Homework 3

CS 4364/5364
Spring 2021

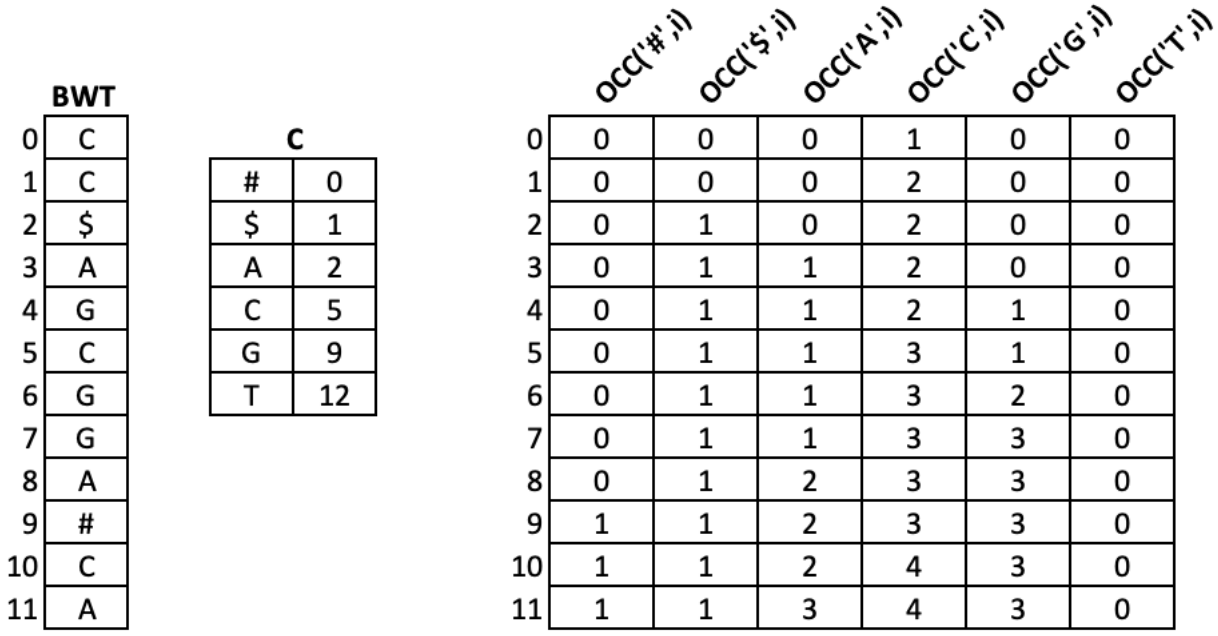
Due: 11 March 2021

Because of the reliance of the particular assignments in this class on mathematical notation, and the fact that all assignments will be submitted electronically, students are encouraged to use LaTeX to formalize their responses. **For those enrolled in the graduate section the use of latex is *required*.** This assignment (like all others) will be posted on the course `github`[1] as source code as well as in PDF form on the course website. Please submit your assignment to the professor via email, either as a link to your assignment online (i.e. overleaf or github) or as an attachment. Graduate students will need to include the `.tex` files as well as a PDF, this is optional but encouraged for undergraduates.

1. **(30 points)** We know that just like the Suffix Tree and Suffix Array an BWT (and in turn an FM Index) can have contain the suffixes for more than one string. Assume we have constructed an FM Index for two strings. Assume that the first string is terminated by the character `#` and the second with `$` (`#` is lexicographically smaller than `$`), neither of which is in Σ.

    - Describe an algorithm to reconstruct all (both) strings contained in an FM Index.
    - Use your algorithm to reconstruct the strings given the FM Index in Figure 1 below.
    - Describe an algorithm to use the same index to find out how many times a pattern P is present in each of the strings (the output will be two different numbers).

[1] `github.com/deblasiolab/CS4364-documents`



	BWT
0	C
1	C
2	$
3	A
4	G
5	C
6	G
7	G
8	A
9	#
10	C
11	A

C	
#	0
$	1
A	2
C	5
G	9
T	12

	OCC('#',i)	OCC('$',i)	OCC('A',i)	OCC('C',i)	OCC('G',i)	OCC('T',i)
0	0	0	0	1	0	0
1	0	0	0	2	0	0
2	0	1	0	2	0	0
3	0	1	1	2	0	0
4	0	1	1	2	1	0
5	0	1	1	3	1	0
6	0	1	1	3	2	0
7	0	1	1	3	3	0
8	0	1	2	3	3	0
9	1	1	2	3	3	0
10	1	1	2	4	3	0
11	1	1	3	4	3	0

Figure 1: Example Ferragina Manzini Index