Building Large Updatable Colored de Bruijn Graphs via Merging

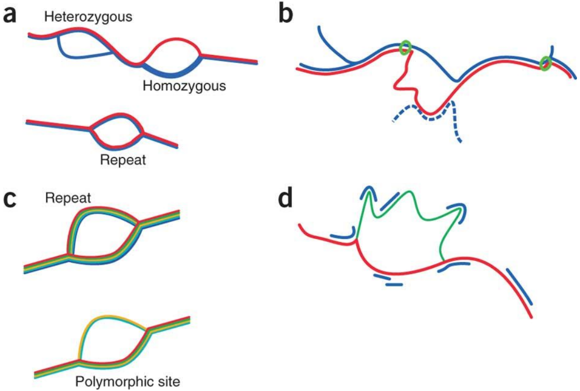
Martin D. Muggli[1] Bahar Alipanahi[2] Christina Boucher[2]

[1]Colorado State University

[2]University of Florida

Colored de Bruijn Graphs

In 2012, Iqbal *et al.* introduced the colored de Bruijn graph with CORTEX. It can detect complex variants within a population without a reference.



Related Work

Efficient de Bruijn graphs

- ABySS
- Conway and Bromage
- Okanohara and Sadakane
- Minia
- BOSS
- MEGAHIT
- Chikhi *et al.*

Efficient colored de Bruijn graphs

- VARI
- Rainbowfish
- Bloom filter trie
- Mantis
- Almodaresi *et al.*

Efficient color representation

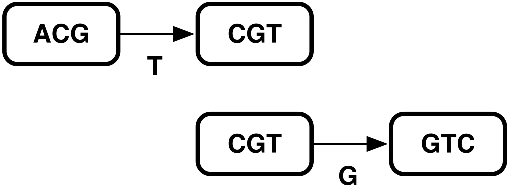
- Mustafa *et al.*
- Mutli-BRWT

Our contribution

- We developed VARIMERGE
  - Construct succinct colored de Bruijn for sub-populations using VARI
  - New algorithm to merge succinct colored de Bruijn graphs
- Advantages
  - Compress data early => Use less and faster memory
  - Reuse previous work => Incremental update
- First to demonstrate incremental update at this scale

Background: de Bruijn Graphs

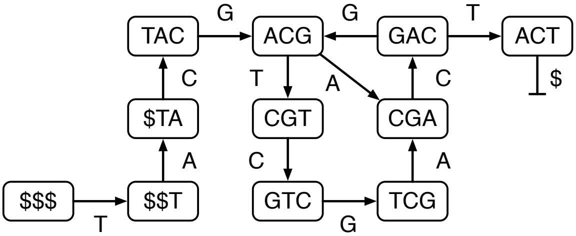
T = TACG**ACGT**CGACT



ACG
CGT

Vertex labels are redundant

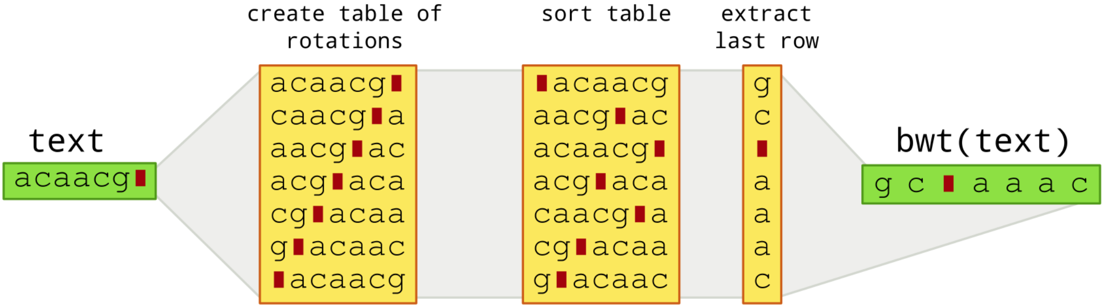
T = TACGACGTCGACT



Burrows-Wheeler Transform (BWT)

