Exercise sheet 10: Introduction to Mapping

Exercise 1

Question 1A Extract the Burrows-Wheeler Transform B(S) of $S = TGGTGGTTGA\$$.

Solution $B(S) = AGTTTGGT\$GG$

i	SA[i]	SA[i]-th suffix	Rotation	Last column
1	11	$	$TGGTGGTTGA	A
2	10	A$	A$TGGTGGTTG	G
3	9	GA$	GA$TGGTGGTT	T
4	2	GGTGGTTGA$	GGTGGTTGA$T	T
5	5	GGTTGA$	GGTTGA$TGGT	T
6	3	GTGGTTGA$	GTGGTTGA$TG	G
7	6	GTTGA$	GTTGA$TGGTG	G
8	8	TGA$	TGA$TGGTGGT	T
9	1	TGGTGGTTGA$	TGGTGGTTGA$	$
10	4	TGGTTGA$	TGGTTGA$TGG	G
11	7	TTGA$	TTGA$TGGTGG	G

Question 1B Invert the Burrows-Wheeler-Transform $B(S) = TCAACT\$AA$ to obtain S.

Hint Step-by-step method:

Add
T
C
A
A
C
T
!
A
A

→

Sort
!
A
A
A
A
C
C
T
T

Add	
T	!
C	A
A	A
A	A
C	A
T	C
!	C
A	T
A	T

→

Sort	
!	C
A	A
A	A
A	T
A	T
C	A
C	A
T	!
T	C

Add		
T	!	C
C	A	A
A	A	A
A	A	T
C	A	T
T	C	A
!	C	A
A	T	!
A	T	C

→

Sort		
!	C	A
A	A	A
A	A	T
A	T	!
A	T	C
C	A	A
C	A	T
T	!	C
T	C	A

Add			
T	!	C	A
C	A	A	A
A	A	A	T
A	A	T	!
C	A	T	C
T	C	A	A
!	C	A	T
A	T	!	C
A	T	C	A

→

Sort			
!	C	A	T
A	A	A	T
A	A	T	!
A	T	!	C
A	T	C	A
C	A	A	A
C	A	T	C
T	!	C	A
T	C	A	A

Add				
T	!	C	A	T
C	A	A	A	T
A	A	A	T	!
A	A	T	!	C
C	A	T	C	A
T	C	A	A	A
!	C	A	T	C
A	T	!	C	A
A	T	C	A	A

→

Sort				
!	C	A	T	C
A	A	A	T	!
A	A	T	!	C
A	T	!	C	A
A	T	C	A	A
C	A	A	A	T
C	A	T	C	A
T	!	C	A	T
T	C	A	A	A

Add					
T	!	C	A	T	C
C	A	A	A	T	!
A	A	A	T	!	C
A	A	T	!	C	A
C	A	T	C	A	A
T	C	A	A	A	T
!	C	A	T	C	A
A	T	!	C	A	T
A	T	C	A	A	A

→

Sort					
!	C	A	T	C	A
A	A	A	T	!	C
A	A	T	!	C	A
A	T	!	C	A	T
A	T	C	A	A	A
C	A	A	A	T	!
C	A	T	C	A	A
T	!	C	A	T	C
T	C	A	A	A	T

Add						
T	!	C	A	T	C	A
C	A	A	A	T	!	C
A	A	A	T	!	C	A
A	A	T	!	C	A	T
C	A	T	C	A	A	A
T	C	A	A	A	T	!
!	C	A	T	C	A	A
A	T	!	C	A	T	C
A	T	C	A	A	A	T

→

Sort						
!	C	A	T	C	A	A
A	A	A	T	!	C	A
A	A	T	!	C	A	T
A	T	!	C	A	T	C
A	T	C	A	A	A	T
C	A	A	A	T	!	C
C	A	T	C	A	A	A
T	!	C	A	T	C	A
T	C	A	A	A	T	!

