

## Exercise sheet 10: Introduction to Mapping

---

### Exercise 1

1a)

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

Hide

**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

1b)

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

Hide

**Hint 1** You can get the first column  $F(S)$  of the Burrows Wheeler Matrix via counting and sorting the letters since it has a very predictable structure.

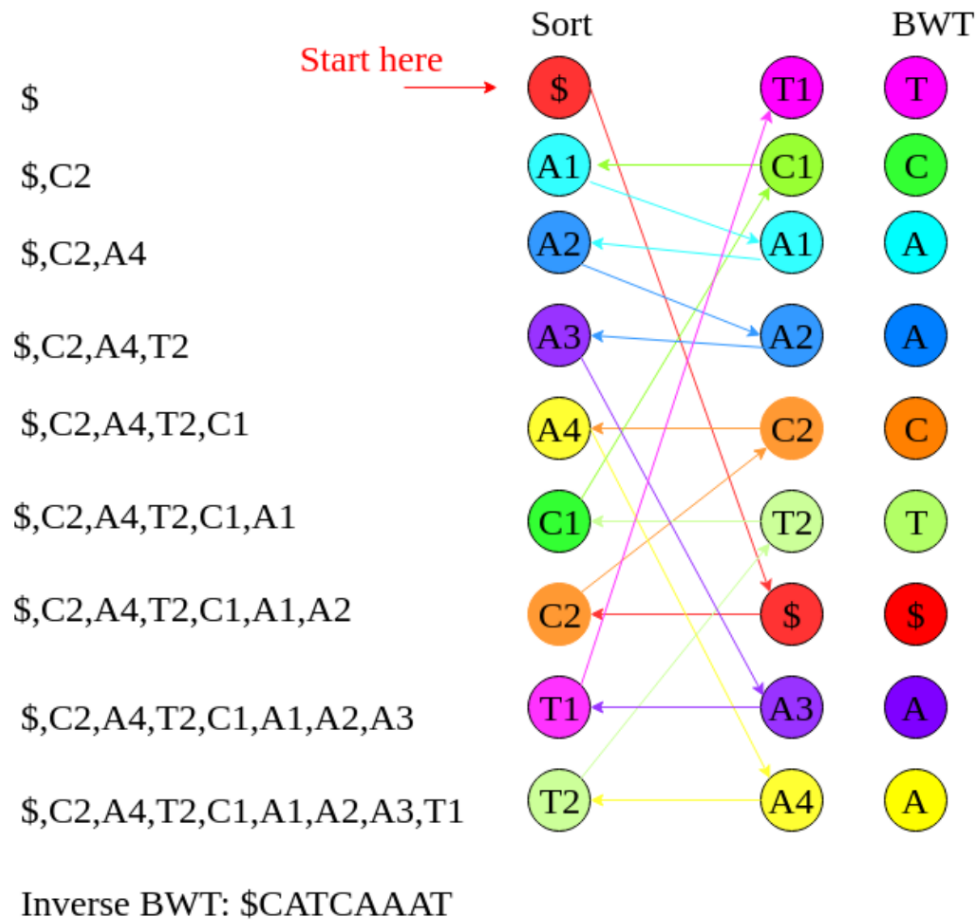
$$F(S) = \$AAAACCTT$$

**Hint 2** Use the last-first mapping to assign indices to the corresponding letters in the first  $F(S)$  and the last column  $B(S)$

