAT
C	AT$AGATA
C	AT$GATA
C	AT$GATTA

BWT	context
G	AT$GATTA
G	ATA$GATTA
G	ATACAT$A
G	ATACAT$
G	ATTACAT$
G	ATTAGAT$
G	ATTAGATA$
A	CAT$AGAT
A	CAT$GAT
A	CAT$GATT
A	GAT$GATT
A	GATA$GATT
A	GATACAT$
$	GATACAT
$	GATTACAT

BWT	context
$	GATTAGAT
$	GATTAGATA
A	T$AGATAC
A	T$GATAC
A	T$GATTAC
A	T$GATTAG
A	TA$GATTAG
A	TACAT$AG
A	TACAT$G
T	TACAT$GA
T	TAGAT$GA
T	TAGATA$GA
A	TTACAT$G
A	TTAGAT$G
A	TTAGATA$G

T A G A T

Toehold Lemma

GATTACAT$ AGATACAT$ GATACAT$ GATTAGAT$ GATTAGATA$

BWT	context
T	$AGATACA
T	$GATACA
T	$GATTACA
T	$GATTAGA
A	$GATTAGAT
T	A$GATTAGA
T	ACAT$AGA
T	ACAT$GA
T	ACAT$GAT
T	AGAT$GAT
T	AGATA$GAT
$	AGATACAT
C	AT$AGATA
C	AT$GATA
C	AT$GATTA

BWT	context
G	AT$GATTA
G	ATA$GATTA
G	ATACAT$A
G	ATACAT$
G	ATTACAT$
G	ATTAGAT$
G	ATTAGATA$
A	CAT$AGAT
A	CAT$GAT
A	CAT$GATT
A	GAT$GATT
A	GATA$GATT
A	GATACAT$
$	GATACAT
$	GATTACAT

BWT	context
$	GATTAGAT
$	GATTAGATA
A	T$AGATAC
A	T$GATAC
A	T$GATTAC
A	T$GATTAG
A	TA$GATTAG
A	TACAT$AG
A	TACAT$G
T	TACAT$GA
T	TAGAT$GA
T	TAGATA$GA
A	TTACAT$G
A	TTAGAT$G
A	TTAGATA$G

T A G A T

	fast counting	fast locating	fast MEM-finding	scalable
FM-index (*JACM*, 2005)	yes	yes	yes	no
RLCSA (*JCB*, 2010)	yes	no	no	yes
Toehold Lemma (*Algorithmica*, 2018)	yes	1	no	yes
r-index (*JACM*, 2020)	yes	yes	no	yes
MONI (*JCB*, 2022)	no	yes	yes	yes

r-index

Theorem

We can store a text $T[1..n]$ over $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}, \$\}$ in $O(r)$ space such that when given a pattern $P[1..m]$, we can

1. *count P's occurrences in T in $O(m)$ time*
2. *list those occurrences' positions in T in constant time per occurrence.*

...but still no MEMs!

	fast counting	fast locating	fast MEM-finding	scalable
FM-index (*JACM*, 2005)	yes	yes	yes	no
RLCSA (*JCB*, 2010)	yes	no	no	yes
Toehold Lemma (*Algorithmica*, 2018)	yes	1	no	yes
r-index (*JACM*, 2020)	yes	yes	no	yes
MONI (*JCB*, 2022)	no	yes	yes	yes

MONI

GATTACAT$ AGATACAT$ GATACAT$ GATTAGAT$ GATTAGAT$

BWT	context
T	$AGATACA
T	$GATACA
T	$GATTACA
T	$GATTAGA
A	$GATTAGAT
T	A$GATTAGA
T	ACAT$AGA
T	ACAT$GA
T	ACAT$GAT
T	AGAT$GAT
T	AGATA$GAT
$	AGATACAT
C	AT$AGATA
C	AT$GATA
C	AT$GATTA

BWT	context
G	AT$GATTA
G	ATA$GATTA
G	ATACAT$A
G	ATACAT$
G	ATTACAT$
G	ATTAGAT$
G	ATTAGATA$
A	CAT$AGAT
A	CAT$GAT
A	CAT$GATT
A	GAT$GATT
A	GATA$GATT
A	GATACAT$
$	GATACAT
$	GATTACAT

