HISAT2

Fast and sensitive alignment against general human population

Daehwan Kim
infphilo@gmail.com

History about BWT, FM, XBWT, GBWT, and GFM

- BWT (1994) BWT for Linear path
  - Burrows M, Wheeler DJ: A Block Sorting Lossless Data Compression Algorithm. Technical Report 124. Palo Alto, CA: Digital Equipment Corporation; 1994.
- FM (2000) BWT + metadata for fast lookup
  - Ferragina, P. & Manzini, G. Proc. 41st Annual Symposium on Foundations of Computer Science 390–398 (IEEE Comput. Soc.; 2000).
- XBWT (2009) BWT for Tree
  - P. Ferragina, F. Luccio, G. Manzini, and S. Muthukrishnan, "Compressing and Indexing Labeled Trees, with Applications," J. ACM, vol. 57, no. 1, p. 4, 2009.
- GCSA (2011, 2014) BWT for Graph
  - Sirén J, Välimäki N, Mäkinen V (2014) Indexing graphs for path queries with applications in genome research. IEEE/ACM Transactions on Computational Biology and Bioinformatics 11: 375–388. doi: 10.1109/tcbb.2013.2297101
- GFM and HGFM (2015) GBWT + metadata for fast lookup
  - Kim D., Paggi J., Salzberg S.L.

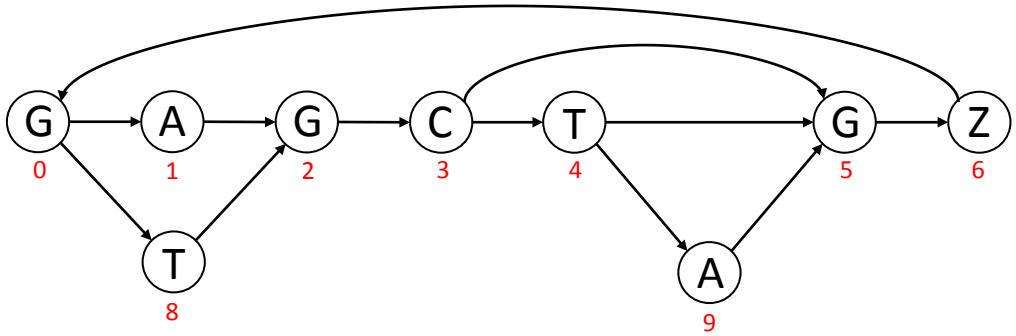
Small example

File #1: kim_example.fa

>kim_example
GAGCTG

File #2: kim_example.snp

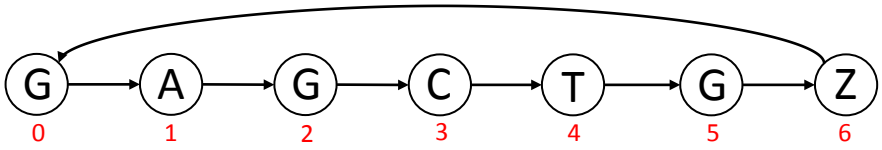
r01	single	kim_example	1	T
r02	deletion	kim_example	4	1
r03	insertion	kim_example	5	A

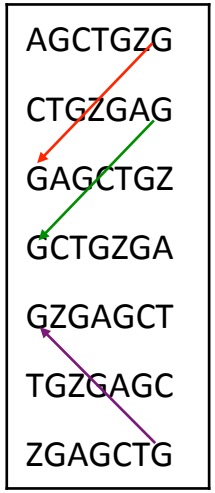


Requirement for GBWT

LF property

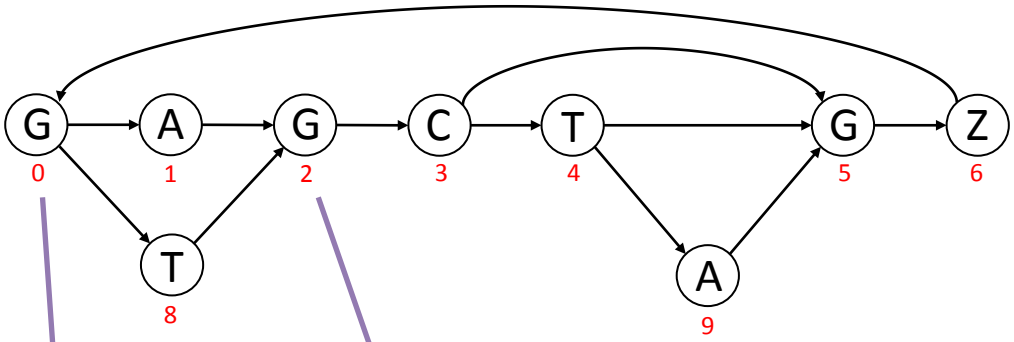
(LF: Last to First)