Advantages:

- Compresses repetitive strings well.
- *Self index*: Encodes original string and can provide an index of the implicit *suffix array*.
- $\mathrm{BWT}[i] = \mathrm{X}[\mathrm{SA}[i] - 1]$ if $\mathrm{SA}[i] > 1$ and $\$$ otherwise.



text

acaacg▮

create table of rotations

acaacg▮
caacg▮a
aacg▮ac
acg▮aca
cg▮acaa
g▮acaac
▮acaacg

sort table

▮acaacg
aacg▮ac
acaacg▮
acg▮aca
caacg▮a
cg▮acaa
g▮acaac

extract last row

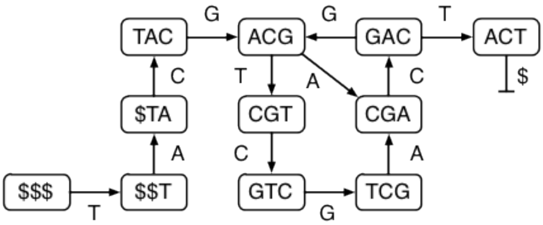
g
c
▮
a
a
a
c

bwt(text)

g c ▮ a a a c

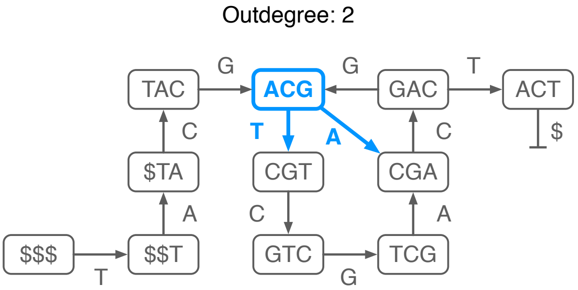
Succinct de Bruijn graphs sort origin labels colex.

Succinct de Bruijn Graphs represent edges as last-to-first mappings in the Burrows-Wheeler transform.



Node			*W*
$	$	$	T
C	G	A	C
$	T	A	C
G	A	C	G
G	A	C	T
T	A	C	G-
G	T	C	G
A	C	G	A
A	C	G	T
T	C	G	A-
$	$	T	A
A	C	T	$
C	G	T	C

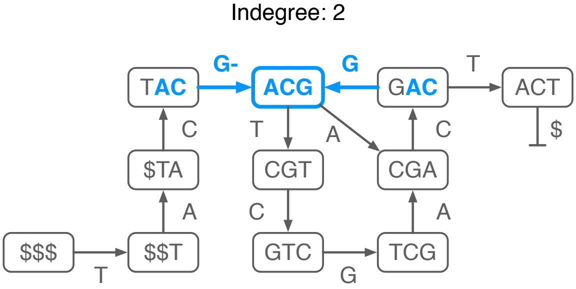
Encoding origins = $(k - 1)$ common vertex suffixes



F	
$	0
A	1
C	3
G	7
T	10

v	*i*	*L*	*Node*			*W*
0	0	1	$	$	$	T
1	1	1	C	G	A	C
2	2	1	$	T	A	C
3	3	0	G	A	C	G
	4	1	G	A	C	T
4	5	1	T	A	C	G-
5	6	1	G	T	C	G
6	**7**	0	A	C	G	A
	8	1	A	C	G	T
7	9	1	T	C	G	A-
8	10	1	$	$	T	A
9	11	1	A	C	T	$
10	12	1	C	G	T	C

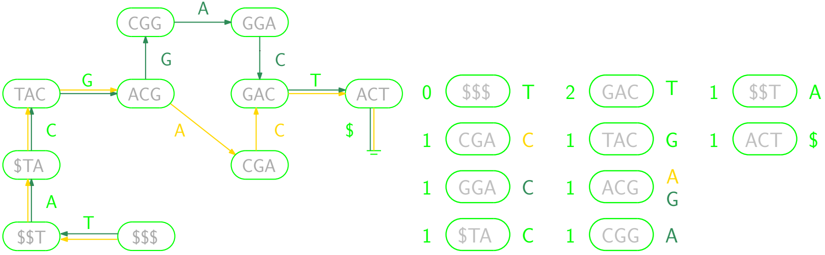
Encoding destinations = $(k-2)$ common vertex suffixes