BWT	context
$	GATTAGAT
$	GATTAGATA
A	T$AGATAC
A	T$GATAC
A	T$GATTAC
A	T$GATTAG
A	TA$GATTAG
A	TACAT$AG
A	TACAT$G
T	TACAT$GA
T	TAGAT$GA
T	TAGATA$GA
A	TTACAT$G
A	TTAGAT$G
A	TTAGATA$G

MONI

GATTACAT$ AGATACAT$ GATACAT$ GATTAGAT$ GATTAGAT$

BWT	context
T	$AGATACA
T	$GATACA
T	$GATTACA
T	$GATTAGA
A	$GATTAGAT
T	A$GATTAGA
T	ACAT$AGA
T	ACAT$GA
T	ACAT$GAT
T	AGAT$GAT
T	AGATA$GAT
$	AGATACAT
C	AT$AGATA
C	AT$GATA
C	AT$GATTA

BWT	context
G	AT$GATTA
G	ATA$GATTA
G	ATACAT$A
G	ATACAT$
G	ATTACAT$
G	ATTAGAT$
G	ATTAGATA$
A	CAT$AGAT
A	CAT$GAT
A	CAT$GATT
A	GAT$GATT
A	GATA$GATT
A	GATACAT$
$	GATACAT
$	GATTACAT

BWT	context
$	GATTAGAT
$	GATTAGATA
A	T$AGATAC
A	T$GATAC
A	T$GATTAC
A	T$GATTAG
A	TA$GATTAG
A	TACAT$AG
A	TACAT$G
T	TACAT$GA
T	TAGAT$GA
T	TAGATA$GA
A	TTACAT$G
A	TTAGAT$G
A	TTAGATA$G

G A T A G A T

MONI

G A T T A C A T $ A G A T A C A T $ G A T A C A T $ G A T T A G A T $ G A T T A G A T A $

BWT	context
T	$AGATACA
T	$GATACA
T	$GATTACA
T	$GATTAGA
A	$GATTAGAT
T	A$GATTAGA
T	ACAT$AGA
T	ACAT$GA
T	ACAT$GAT
T	AGAT$GAT
T	AGATA$GAT
$	AGATACAT
C	AT$AGATA
C	AT$GATA
C	AT$GATTA

BWT	context
G	AT$GATTA
G	ATA$GATTA
G	ATACAT$A
G	ATACAT$
G	ATTACAT$
G	ATTAGAT$
G	ATTAGATA$
A	CAT$AGAT
A	CAT$GAT
A	CAT$GATT
A	GAT$GATT
A	GATA$GATT
A	GATACAT$
$	GATACAT
$	GATTACAT

BWT	context
$	GATTAGAT
$	GATTAGATA
A	T$AGATAC
A	T$GATAC
A	T$GATTAC
A	T$GATTAG
A	TA$GATTAG
A	TACAT$AG
A	TACAT$G
T	TACAT$GA
T	TAGAT$GA
T	TAGATA$GA
A	TTACAT$G
A	TTAGAT$G
A	TTAGATA$G

G A T A G A T

5 4 3 2 1

MONI

GATTACAT$ AGATACAT$ GATACAT$ GATTAGAT$ GATTAGAT$

BWT	context
T	$AGATACA
T	$GATACA
T	$GATTACA
T	$GATTAGA
A	$GATTAGAT
T	A$GATTAGA
T	ACAT$AGA
T	ACAT$GA
T	ACAT$GAT
T	AGAT$GAT
T	AGATA$GAT
$	AGATACAT
C	AT$AGATA
C	AT$GATA
C	AT$GATTA

BWT	context
G	AT$GATTA
G	ATA$GATTA
G	ATACAT$A
G	ATACAT$
G	ATTACAT$
G	ATTAGAT$
G	ATTAGATA$
A	CAT$AGAT
A	CAT$GAT
A	CAT$GATT
A	GAT$GATT
A	GATA$GATT
A	GATACAT$
$	GATACAT
$	GATTACAT