Requirement for GBWT

LF property

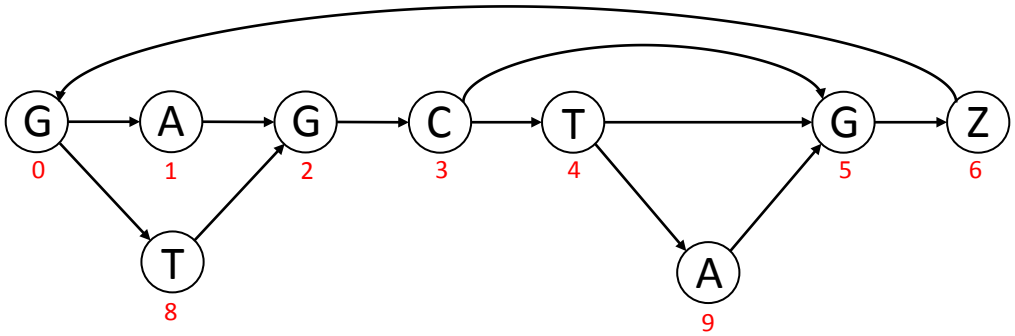


0	GAGCGZ
	GAGCTAGZ
	GAGCTGZ
	GTGCGZ
	GTGCTAGZ
	GTGCTGZ

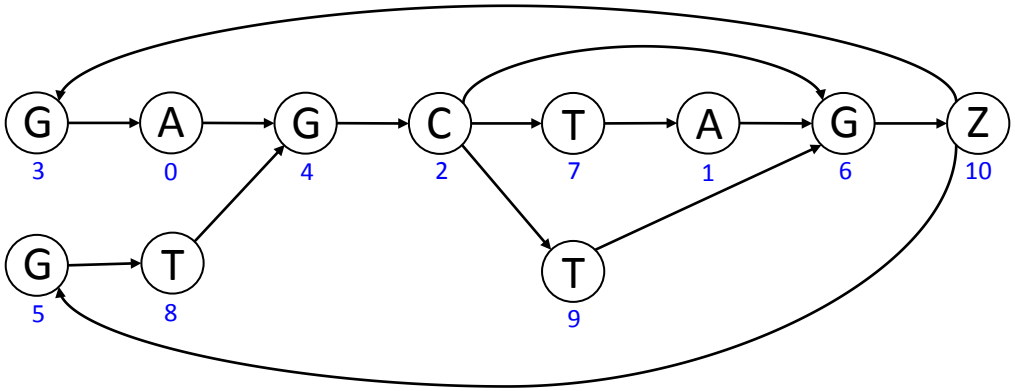
2	GCGZ
	GCTGZ
	GCTAGZ

First		Last
A (1)		G
…		…
C (3)		G
…		…
G (0)		…
G (2)		…
…		…
T (8)		G

Small example



Prefix-doubling and pruning



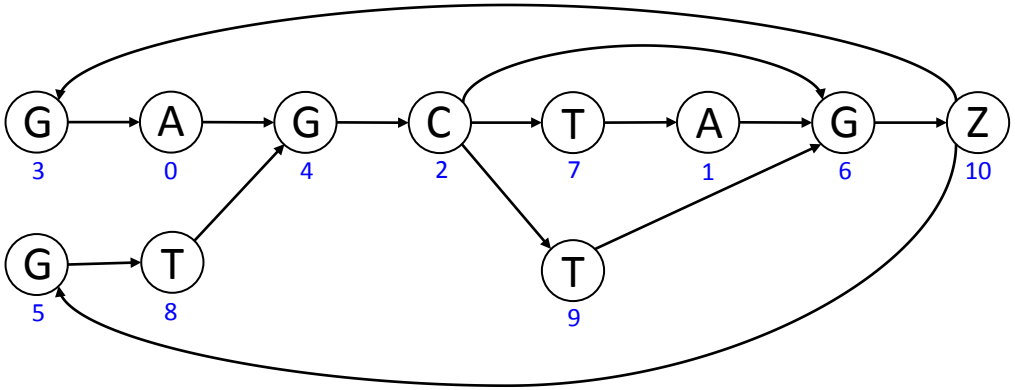
3	GAGCGZ
	GAGCTAGZ
	GAGCTGZ

4	GCGZ
	GCTAGZ
	GCTGZ

5	GTGCGZ
	GTGCTAGZ
	GTGCTGZ

6	GZ

Small example



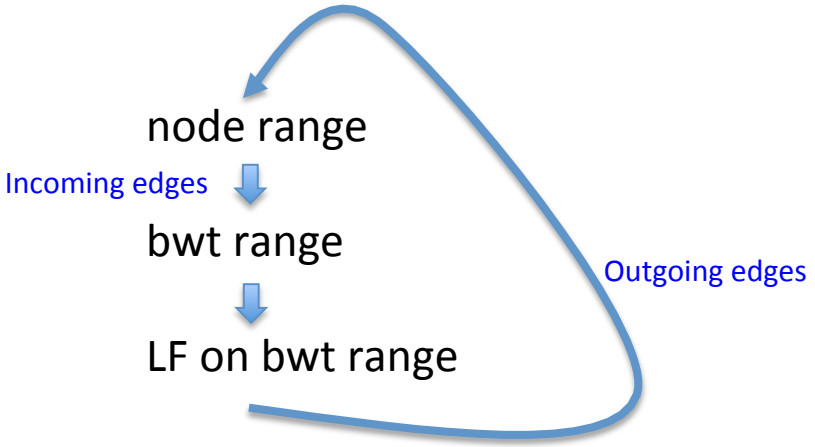
	Outgoing edge			Incoming edge	
	Node ID	First		Last	Node ID
0	0	A		G	0
1	1	A		T	1
2	2	C		G	2
3				Z	3
4				A	4
5	3	G		T	
6	4	G		Z	5
7	5	G		A	6
8	6	G		C	
9	7	T		T	
10	8	T		C	7
11	9	T		G	8
12	10	Z		C	9
13				G	10

Search basics in GBWT

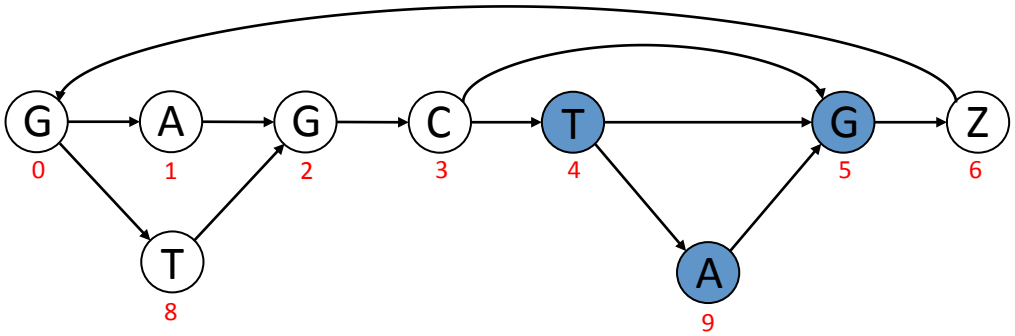
	Outgoing edge			Incoming edge	
	Node ID	First		Last	Node ID
0	0	A		G	0
1	1	A		T	1
2	2	C		G	2
3				Z	3
4				A	4
5	3	G		T	
6	4	G		Z	5
7	5	G		A	6
8	6	G		C	
9	7	T		T	
10	8	T		C	7
11	9	T		G	8
12	10	Z		C	9
13				G	10