F	
$	0
A	1
C	3
G	7
T	10

v	*i*	*L*	*Node*			*W*
0	0	1	$	$	$	T
1	1	1	C	G	A	C
2	2	1	$	T	A	C
3	3	0	G	A	C	G
	4	1	G	A	C	T
4	5	1	T	A	C	G-
5	6	1	G	T	C	G
6	**7**	0	A	C	G	A
	8	1	A	C	G	T
7	9	1	T	C	G	A-
8	10	1	$	$	T	A
9	11	1	A	C	T	$
10	12	1	C	G	T	C

VARI method: e.g. a two colored de Bruijn graph and its representation



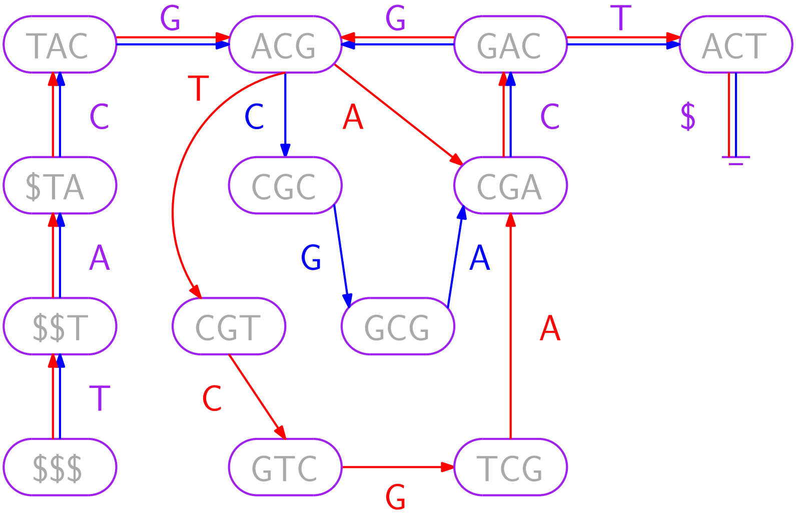
$$\mathrm{EBWT}(G) = \quad \texttt{TCCCTGAGAA\$}$$

$$C^{\mathrm{T}} = \quad \begin{array}{l} \texttt{11011110011} \\ \texttt{10111101111} \end{array}$$

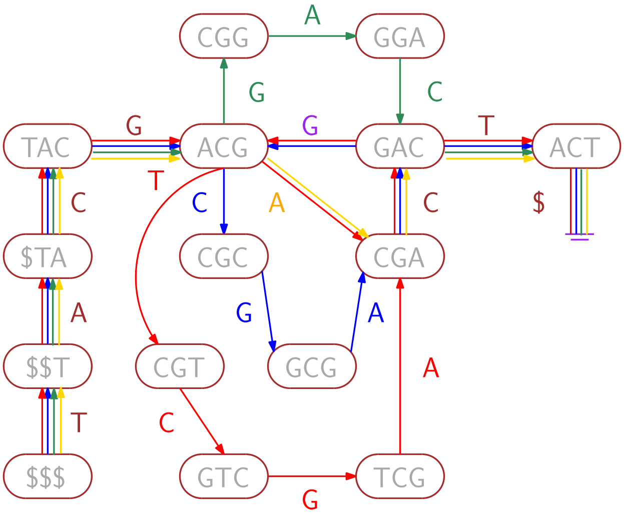
VARIMERGE: Main Algorithm

1. Consider the final population as a collection of sub-populations
2. Run VARI on each sub-population
3. Run our new algorithm, MERGE on the succinct de Bruijn graphs