BWT	context
$	GATTAGAT
$	GATTAGATA
A	T$AGATAC
A	T$GATAC
A	T$GATTAC
A	T$GATTAG
A	TA$GATTAG
A	TACAT$AG
A	TACAT$G
T	TACAT$GA
T	TAGAT$GA
T	TAGATA$GA
A	TTACAT$G
A	TTAGAT$G
A	TTAGATA$G

G	A	T	A	G	A	T
	3	5	4	3	2	1

MONI

GATTACAT$ AGATACAT$ GATACAT$ GATTAGAT$ GATTAGAT$

BWT	context
T	$AGATACA
T	$GATACA
T	$GATTACA
T	$GATTAGA
A	$GATTAGAT
T	A$GATTAGA
T	ACAT$AGA
T	ACAT$GA
T	ACAT$GAT
T	AGAT$GAT
T	AGATA$GAT
$	AGATACAT
C	AT$AGATA
C	AT$GATA
C	AT$GATTA

BWT	context
G	AT$GATTA
G	ATA$GATTA
G	ATACAT$A
G	ATACAT$
G	ATTACAT$
G	ATTAGAT$
G	ATTAGATA$
A	CAT$AGAT
A	CAT$GAT
A	CAT$GATT
A	GAT$GATT
A	GATA$GATT
A	GATACAT$
$	GATACAT
$	GATTACAT

BWT	context
$	GATTAGAT
$	GATTAGATA
A	T$AGATAC
A	T$GATAC
A	T$GATTAC
A	T$GATTAG
A	TA$GATTAG
A	TACAT$AG
A	TACAT$G
T	TACAT$GA
T	TAGAT$GA
T	TAGATA$GA
A	TTACAT$G
A	TTAGAT$G
A	TTAGATA$G

G	A	T	A	G	A	T
4	3	5	4	3	2	1

MONI

GATTACAT$ AGATACAT$ GATACAT$ GATTAGAT$ GATTAGAT$

BWT	context
T	$AGATACA
T	$GATACA
T	$GATTACA
T	$GATTAGA
A	$GATTAGAT
T	A$GATTAGA
T	ACAT$AGA
T	ACAT$GA
T	ACAT$GAT
T	AGAT$GAT
T	AGATA$GAT
$	AGATACAT
C	AT$AGATA
C	AT$GATA
C	AT$GATTA

BWT	context
G	AT$GATTA
G	ATA$GATTA
G	ATACAT$A
G	ATACAT$
G	ATTACAT$
G	ATTAGAT$
G	ATTAGATA$
A	CAT$AGAT
A	CAT$GAT
A	CAT$GATT
A	GAT$GATT
A	GATA$GATT
A	GATACAT$
$	GATACAT
$	GATTACAT

BWT	context
$	GATTAGAT
$	GATTAGATA
A	T$AGATAC
A	T$GATAC
A	T$GATTAC
A	T$GATTAG
A	TA$GATTAG
A	TACAT$AG
A	TACAT$G
T	TACAT$GA
T	TAGAT$GA
T	TAGATA$GA
A	TTACAT$G
A	TTAGAT$G
A	TTAGATA$G

G	A	T	A	G	A	T
4	3	5	4	3	2	1

MONI

Theorem

We can store a text $T[1..n]$ over $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}, \$\}$ in $O(r + |\mathsf{LCE}|)$ space, where $|\mathsf{LCE}|$ is the space for an LCE data structure, such that when given a pattern $P[1..m]$ we can

1. *find the MEMs of P with respect to T in $O(m \log n)$ time*
2. *list their occurrences' positions in T in constant time per occurrence.*

Corollary

We can store T in $O(r + t + |\mathsf{LCE}|)$ space such that when given P, we can list all the distinct **vertices** *where we start processing occurrences of P in T, in $O(m \log n)$ time plus constant time per* **occurrence**.

Can we find the interval directly? DO WE NEED IT?