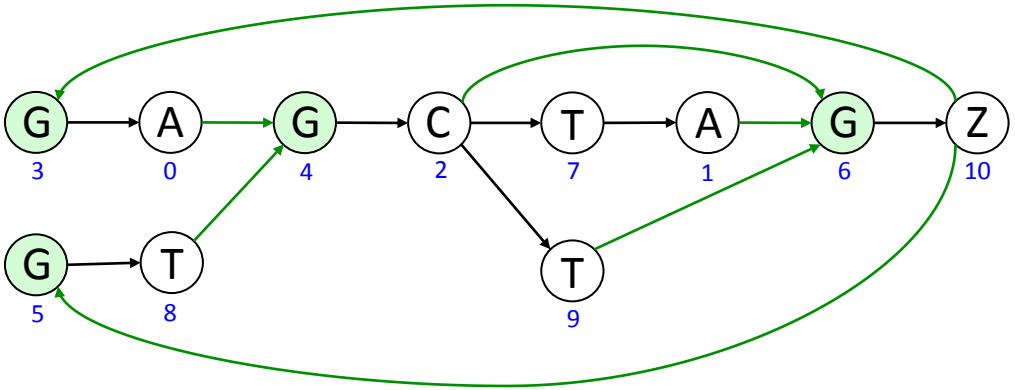
For a query string,

- we search base by base from right to left.
- two different ranges exist
  - node range [0, 10]
  - bwt range [0, 13]
- search:



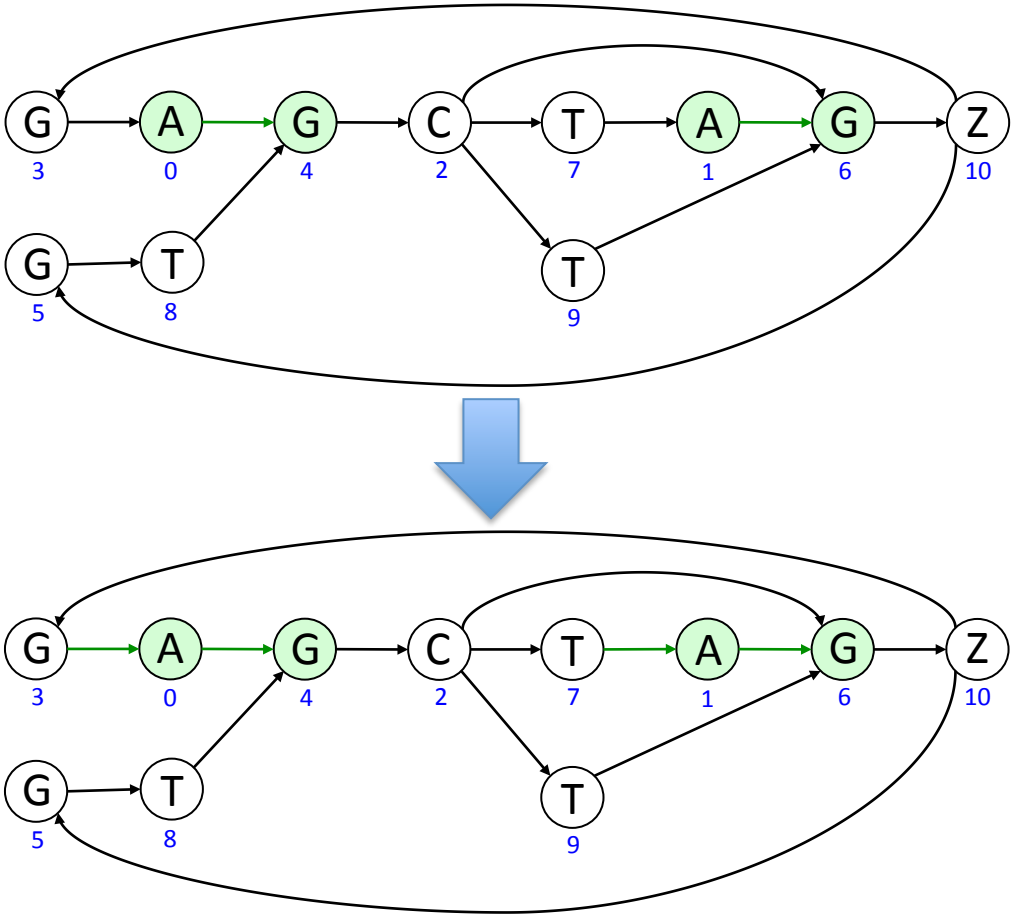
Small example - TAG





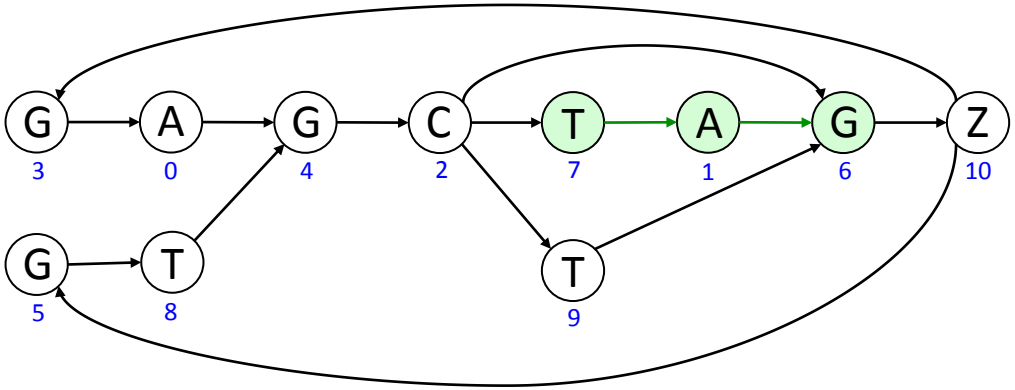
	Outgoing edge			Incoming edge	
	Node ID	First		Last	Node ID
0	0	A		G	0
1	1	A		T	1
2	2	C		G	2
3				Z	3
4				A	4
5	3	G		T	
6	4	G		Z	5
7	5	G		A	6
8	6	G		C	
9	7	T		T	
10	8	T		C	7
11	9	T		G	8
12	10	Z		C	9
13				G	10

Small example - TAG



	Outgoing edge			Incoming edge	
	Node ID	First		Last	Node ID
0	0	A		G	0
1	1	A		T	1
2	2	C		G	2
3				Z	3
4				A	4
5	3	G		T	
6	4	G		Z	5
7	5	G		A	6
8	6	G		C	
9	7	T		T	
10	8	T		C	7
11	9	T		G	8
12	10	Z		C	9
13				G	10

Small example - TAG



	Outgoing edge			Incoming edge	
	Node ID	First		Last	Node ID
0	0	A		G	0
1	1	A		T	1
2	2	C		G	2
3				Z	3
4				A	4
5	3	G		T	
6	4	G		Z	5
7	5	G		A	6
8	6	G		C	
9	7	T		T	
10	8	T		C	7
11	9	T		G	8
12	10	Z		C	9
13				G	10