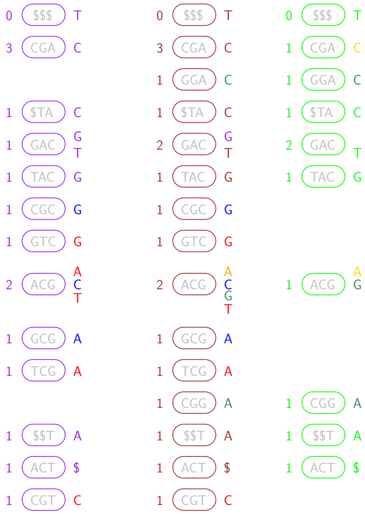
Another two colored de Bruijn graph



A four colored merged de Bruijn graph

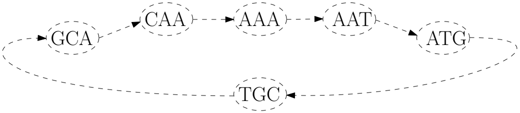


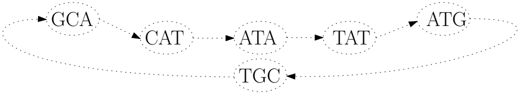
Merging edge labels requires origin vertex label



0	$$$	T	0	$$$	T	0	$$$	T
3	CGA	C	3	CGA	C	1	CGA	C
			1	GGA	C	1	GGA	C
1	$TA	C	1	$TA	C	1	$TA	C
1	GAC	G T	2	GAC	G T	2	GAC	T
1	TAC	G	1	TAC	G	1	TAC	G
1	CGC	G	1	CGC	G			
1	GTC	G	1	GTC	G			
2	ACG	A C T	2	ACG	A C G T	1	ACG	A G
1	GCG	A	1	GCG	A			
1	TCG	A	1	TCG	A			
			1	CGG	A	1	CGG	A
1	$$T	A	1	$$T	A	1	$$T	A
1	ACT	$	1	ACT	$	1	ACT	$
1	CGT	C	1	CGT	C			

Two succinct de Bruijn graphs, ignoring color



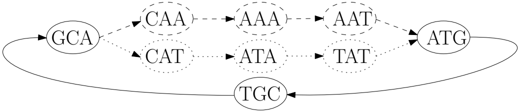


			MSD	LSD
Edge k-mer 1	A	A	A	T
	C	A	A	A
	G	C	A	A
	T	G	C	A
	A	T	G	C
Edge k-mer 6	A	A	T	G

		MSD	LSD
G	C	A	T
A	T	A	T
T	G	C	A
A	T	G	C
C	A	T	A
T	A	T	G

The merged graph

Can we generate the merged succinct graph without reconstructing full vertex *k*-mers?



Prior BWT merge methods

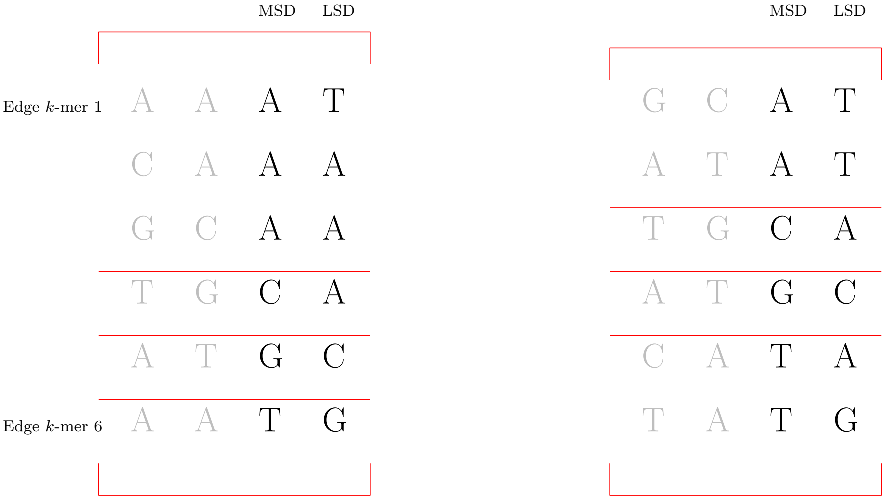
- Holt and McMillan
- Sirén

Generate Most Significant Digit (MSD) column

			MSD	LSD
Edge k-mer 1	A	A	A	T
	C	A	A	A
	G	C	A	A
	T	G	C	A
	A	T	G	C
Edge k-mer 6	A	A	T	G