MARIA

(unpublished joint work with Andrej Baláž, Adrián Goga and Alessia Petescia)

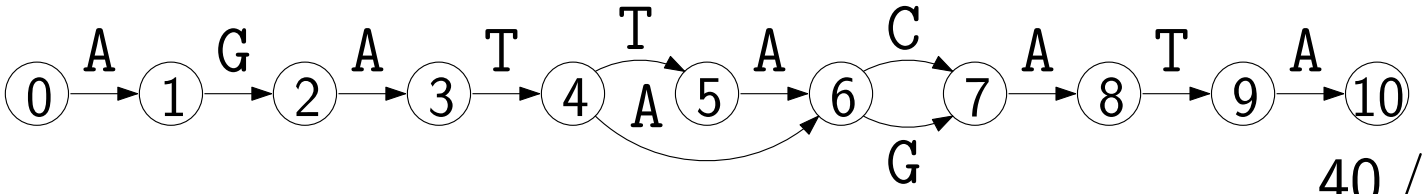
data structure

GATTACAT$ AGATACAT$ GATACAT$ GATTAGAT$ GATTAGATA$

tag	context
9	$AGATACA
9	$GATACA
9	$GATTACA
9	$GATTAGA
10	$GATTAGAT
9	A$GATTAGA
4	ACAT$AGA
4	ACAT$GA
5	ACAT$GAT
5	AGAT$GAT
5	AGATA$GAT
0	AGATACAT
7	AT$AGATA
7	AT$GATA
7	AT$GATTA

tag	context
7	AT$GATTA
7	ATA$GATTA
2	ATACAT$A
2	ATACAT$
2	ATTACAT$
2	ATTAGAT$
2	ATTAGATA$
6	CAT$AGAT
6	CAT$GAT
6	CAT$GATT
6	GAT$GATT
6	GATA$GATT
1	GATACAT$
1	GATACAT
1	GATTACAT

tag	context
1	GATTAGAT
1	GATTAGATA
8	T$AGATAC
8	T$GATAC
8	T$GATTAC
8	T$GATTAG
8	TA$GATTAG
3	TACAT$AG
3	TACAT$G
4	TACAT$GA
4	TAGAT$GA
4	TAGATA$GA
3	TTACAT$G
3	TTAGAT$G
3	TTAGATA$G



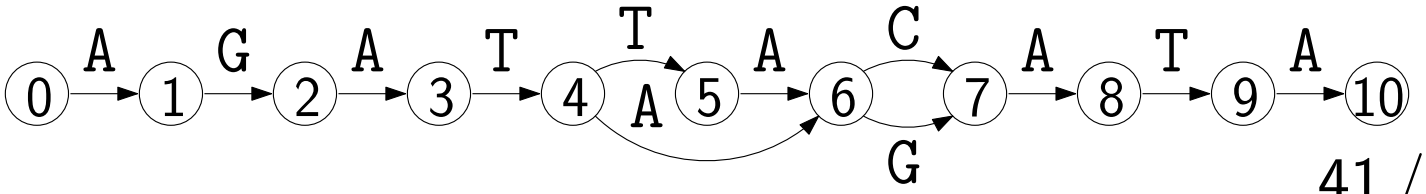
examples

GATTACAT$ AGATACAT$ GATACAT$ GATTAGAT$ GATTAGAT$

tag	context
9	$AGATACA
9	$GATACA
9	$GATTACA
9	$GATTAGA
10	$GATTAGAT
9	A$GATTAGA
4	ACAT$AGA
4	ACAT$GA
5	ACAT$GAT
5	AGAT$GAT
5	AGATA$GAT
0	AGATACAT
7	AT$AGATA
7	AT$GATA
7	AT$GATTA

tag	context
7	AT$GATTA
7	ATA$GATTA
2	ATACAT$A
2	ATACAT$
2	ATTACAT$
2	ATTAGAT$
2	ATTAGATA$
6	CAT$AGAT
6	CAT$GAT
6	CAT$GATT
6	GAT$GATT
6	GATA$GATT
1	GATACAT$
1	GATACAT
1	GATTACAT

tag	context
1	GATTAGAT
1	GATTAGATA
8	T$AGATAC
8	T$GATAC
8	T$GATTAC
8	T$GATTAG
8	TA$GATTAG
3	TACAT$AG
3	TACAT$G
4	TACAT$GA
4	TAGAT$GA
4	TAGATA$GA
3	TTACAT$G
3	TTAGAT$G
3	TTAGATA$G