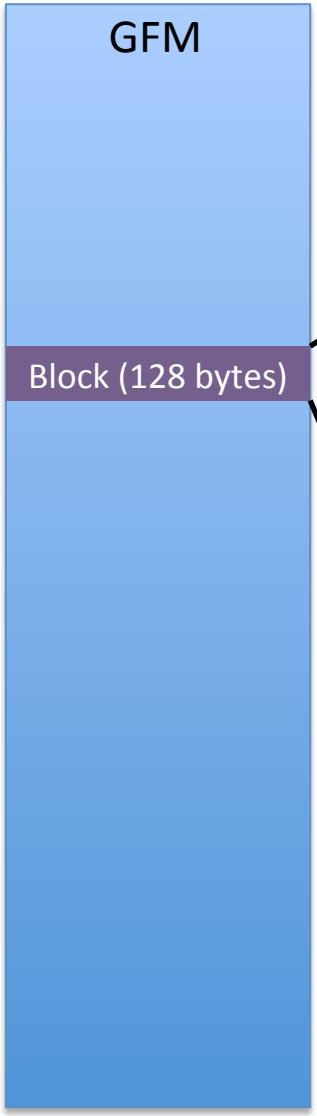
Node ID & BWT char **to** bits

	Outgoing edge			Incoming edge	
	Node ID	First		Last	Node ID
0	0	A		G	0
1	1	A		T	1
2	2	C		G	2
3				Z	3
4				A	4
5	3	G		T	
6	4	G		Z	5
7	5	G		A	6
8	6	G		C	
9	7	T		T	
10	8	T		C	7
11	9	T		G	8
12	10	Z		C	9
13				G	10

	Outgoing edge			Incoming edge	
	Node ID	First		Last	Node ID
0	1	2		10	1
1	1			11	1
2	1 0 0	3		10	1
3				00 (Z)	1
4				00	1 0
5	1	4		11	
6	1			00 (Z)	1
7	1			00	1 0 0
8	1			01	
9	1	3		11	
10	1			01	1
11	1			10	1
12	1 0	2		01	1
13				10	1

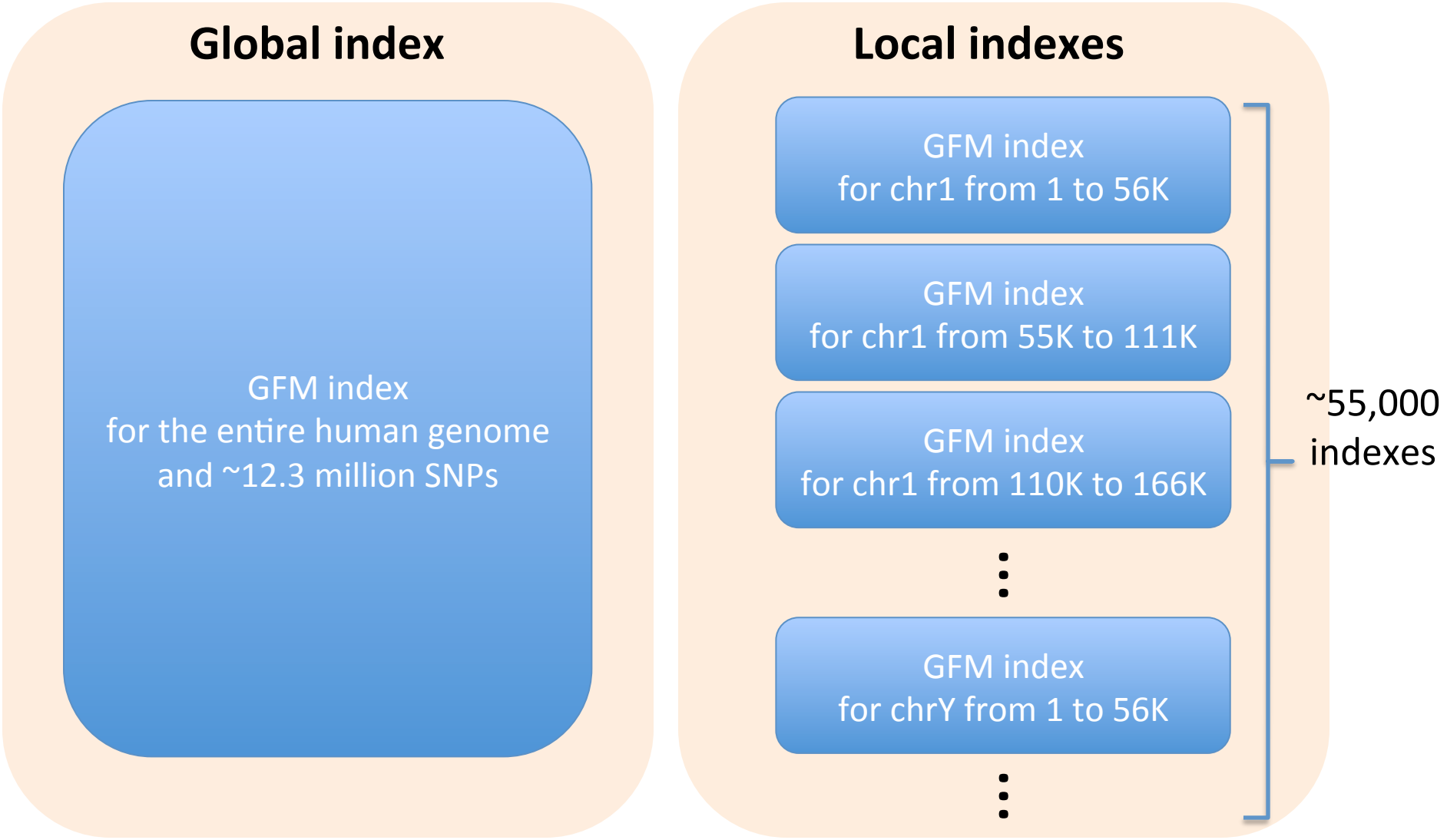
Graph FM index (GFM)

- Block size is 128 bytes (each cell represents four bytes in the table below, there are 32 cells).
  - Five 4-bytes (a total of 20 bytes) used to store numbers such as accumulated occurrences of A, C, G, T, and 1 bits in O array.
  - One 4-bytes used to store the corresponding location of 1 in IN for 1 in OUT.
  - Remaining 104 bytes used to represent 208 gbwt characters along, 208 bits from IN and OUT arrays each.