Add							
T	!	C	A	T	C	A	A
C	A	A	A	T	!	C	A
A	A	A	T	!	C	A	T
A	A	T	!	C	A	T	C
C	A	T	C	A	A	A	T
T	C	A	A	A	T	!	C
!	C	A	T	C	A	A	A
A	T	!	C	A	T	C	A
A	T	C	A	A	A	T	!

→

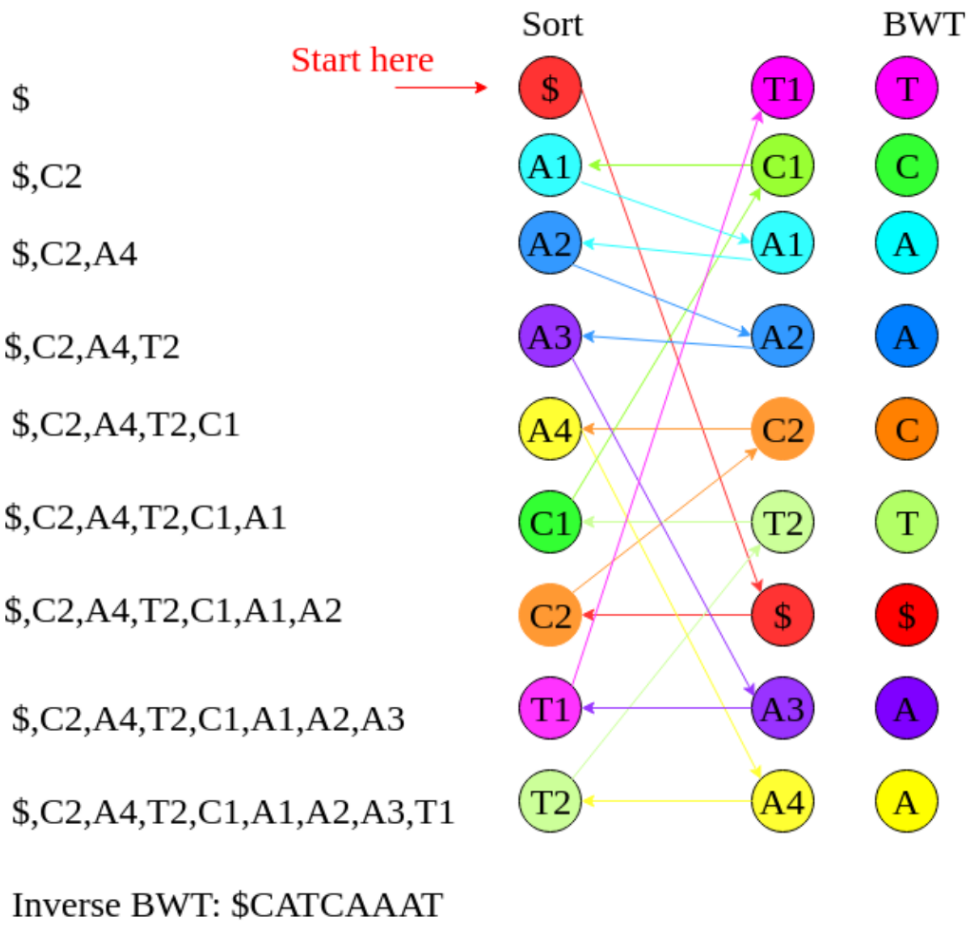
Sort							
!	C	A	T	C	A	A	A
A	A	A	T	!	C	A	T
A	A	T	!	C	A	T	C
A	T	!	C	A	T	C	A
A	T	C	A	A	A	T	!
C	A	A	A	T	!	C	A
C	A	T	C	A	A	A	T
T	!	C	A	T	C	A	A
T	C	A	A	A	T	!	C

Add								
T	!	C	A	T	C	A	A	A
C	A	A	A	T	!	C	A	T
A	A	A	T	!	C	A	T	C
A	A	T	!	C	A	T	C	A
C	A	T	C	A	A	A	T	!
T	C	A	A	A	T	!	C	A
!	C	A	T	C	A	A	A	T
A	T	!	C	A	T	C	A	A
A	T	C	A	A	A	T	!	C