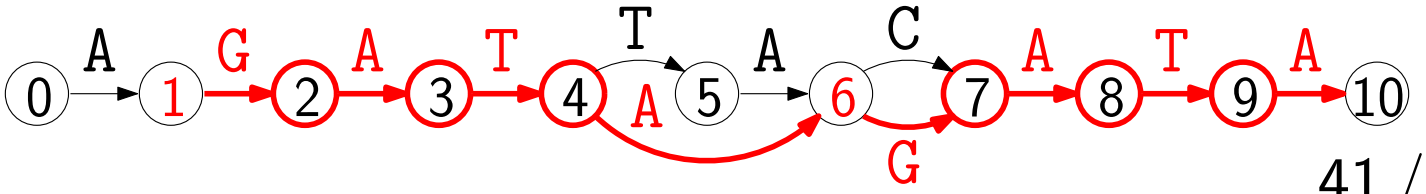
examples

GATTACAT$ AGATACAT$ GATACAT$ GATTAGAT$ GATTAGATA$

tag	context
9	$AGATACA
9	$GATACA
9	$GATTACA
9	$GATTAGA
10	$GATTAGAT
9	A$GATTAGA
4	ACAT$AGA
4	ACAT$GA
5	ACAT$GAT
5	AGAT$GAT
5	AGATA$GAT
0	AGATACAT
7	AT$AGATA
7	AT$GATA
7	AT$GATTA

tag	context
7	AT$GATTA
7	ATA$GATTA
2	ATACAT$A
2	ATACAT$
2	ATTACAT$
2	ATTAGAT$
2	ATTAGATA$
6	CAT$AGAT
6	CAT$GAT
6	CAT$GATT
6	GAT$GATT
6	GATA$GATT
1	GATACAT$
1	GATACAT
1	GATTACAT

tag	context
1	GATTAGAT
1	GATTAGATA
8	T$AGATAC
8	T$GATAC
8	T$GATTAC
8	T$GATTAG
8	TA$GATTAG
3	TACAT$AG
3	TACAT$G
4	TACAT$GA
4	TAGAT$GA
4	TAGATA$GA
3	TTACAT$G
3	TTAGAT$G
3	TTAGATA$G



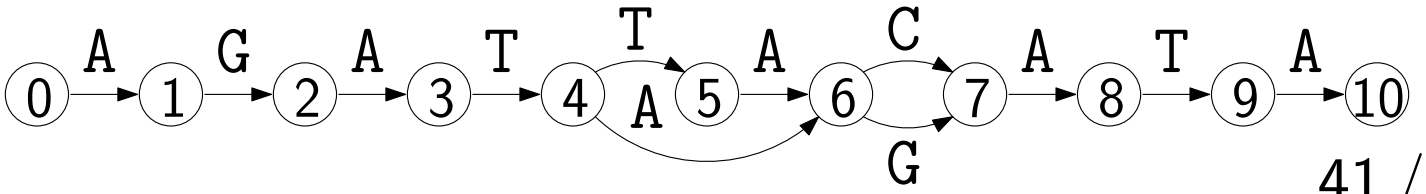
examples

GATTACAT$ AGATACAT$ GATACAT$ GATTAGAT$ GATTAGATA$

tag	context
9	$AGATACA
9	$GATACA
9	$GATTACA
9	$GATTAGA
10	$GATTAGAT
9	A$GATTAGA
4	ACAT$AGA
4	ACAT$GA
5	ACAT$GAT
5	AGAT$GAT
5	AGATA$GAT
0	AGATACAT
7	AT$AGATA
7	AT$GATA
7	AT$GATTA

tag	context
7	AT$GATTA
7	ATA$GATTA
2	ATACAT$A
2	ATACAT$
2	ATTACAT$
2	ATTAGAT$
2	ATTAGATA$
6	CAT$AGAT
6	CAT$GAT
6	CAT$GATT
6	GAT$GATT
6	GATA$GATT
1	GATACAT$
1	GATACAT
1	GATTACAT

tag	context
1	GATTAGAT
1	GATTAGATA
8	T$AGATAC
8	T$GATAC
8	T$GATTAC
8	T$GATTAG
8	TA$GATTAG
3	TACAT$AG
3	TACAT$G
4	TACAT$GA
4	TAGAT$GA
4	TAGATA$GA
3	TTACAT$G
3	TTAGAT$G
3	TTAGATA$G



