

# Assignment 3, 3620 - Solutions

① a) (i)

00001 (p)  
-----0[0]... (t)  
          ↓  
          bad symbol.  
Shift is 1.

text: 0000[0] ..... 0  
      0000[1] ..... 0  
          ↑  
      first comparison

Since first comparison is failure, one comparison for each position of the pattern

⇒ 996 comparisons

ii) For pattern 1000, the shift for bad symbol '0' is also 1, we align pattern to text also in 996 positions. However, for each position we do 4 successful comparisons before the fifth is a failure.

⇒  $5 \times 996 = 4980$  comparisons

iii) For pattern 01010, the shift for symbol 0 is 2, so we only align pattern  $996/2 = 498$  times. For each alignment there are 2 comparisons ⇒ 996 comparisons.

ii) i) For pattern 00001, we do not have successful comparison, so this case is same as for Horpool's algorithm: 996 comparisons.

ii) pattern 10000, we will have 4 successful matches. Since the 4-char suffix of pattern doesn't occur anywhere else in pattern, we shift by 5 characters. Counting the unsuccessful comparison, we have 5 comparisons given one alignment of the pattern, so the total # of comparisons equals the # of characters in text = 1000.

iii) For 01010 the good shift for 1 successful comparison is 2 as in bad symbol table, so the # of comparisons is same as in a) = 996.

②. a)

Pattern: TCCTATTCTT

Horpool's table	Bad symbol	Shift
	A	5
	C	2
	G	10
	T	1

2. b)

k	ref(k)	shift
1	T	1
2	TT	3
3	CTT	9
4	TCTT	9
5	⋮	⋮
6	⋮	⋮
7	⋮	⋮
8	⋮	⋮
9	⋮	9

c)

TTATAGATCTCGTATTCTT  
 TCCTATTCTT  
 ...CTT  
 ..CTT  
 ...CTT  
 ...CTATTCTT

mismatch  $\rightarrow$  successful

TTATAGATCTCCTATTCTT  
 ...CTT  
 ...CTT  
 ...CTT  
 CTATTCTT  
 ...CTT  
 ...CTT

found!



## Marking scheme

### Problem 1

- 1 pt of answer in each combination pattern - algorithm is about 60% correct
- 0.5 pts if answer is partially correct (< 60%)
- 0 pts if answer is not correct.

### Problem 2

- 2 pts max for each subproblem
  - for a & b allocate - 1 pt if general outline of table is OK (or for a) there should be 4 entries in table)
  - 1 pt allocated for content.

### Problem 3

- 2 pts for each subproblem