16 gbwt chars	16 gbwt chars	16 gbwt chars	16 gbwt chars
16 gbwt chars	16 gbwt chars	16 gbwt chars	16 gbwt chars
16 gbwt chars	16 gbwt chars	16 gbwt chars	16 gbwt chars
16 gbwt chars	32 IN bits	32 IN bits	32 IN bits
32 IN bits	32 IN bits	32 IN bits	16 IN bits / 16 OUT bits
32 OUT bits	32 OUT bits	32 OUT bits	32 OUT bits
32 OUT bits	32 OUT bits	Location of 1 in IN corresponding to 1 in OUT	Occurrences of 1 bits in OUT so far
Occurrences of A so far	Occurrences of C so far	Occurrences of G so far	Occurrences of T so far

HGFM – Hierarchical Graph FM index



HISAT2 index files

- Out1 (GFM)
  - Ref. names written at the end
- Out2 (Offset)
- Out3 (Ref. seq. contig information)
- Out4 (Ref. seq.)
- Out5 (Local GFMs)
- Out6 (Local offsets)
- Out7 (SNPs)
- Out8 (SNP IDs)

Index sizes for HISAT2, HISAT and Bowtie2

based on GRCh37 and ~12.3 million SNPs

HISAT2 (HGFM)		HISAT2 (HFM)		HISAT (HFM)		Bowtie2 (FM)	
genome.1.ht2	1863 MB	genome_FM.1.ht2	783 MB	genome.1.bt2	913 MB	genome.1.bt2	913 MB
genome.2.ht2	752 MB	genome_FM.2.ht2	682 MB	genome.2.bt2	682 MB	genome.2.bt2	682 MB
genome.3.ht2	1 MB	genome_FM.3.ht2	1 MB	genome.3.bt2	1 MB	genome.3.bt2	1 MB
genome.4.ht2	682 MB	genome_FM.4.ht2	682 MB	genome.4.bt2	682 MB	genome.4.bt2	682 MB
genome.5.ht2	1999 MB	genome_FM.5.ht2	1146 MB	genome.5.bt2	1145 MB		
genome.6.ht2	721 MB	genome_FM.6.ht2	695 MB	genome.6.bt2	693 MB		
genome.7.ht2	236 MB	genome_FM.7.ht2	1 M				
genome.8.ht2	129 MB	genome_FM.8.ht2	1 M				
						genome.rev,1.bt2	913 MB
						genome.rev.2.bt2	682 MB
Total	6.2 GB		3.9 GB		4.0 GB		3.8 GB

SAM output from HISAT2

read1 1 673181 100M

CTGAGGAAAGATGTTGAAATGTGACAAGTAAAGTAATATGAGTTCTTTTGACTATGTAAAATAATCAAACAAAAAATGACTTACTAAATTATAATACCCT

AS:i:0 XM:i:0 NM:i:0 MD:Z:50T49 Zs:Z:50|S|rs529937446

read2 1 769089 50M2D50M

ATTTCCTGAAAATAATATCCAAGATGCAAAGCATATGGCTCTGGTGAGACGTGTGAGGAGCTGAGAATGAGACGGCTGAGTGTCGGGGGCAGATCACGA

AS:i:0 XM:i:0 NM:i:0 MD:Z:50^AT50 Zs:Z:50|D|rs59306077

read3 1 843891 50M3I47M

GACAGGGGAGGTGACAGAGGGAGGGGAGGGGGAGGAGGGGCGGGGGAGAGGATGAGGGAAAAGGGGGAGGCGATGGGACGGGGGAGGGAATGGGGGAACA

AS:i:0 XM:i:0 NM:i:0 MD:Z:97 Zs:Z:50|I|rs199636838

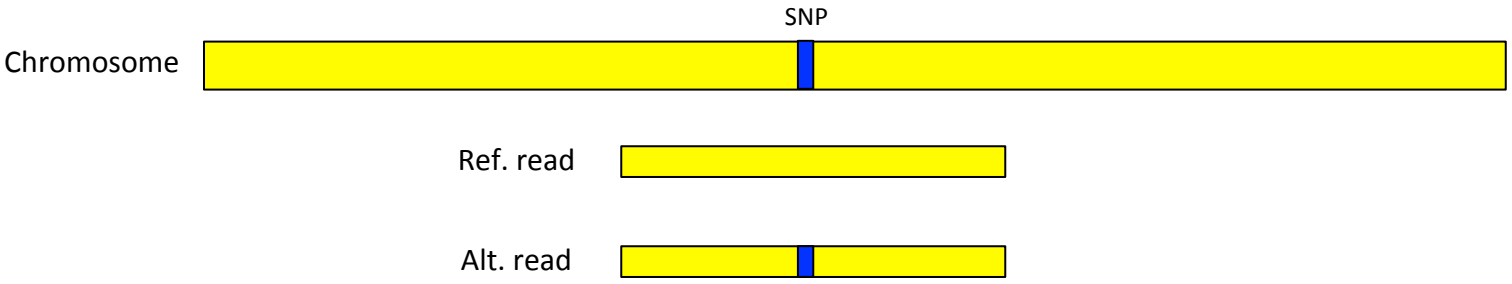
Two simulated data sets

From 12.3 million SNPs,

1) ref. reads: 12.3 million reads that are exactly the same as reference genome.

2) alt. reads: 12.3 million reads that contain a SNP in the middle and otherwise are exactly the same as reference genome.

(Reads are 100-bp long and single-end.)