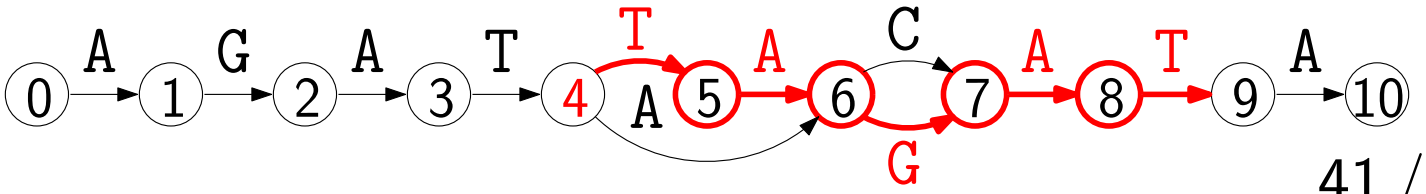
examples

GATTACAT$ AGATACAT$ GATACAT$ GATTAGAT$ GATTAGATA$

tag	context
9	$AGATACA
9	$GATACA
9	$GATTACA
9	$GATTAGA
10	$GATTAGAT
9	A$GATTAGA
4	ACAT$AGA
4	ACAT$GA
5	ACAT$GAT
5	AGAT$GAT
5	AGATA$GAT
0	AGATACAT
7	AT$AGATA
7	AT$GATA
7	AT$GATTA

tag	context
7	AT$GATTA
7	ATA$GATTA
2	ATACAT$A
2	ATACAT$
2	ATTACAT$
2	ATTAGAT$
2	ATTAGATA$
6	CAT$AGAT
6	CAT$GAT
6	CAT$GATT
6	GAT$GATT
6	GATA$GATT
1	GATACAT$
1	GATACAT
1	GATTACAT

tag	context
1	GATTAGAT
1	GATTAGATA
8	T$AGATAC
8	T$GATAC
8	T$GATTAC
8	T$GATTAG
8	TA$GATTAG
3	TACAT$AG
3	TACAT$G
4	TACAT$GA
4	TAGAT$GA
4	TAGATA$GA
3	TTACAT$G
3	TTAGAT$G
3	TTAGATA$G



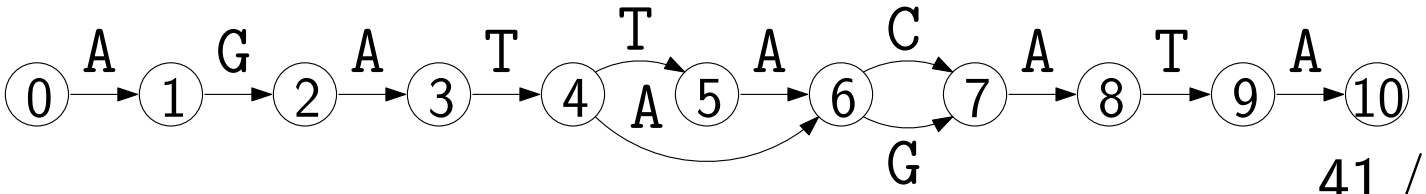
examples

GATTACAT$ AGATACAT$ GATACAT$ GATTAGAT$ GATTAGATA$

tag	context
9	$AGATACA
9	$GATACA
9	$GATTACA
9	$GATTAGA
10	$GATTAGAT
9	A$GATTAGA
4	ACAT$AGA
4	ACAT$GA
5	ACAT$GAT
5	AGAT$GAT
5	AGATA$GAT
0	AGATACAT
7	AT$AGATA
7	AT$GATA
7	AT$GATTA

tag	context
7	AT$GATTA
7	ATA$GATTA
2	ATACAT$A
2	ATACAT$
2	ATTACAT$
2	ATTAGAT$
2	ATTAGATA$
6	CAT$AGAT
6	CAT$GAT
6	CAT$GATT
6	GAT$GATT
6	GATA$GATT
1	GATACAT$
1	GATACAT
1	GATTACAT

tag	context
1	GATTAGAT
1	GATTAGATA
8	T$AGATAC
8	T$GATAC
8	T$GATTAC
8	T$GATTAG
8	TA$GATTAG
3	TACAT$AG
3	TACAT$G
4	TACAT$GA
4	TAGAT$GA
4	TAGATA$GA
3	TTACAT$G
3	TTAGAT$G
3	TTAGATA$G