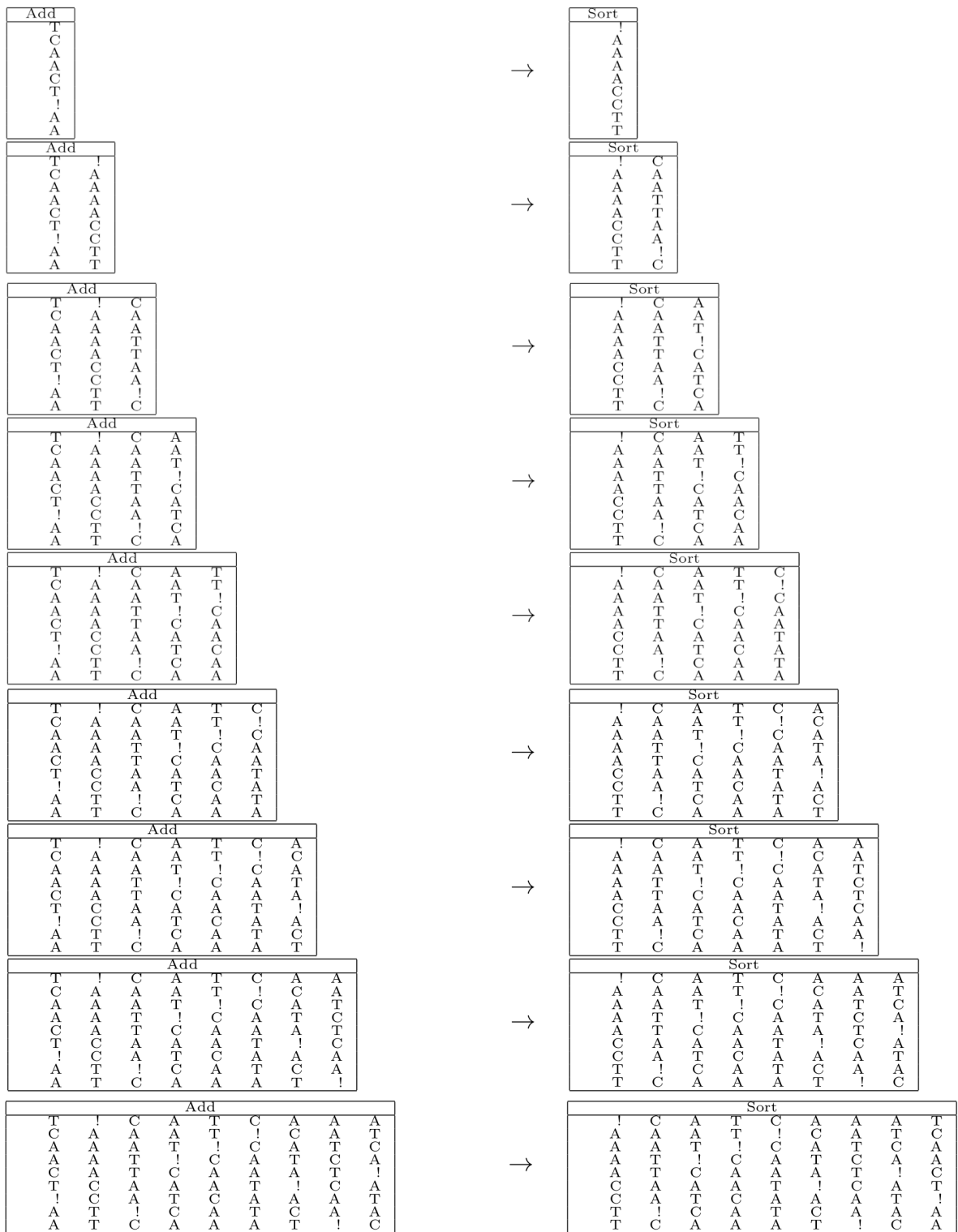
$$F(S) = \$_1 A_1 A_2 A_3 A_4 C_1 C_2 T_3 T_4$$

$$B(S) = T_1 C_1 A_1 A_2 C_2 T_2 \$ _1 A_3 A_4$$

**Solution 1**



**Solution 2** Step-by-step method:



1c)

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

Hide

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  $R = CCCCCGTTTT$  and a reference genome  $T = \dots CCCCATTTT \dots CCCCAGA \dots AGTTTT \dots$  and explain step-by-step how re-seeding and chaining let bwa-mem let recover the correct best alignment of  $R$  to  $T$ .

2a)

What are the original SMEMs that get generated?

Hide

**Solution** The original MEMs are CCCCCG and GTTTT.

**2b)**

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

**Hide**

**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.

**2c)**

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

**Hide**

**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).

**2d)**

What is the effect of chaining of colinear seeds?

**Hide**

**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.

---

## Exercise 3 - Programming assignment

Under Construction

---