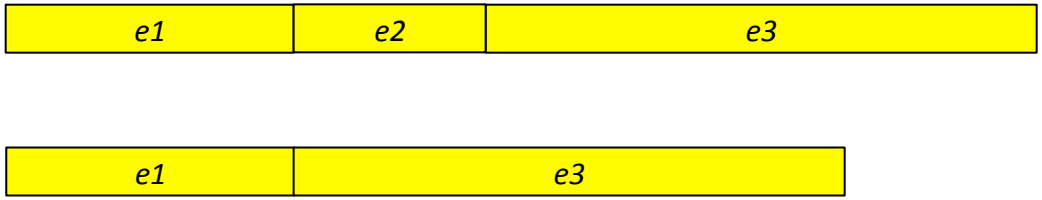


Runtime and alignment sensitivity for HISAT2, HISAT and Bowtie2

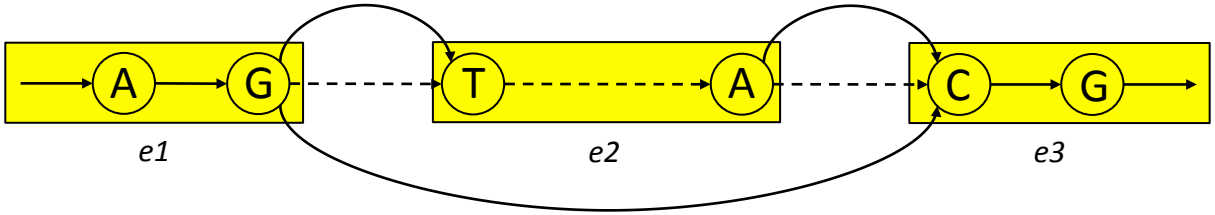
	Ref. reads				Alt. reads			
	Runtime	Memory	Alignment sensitivity (first)	Alignment sensitivity (+others)	Runtime	Memory	Alignment sensitivity (first)	Alignment sensitivity (+others)
HISAT2 (HGFM)	8 m 53 s	6.7 GB	98.56 %	99.14 %	9 m 54 s	6.7 GB	98.95 %	99.31 %
HISAT2 (HFM)	4 m 17 s	4.0 GB	98.92 %	99.48 %	7 m 40 s	4.0 GB	94.77 %	95.15 %
HISAT (HFM)	5 m 16 s	4.1 GB	98.71 %	99.25 %	8 m 25 s	4.1 GB	94.98 %	95.41 %
Bowtie2 (FM) **-k 5**	5 m 30 s	3.1 GB	98.93 %	99.54 %	<===== **--score-min C,0** to use FM index only (without seeding and dynamic programming)			
Bowtie2 (FM) **-k 5**	=====> Use the default setting to allow for mismatches and indels				1 h 48 m	3.1 GB	97.90 %	98.94 %