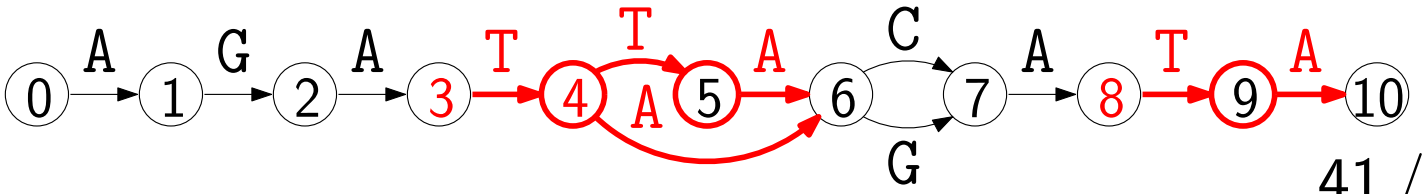
examples

GATTACAT$ AGATACAT$ GATACAT$ GATTAGAT$ GATTAGATA$

tag	context
9	$AGATACA
9	$GATACA
9	$GATTACA
9	$GATTAGA
10	$GATTAGAT
9	A$GATTAGA
4	ACAT$AGA
4	ACAT$GA
5	ACAT$GAT
5	AGAT$GAT
5	AGATA$GAT
0	AGATACAT
7	AT$AGATA
7	AT$GATA
7	AT$GATTA

tag	context
7	AT$GATTA
7	ATA$GATTA
2	ATACAT$A
2	ATACAT$
2	ATTACAT$
2	ATTAGAT$
2	ATTAGATA$
6	CAT$AGAT
6	CAT$GAT
6	CAT$GATT
6	GAT$GATT
6	GATA$GATT
1	GATACAT$
1	GATACAT
1	GATTACAT

tag	context
1	GATTAGAT
1	GATTAGATA
8	T$AGATAC
8	T$GATAC
8	T$GATTAC
8	T$GATTAG
8	TA$GATTAG
3	TACAT$AG
3	TACAT$G
4	TACAT$GA
4	TAGAT$GA
4	TAGATA$GA
3	TTACAT$G
3	TTAGAT$G
3	TTAGATA$G



theorem

Theorem

We can store a text $T[1..n]$ in $O(r + t + |\mathsf{LCE}|)$ space such that when given a pattern $P[1..m]$, we can

1. *find the MEMs of P with respect to T in $O(m \log n)$ time*
2. *list all the distinct vertices where we start processing occurrences in T of each MEM, in constant time per vertex listed.*

conclusion

future work

With MONI and MARIA, we have a pipeline for our goal:

1. index the dataset as a set of strings *losslessly* — **without adding or excluding any variations**
2. find good seeds in the reads exactly matching **substrings of the dataset**
3. map those seeds **directly** onto walks in the graph* — **in time independent of the number of matches in the dataset (which we don't even compute)**.

When MARIA is finished, someone (else!) should still extend the MEMs to approximate matches for the reads, build a consensus sequence, etc.

*The seeds needn't match the walks exactly! We can simplify the graph — as a *coordinate system* — without weakening the index.

thanks to...

Omar Ahmed
Jarno Alanko
Andrej Baláž
Hideo Bannai
Paola Bonizzoni
Christina Boucher
Nate Brown
Nicola Cotumaccio
Adrián Goga
Simon Heumos
Tomohiro I
Dominik Köppl
Alan Kuhnle
Ben Langmead
Veli Mäkinen
Giovanni Manzini
Taher Mun
Gonzalo Navarro
Alessia Petescia
Nicola Prezza
Massimiliano Rossi
Jouni Sirén

DALHOUSIE UNIVERSITY
FACULTY OF COMPUTER SCIENCE

NSERC CRSNG

ICG INSTITUTE FOR COMPARATIVE GENOMICS

NIH

CeBiB Centre for Biotechnology and Bioengineering

NSF

questions?