Succinct data structures.

Rank & Select

- rank (i) - #1 upto position i — obě jdou udělat v $O(1)$
- select (i) - i-tá #1 (pozice)

reprezentace binárního stromu (pouze jeho tvaru)

- každému uzlu přidám dva potomky
- vytvořím binární vektor kde původní uzly jsou 1, moje potomci 0
  → indexuji rovnou po řádcích

left_child = 2. rank (i)
right_child (i) = 2. rank (i) + 1
parent(i) = select $\left(\lfloor \frac{i}{2} \rfloor\right)$

- reprezentace má $2n+1$ bitů

general trees - každý uzel reprezentován $1^m 0$ kde m je počet potomků

...

Data structure - who knows exactly

- several levels (constant) of indirect access
- logarithmic blocks with superblocks and precalculated table

Wavelet Trees

- rozšíření rank & select na větší abecedu
- více vektorů v jednom stromu výšky $O(\lg|\Sigma|)$

- rank shora dolů - vždy používám výsledek jako vstup další query
- select směrem nahoru

→ compressed suffix arrays

- každý level polovina prvků předchozího
- sudé hodnoty SA_k budou půlení i v levelu $k+1$
  - liché mají neighborhood funkci Φ kdovou, že $SA_k[\phi_k(i)] - 1$
    (tj $\Phi(i)$ obsahuje index prvku jehož $SA = SA[i] + 1$)

Burrows - Wheeler & FM - Index

T = abracadabra P = bdabra

i	1	2	3	4	5	6	7	8	9	10	11
T[i]	a	b	r	a	c	a	d	a	b	r	a
SA	11	8	1	4	6	9	2	5	7	10	3
F	a	a	a	a	a	b	b	c	d	r	r
SA'	10	7	11	3	5	8	1	4	6	9	2
L	r	d	a	r	c	a	a	a	a	b	b

$\sim T[SA[i]]$ = první symboly lexikograficky seřazených rotací

$\sim SA[i] - 1$ - poslední symboly lex. seřazených rotací

$\sim T[SA'[i]] = BWT(T)$

C	a	b	c	d	r
	0	5	7	8	9

- kumulativní frekvence symbolů (první výskyt daného symbolu) v F

FM-Index:

$Occ(X, c, i)$ - # počet výskytů c v prefixu $X[1 \dots i]$

$Sp_i = C[c] + Occ(L, c, Sp_{i-1} - 1) + 1$

$Ep_i = C[c] + Occ(L, c, Ep_{i-1})$

$occ_i = Ep_i - Sp_i + 1$ - počet výskytů symbolů (daného řádku)

interval $[Sp_i, Ep_i]$ říká indexy v SA → pozice ve stringu

příklad: $Sp_0 = 1$

$Ep_0 = 11$

$Occ_0 = 11 - 1 + 1 = 11$ výskytů prázdného řádku

- pattern jde odzadu! → c = a

$Sp_1 = 0 + 0 + 1 = 1$

$Ep_1 = 0 + 5 = 5$

} interval $[1, 5]$

c = r

$Sp_2 = 9 + 0 + 1 = 10$

$Ep_2 = 9 + 2 = 11$

} interval $[10, 11]$