		MSD	LSD
G	C	A	T
A	T	A	T
T	G	C	A
A	T	G	C
C	A	T	A
T	A	T	G

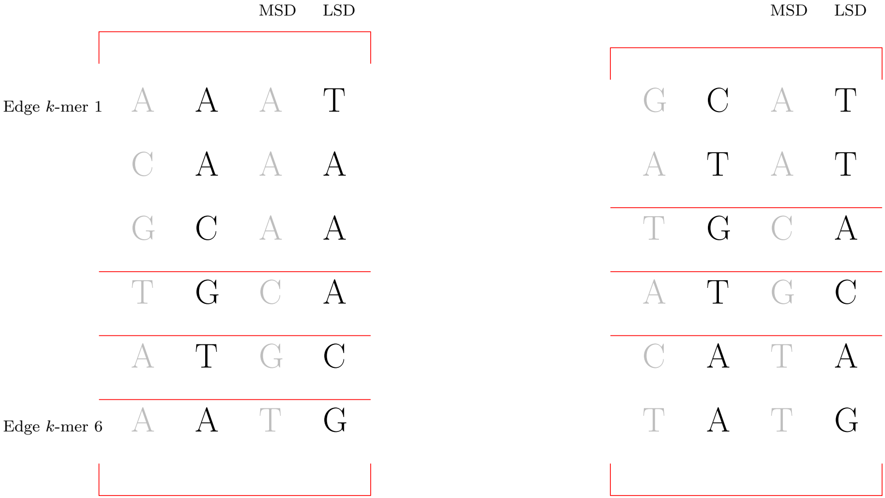
Generate common suffix groups for vertex labels



			MSD	LSD
Edge k-mer 1	A	A	A	T
	C	A	A	A
	G	C	A	A
	T	G	C	A
	A	T	G	C
Edge k-mer 6	A	A	T	G

		MSD	LSD
G	C	A	T
A	T	A	T
T	G	C	A
A	T	G	C
C	A	T	A
T	A	T	G

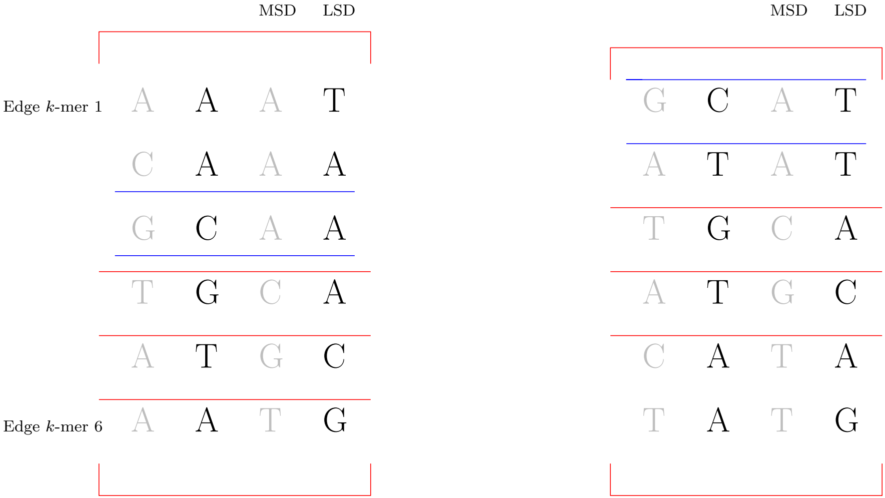
Generate next most significant digit column