Transcriptome mapping

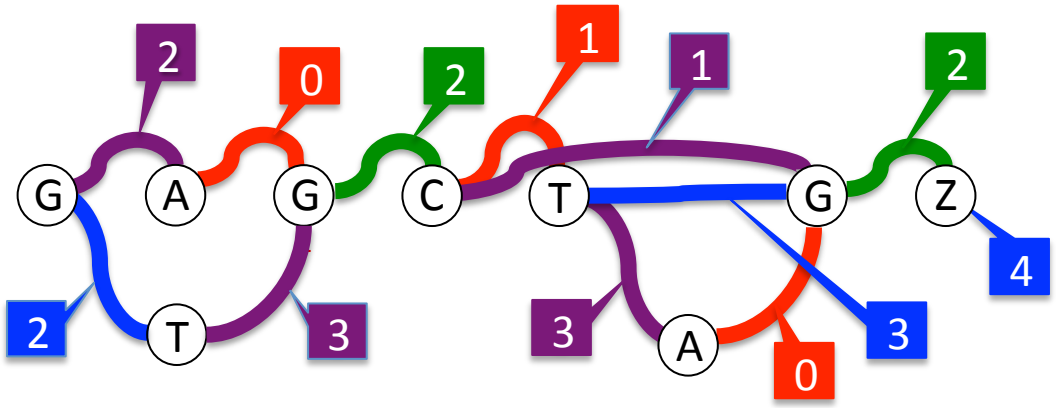
TopHat2

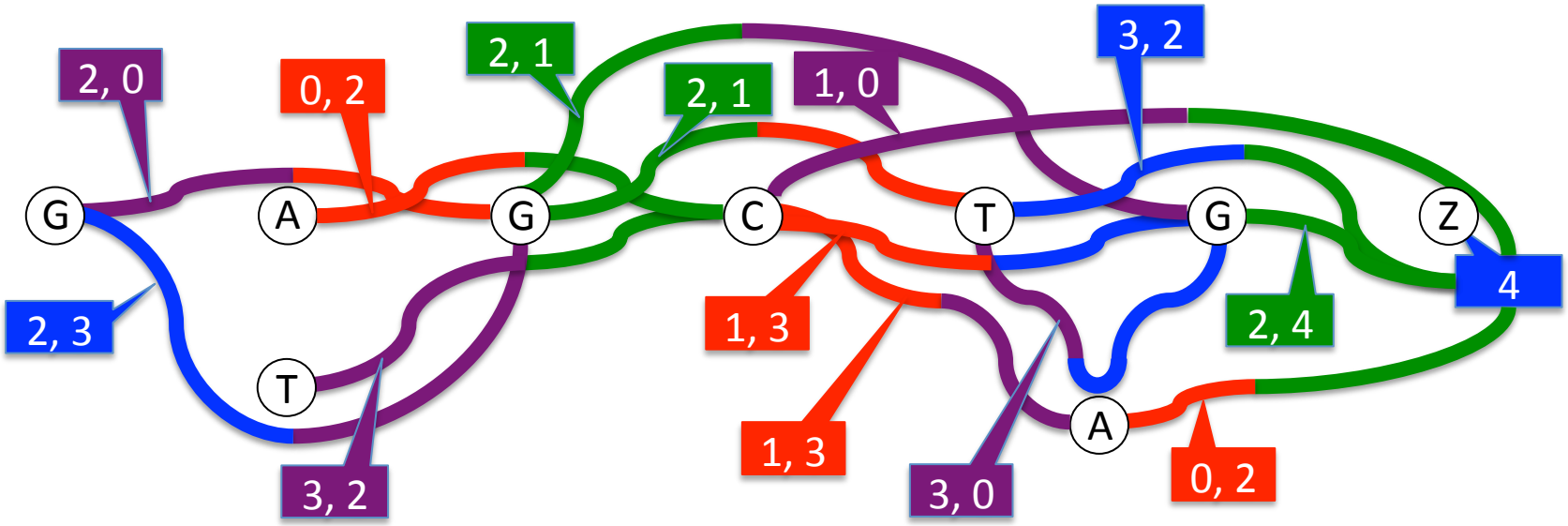


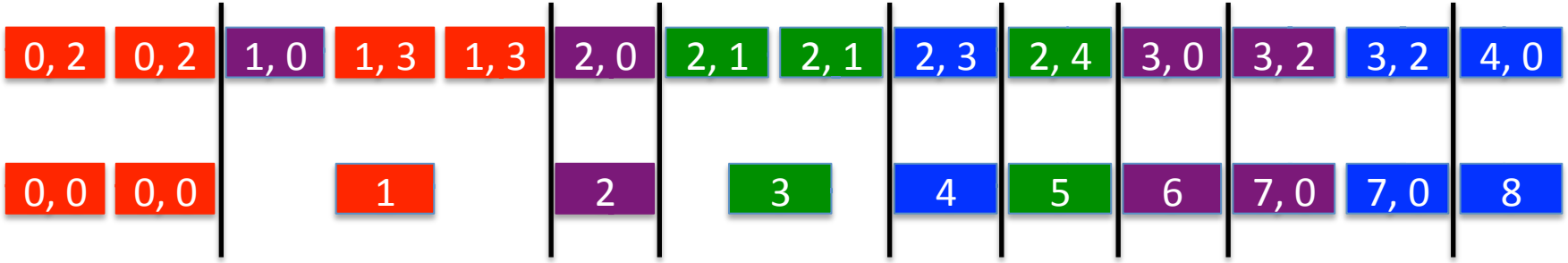
HISAT2

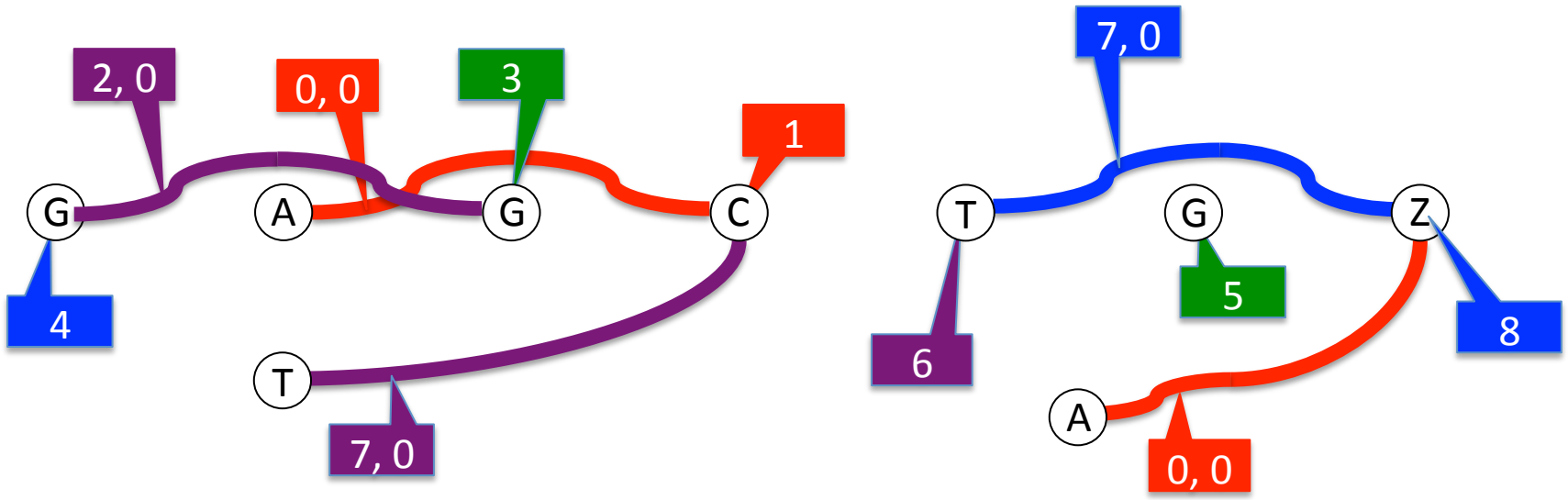


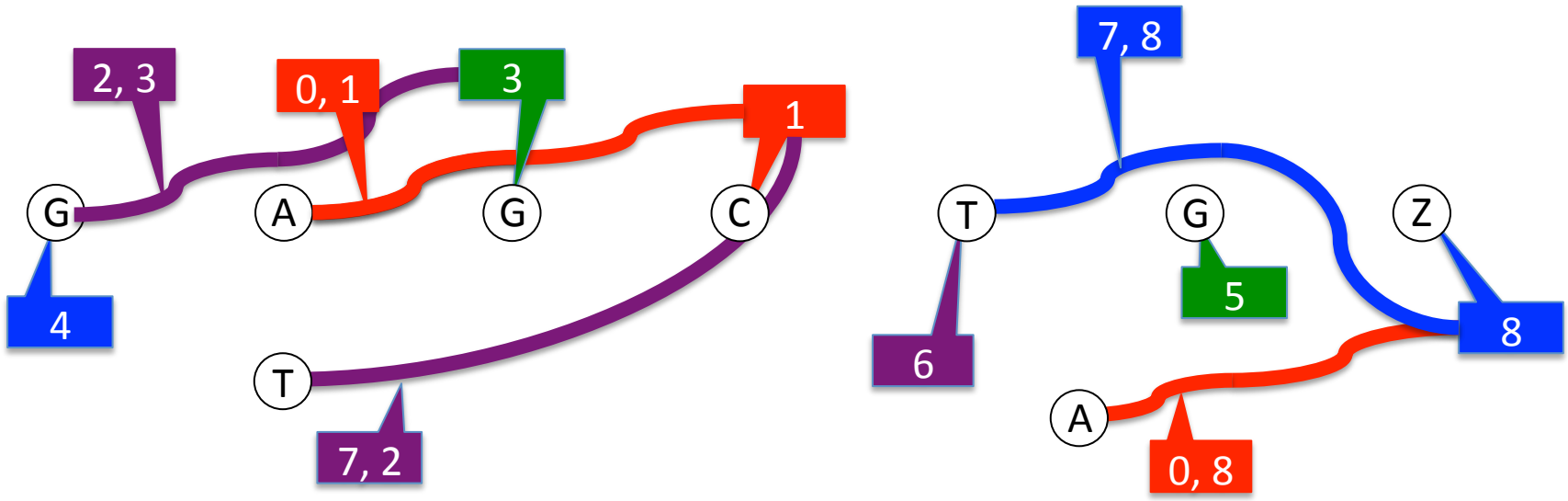
Path-doubling example

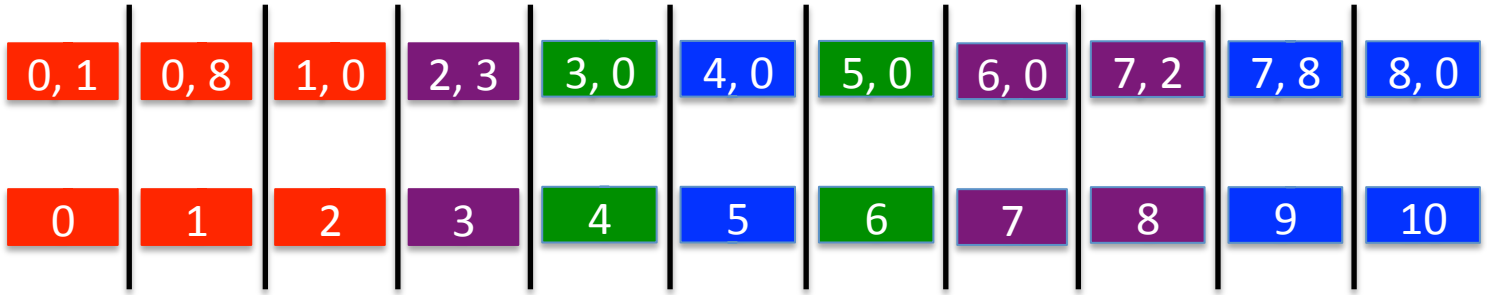


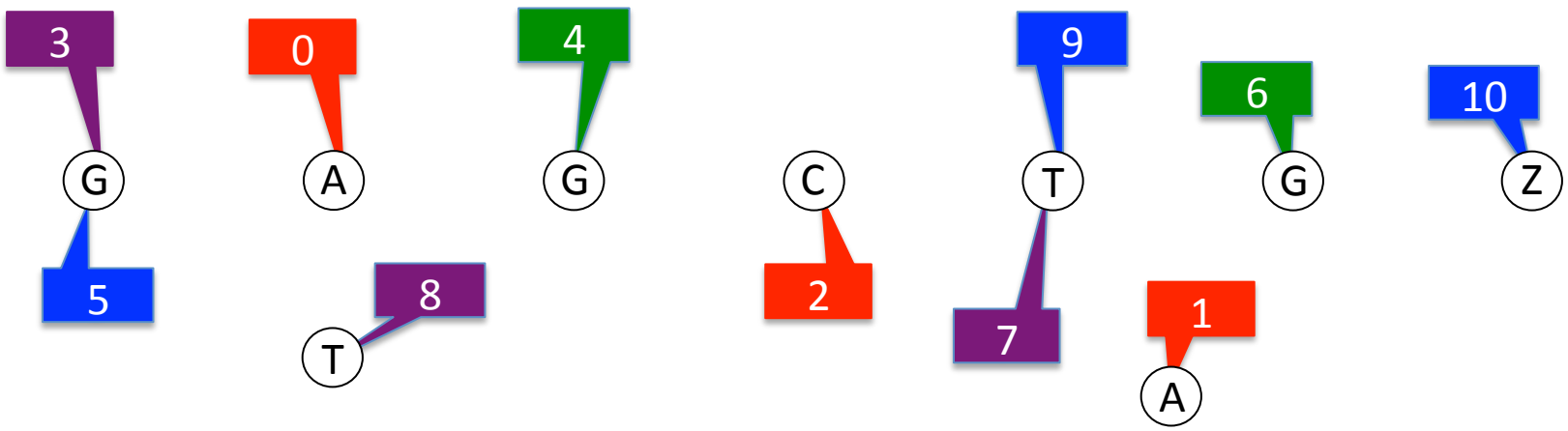

2, 0
0, 2
2, 1
2, 1
1, 0
3, 2
G
A
G
C
T
G
Z
2, 3
1, 3
2, 4
4
T
1, 3
A
3, 2
3, 0