→

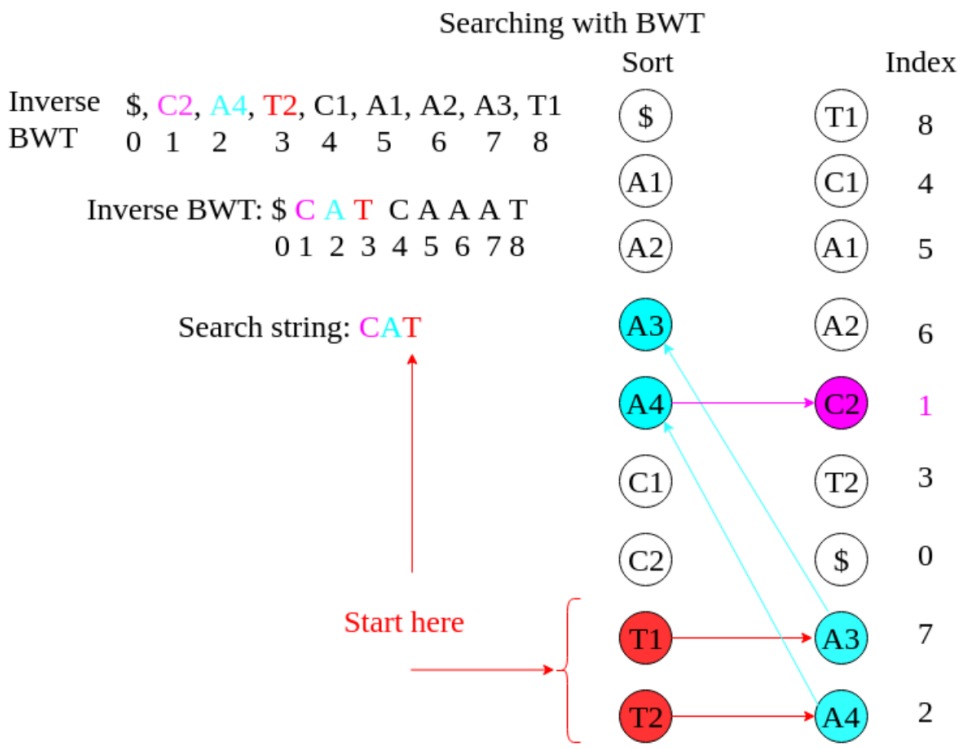
Sort								
!	C	A	T	C	A	A	A	T
A	A	A	T	!	C	A	T	C
A	A	T	!	C	A	T	C	A
A	T	!	C	A	T	C	A	A
A	T	C	A	A	A	T	!	C
C	A	A	A	T	!	C	A	T
C	A	T	C	A	A	A	T	!
T	!	C	A	T	C	A	A	A
T	C	A	A	A	T	!	C	A

Solution



Question 1C Search for CAT in $B(S) = TCAACT\$AA$.

Solution



Exercise 2

Read the publication on the bwa-mem aligner at https://arxiv.org/abs/1303.3997 and pay particular attention to the re-seeding and chaining features of the algorithm. Now consider a read $R = CCCCGTTTT$ and a reference genome $T = ...CCCCATTTT...CCCCGA...AGTTTT...$ and explain step-by-step how re-seeding and chaining let bwa-mem let recover the correct best alignment of R to T.

Question 2A What are the original SMEMs that get generated?

Solution The original MEMs are CCCCG and GTTTT.

Question 2B Would these SMEMs lead to discovery of the best possible alignment?

Solution No, because it would not lead to a best possible match in the reference genome. The current seeds are too specific and we may miss the seeds that lead to the best mapping (CCC- CATTTT), therefore we need to reseed.

Question 2C Which shortened new SMEMs are discovered with re-seeding (assume re-seeding gets performed despite the below-threshold length of the SMEMs)?

Solution Re-seeding around the central base of each MEM leads to discovery of CCCC and TTTT (both occur once more often than the originals).

Question 2D What is the effect of chaining of colinear seeds?

Solution The two new seeds are colinear on R and T , and can be merged into a chain, so only one local alignment has to be performed.