			MSD	LSD
Edge k-mer 1	A	A	A	T
	C	A	A	A
	G	C	A	A
	T	G	C	A
	A	T	G	C
Edge k-mer 6	A	A	T	G

		MSD	LSD
G	C	A	T
A	T	A	T
T	G	C	A
A	T	G	C
C	A	T	A
T	A	T	G

Recursively subdivide existing groups using new column



			MSD	LSD
Edge k-mer 1	A	A	A	T
	C	A	A	A
	G	C	A	A
	T	G	C	A
	A	T	G	C
Edge k-mer 6	A	A	T	G

		MSD	LSD
G	C	A	T
A	T	A	T
T	G	C	A
A	T	G	C
C	A	T	A
T	A	T	G

Constructing 8,000 Salmonella sample graph via merging

Sub population size = 4,000 assemblies

VARIMERGE(8000) = MERGE(VARI(4000), VARI(4000))

Program	Time	External Memory	RAM
VARI(8000)	37 h 27 m	4.6 TB	271 GB
VARIMERGE(8000)	26 h 30 m	1.5 TB	137 GB

Incremental update performance

Program	Time	External Memory	RAM
VariMerge(16000)	69 h 8 m	2.34 TB	254 GB
Vari(1)	7 s	460 MB	2.3 GB
Merge(16000, 1)	7 h 9 m	0	254 GB

Comparison of space-efficient colored graph construction methods

Dataset	No. of k-mers	Program	Output Size	Time	RAM(RSS)
16,000	5.8 Billion	VARI / Rainbowfish	N/A	N/A	N/A
		Bloom Filter Trie	N/A	N/A	N/A
		Multi-BRWT	N/A	N/A	N/A
		Mantis / Method of Almodaresi et al.	256 GB	36 h 12 m	316 GB
		VARIMERGE	233 GB	69 h 8 m	254 GB

Conclusion

- Uncompressed work in small chunk reduces external memory
- Reusing previous computational work lets us build an updated version
- Radix based method satisfies metadata consistency and has no random access

Future work

- What is the optimal sub-population size for initial succinct colored de Bruijn graphs?
- VARIMERGE is radix based, would a trie based merge like bwt-merge be superior under some circumstances?

Acknowledgements

This work was supported by

- National Science Foundation (NSF) IIS Grant No. 1618814
- National Institute of Allergy and Infectious Diseases Institute of the National Institutes of Health (NIH) Grant No. 70R01AI141810-01

Questions

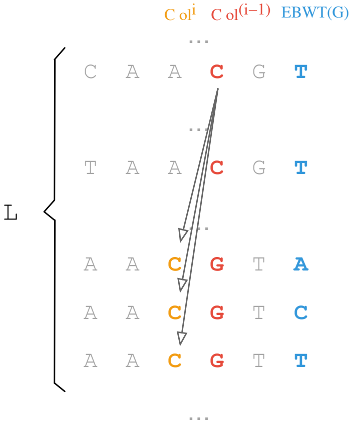
Questions?

Validation

1. Build succinct colored de Bruijn graphs for *E. coli* genome datasets A and B independently using VARI
2. Merge them using VARIMERGE
3. Build a succinct colored de Bruijn graph for all the data in A and B using VARI
4. Compare merged graph from step 2 with directly constructed graph from step 3

Result: bit-for-bit identical files on disk

Efficiently computing one column at a time



Col^i $Col^{(i-1)}$ EBWT(G)

...

C A A C G T

...

T A A C G T

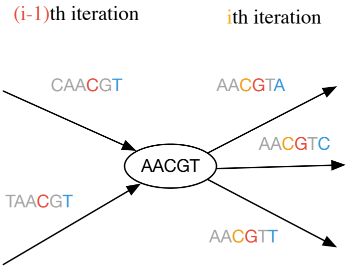
...

A A C G T A

A A C G T C

A A C G T T

...