0, 2

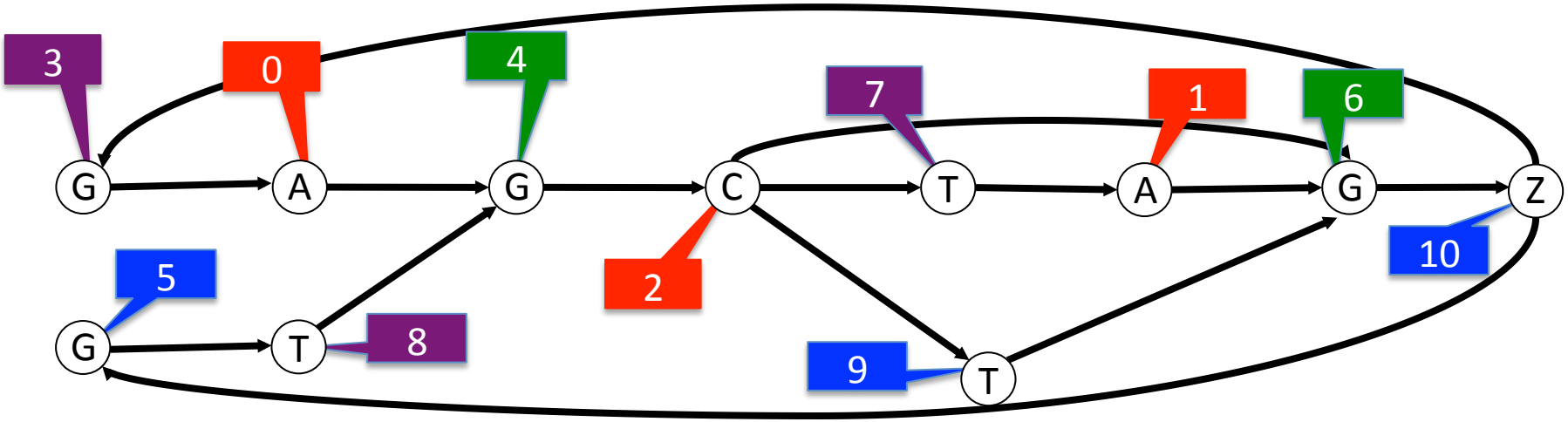

0, 2
0, 2
1, 0
1, 3
1, 3
2, 0
2, 1
2, 1
2, 3
2, 4
3, 0
3, 2
3, 2
4, 0
0, 0
0, 0
1
2
3
4
5
6
7, 0
7, 0
8


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


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


0, 1
0, 8
1, 0
2, 3
3, 0
4, 0
5, 0
6, 0
7, 2
7, 8
8, 0
0
1
2
3
4
5
6
7
8
9
10


3
G
5
0
A
8
T
4
G
C
2
9
T
7
6
G
1
A
10
Z


3
0
4
7
1
6
G
A
G
C
T
A
G
Z
5
2
10
G
T
8
9
T