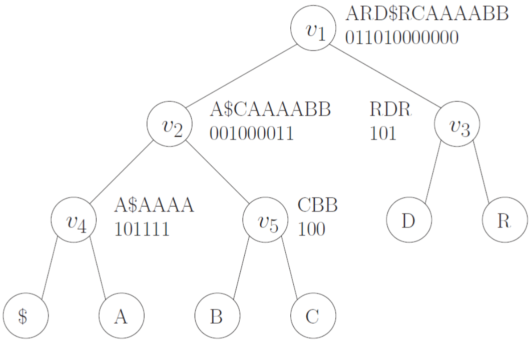
582487 Data Compression Techniques (Spring 2015)

Exercise 5 (February 18th)

Solve the following problem before the exercise session and be prepared to present your solutions at the session.

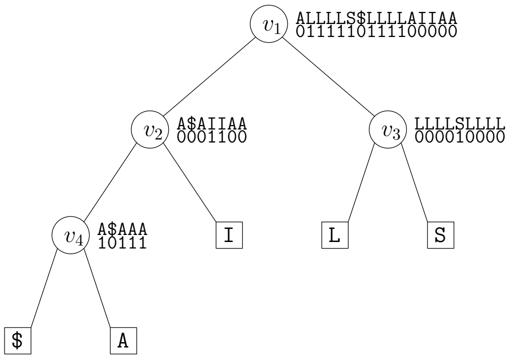
Below are a wavelet tree for the BWT of `ABRACADABRA$` and the function $C(a)$ that returns the partial sums of the characters' frequencies. Suppose each node v supports the queries $v.\mathsf{rank}_0(i)$ and $v.\mathsf{rank}_1(i)$ on the binary sequence it stores. Considering the tree as an FM-index, list the rank queries used to count the number of occurrences of `BRA` in `ABRACADABRA$`.



a	$	A	B	C	D	R
$C(a)$	0	1	6	8	9	10

Note: Travis still hasn't had a chance to update the slides from a few years ago but they're posted anyway. Since he really wants you to understand this stuff, though, on Tuesday the 17th there'll be an optional tutorial (replacing Simon's cancelled class) in which he'll go through the following example:

Problem: Below are a wavelet tree for the BWT of `ILLALLASILLALLA$` and the function $C(a)$ that returns the partial sums of the characters' frequencies. Suppose each node v supports the queries $v.\mathsf{rank}_0(i)$ and $v.\mathsf{rank}_1(i)$ on the binary sequence it stores. Considering the tree as an FM-index, list the rank queries used to count the number of occurrences of `LAL` in `ILLALLASILLALLA$`. (The solution is on the other side of the page.)



a	$	A	I	L	S
$C(a)$	0	1	5	7	15

Solution: We can see from C that there are 8 Ls in $\mathsf{BWT}(S_1)$, so the ranks of the first and last in $\mathsf{BWT}(S_1)[1..16]$ are 1 and 8. Adding $C[\mathrm{L}] = 7$ to 1 and 8 we get 8 and 15, meaning the lexicographically 8th through 15th suffixes of S_1 start with L.

To compute the rank $\mathsf{BWT}(S_1).\mathsf{rank}_{\mathrm{A}}(8-1)+1 = 2$ of the first A in $\mathsf{BWT}(S_1)[8..15]$, we compute

- $v_1.\mathsf{rank}_0(7) = 2$,
- $v_2.\mathsf{rank}_0(2) = 2$,
- $v_4.\mathsf{rank}_1(2) = 1$.

To compute the rank $\mathsf{BWT}(S_1).\mathsf{rank}_{\mathrm{A}}(15) = 3$ of the last A in $\mathsf{BWT}(S_1)[8..15]$, we compute

- $v_1.\mathsf{rank}_0(15) = 6$,
- $v_2.\mathsf{rank}_0(6) = 4$,
- $v_4.\mathsf{rank}_1(4) = 3$.

Adding $C[\mathrm{A}] = 1$ to 2 and 3 we get 3 and 4, meaning the lexicographically 3rd through 4th suffixes of S_1 start with AL.

To compute the rank $\mathsf{BWT}(S_1).\mathsf{rank}_{\mathrm{L}}(3-1)+1 = 2$ of the first L in $\mathsf{BWT}(S_1)[3..4]$, we compute

- $v_1.\mathsf{rank}_1(2) = 1$,
- $v_3.\mathsf{rank}_0(1) = 1$.

To compute the rank $\mathsf{BWT}(S_1).\mathsf{rank}_{\mathrm{L}}(4) = 3$ of the last L in $\mathsf{BWT}(S_1)[3..4]$, we compute

- $v_1.\mathsf{rank}_1(4) = 3$,
- $v_3.\mathsf{rank}_0(3) = 3$.

Adding $C[\mathrm{L}] = 7$ to 2 and 3 we get 9 and 10, meaning the lexicographically 9th and 10th suffixes of S_1 start with LAL.

Since 10 - 9 + 1 = 2, we report that there are 2 occurrences of LAL.