(i-1)th iteration

ith iteration

Constructing 8,000 color graph via merging

	Input Stats		de Bruijn Graph			Color Matrix			Combined Requirements			
Program and Dataset	***k*-mers**	**Colors**	**RAM**	**Time**	**Size**	**RAM**	**Time**	**Size**	**RAM**	**Ext. Mem.**	**Time**	**Size**
VARI(4A)	1.1 B	4,000	136 GB	8 h 46 m	0.31 GB	52 GB	1 h 39 m	51.2 GB	136 GB	1 TB	10 h 25 m	51 GB
VARI(4B)	1.5 B	4,000	137 GB	10 h 40 m	0.52 GB	54 GB	2 h 22 m	52.5 GB	137 GB	1.5 TB	13 h 2 m	53 GB
MERGE(4A, 4B)	2.4 B	8,000	10 GB	2 h 1 m	0.63 GB	117 GB	1 h 2 m	106 GB	117 GB	0 TB	3 h 3 m	106 GB
VARIMERGE	2.4	8,000	137 GB	21 h 27 m	0.63 GB	117 GB	5 h 3 m	117 GB	137 GB	1.5 TB	26 h 30 m	106 GB

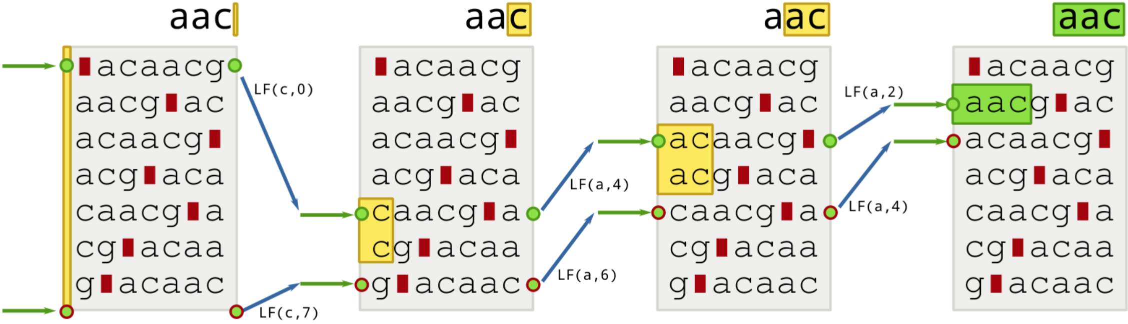
Constructing 16,000 color graph via merging

	Input Stats		de Bruijn Graph			Color Matrix			Combined Requirements			
Program and Dataset	***k*-mers**	**Colors**	**RAM**	**Time**	**Size**	**RAM**	**Time**	**Size**	**RAM**	**Ext. Mem.**	**Time**	**Size**
Vari(4C)	1.7 B	4,000	135 GB	10 h 53 m	0.46 GB	53 GB	2 h 34 m	51.8 GB	135 GB	1.6 TB	13 h 27 m	52 GB
Vari(4D)	2.4 B	4,000	137 GB	14 h 35 m	0.67 GB	59 GB	3 h 37 m	57.9 GB	137 GB	2.34 TB	18 h 12 m	59 GB
Merge(4C, 4D)	3.8 B	8,000	17 GB	2 h 59 m	1.00 GB	118 GB	57 m	107 GB	118 GB	0 TB	3 h 56 m	108 GB
Merge(8AB, 8CD)	5.8 B	16,000	25 GB	4 h 53 m	1.60 GB	254 GB	2 h 10 m	232 GB	254 GB	0 TB	7 h 3 m	233 GB
VariMerge	5.8 B	16,000	137 GB	54 h 47 m	1.60 GB	254 GB	14 h 21 m	232 GB	254 GB	2.34 TB	69 h 8 m	233 GB

FM-Index and Backward Search

Advantages:

- Exact search in O(n).
- Compressed Suffix Array (CSA)


aac
aac
aac
aac
▮acaacg
aacg▮ac
acaacg▮
acg▮aca
caacg▮a
cg▮acaa
g▮acaac
LF(c,0)
LF(c,7)
LF(a,4)
LF(a,6)
LF(a,2)
LF(a,4)