

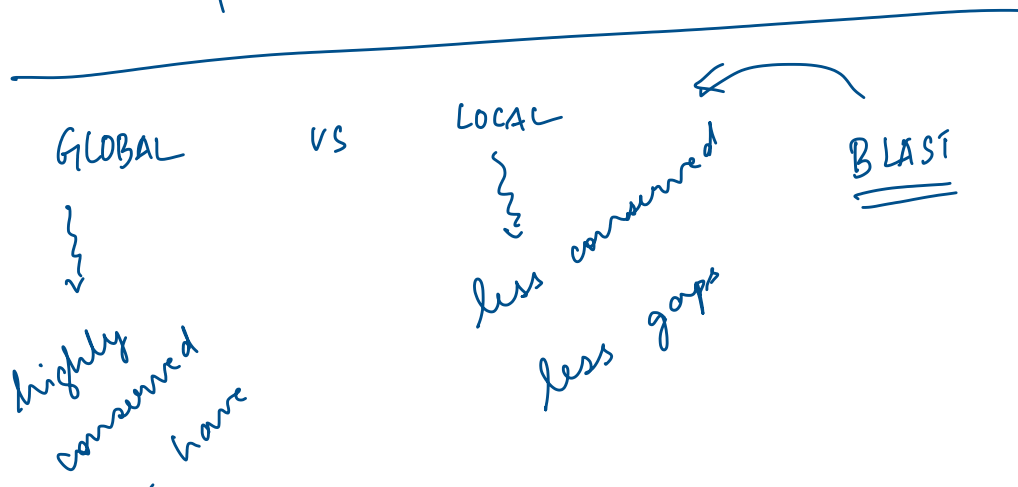
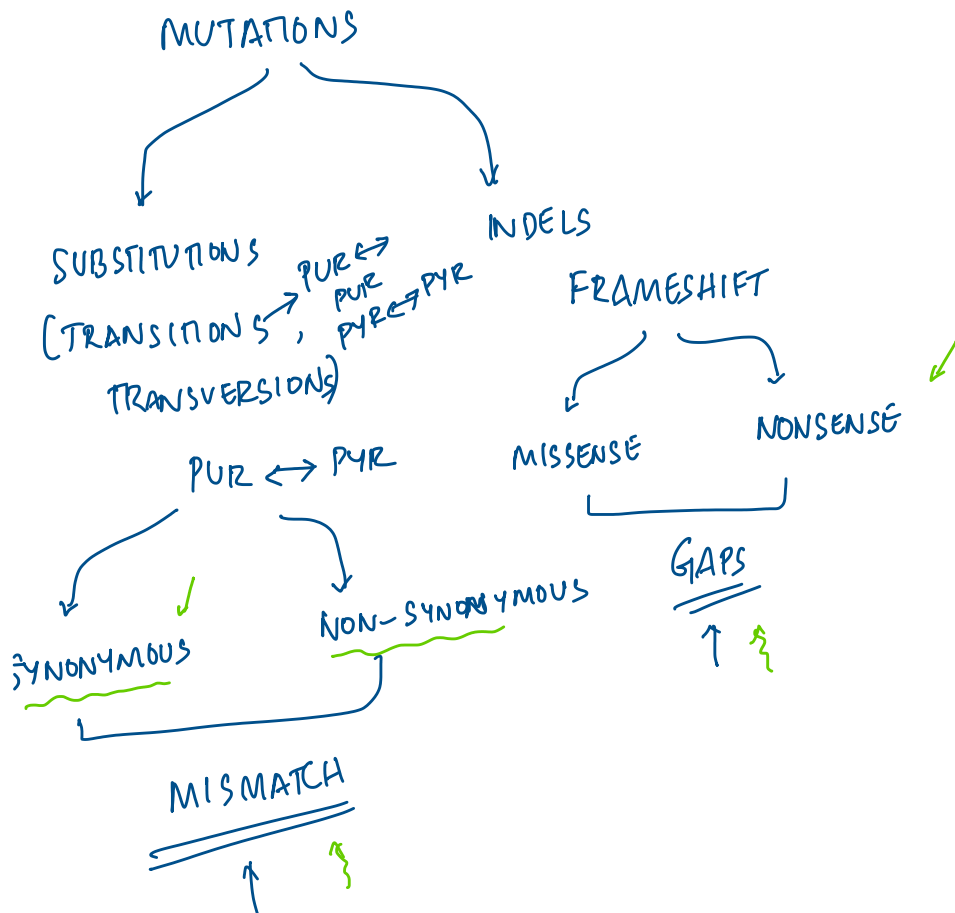
Week 3 Guided Assembly

Friday, 4 February 2022 8:08 AM

BOWTIE, BWA-MEM

⇒ efficiently align reads to a reference

ALIGNMENT → Homology → MATCH



might
(gap)

NEEDLEMAN & WUNSCH

seq. 1 - m

seq 2 - n

STEP 1: create a matrix $(m+1) \times (n+1)$ ↙

STEP 2: fill up gap penalties

STEP 3: use a scoring scheme:

$$S_{i,j} = \max \begin{cases} F(i-1, j-1) + s_{i,j} \\ F(i-1, j) - \text{gap} \\ F(i, j-1) - \text{gap} \end{cases} \quad \begin{matrix} \text{(match)} \\ \text{DIAGONAL} \end{matrix}$$

eg. MATCH = +1, MISMATCH = -1, GAP = -2

seq 1 = AGC (m)

seq 2 = AAAC (n)

		A	G	C
1	0	-2	-4	-6
A	-2	-4	-1	-3
A	-4	-1	0	-2
A	-6	-3	-2	-1

A	A	A	C
1	X		1
A	G	—	C

C | -8 | -5 | -4 | -1

SLW \rightarrow if you a -ve score, you push it to 0

PROBLEM: align 'billions' of reads to a reference genome

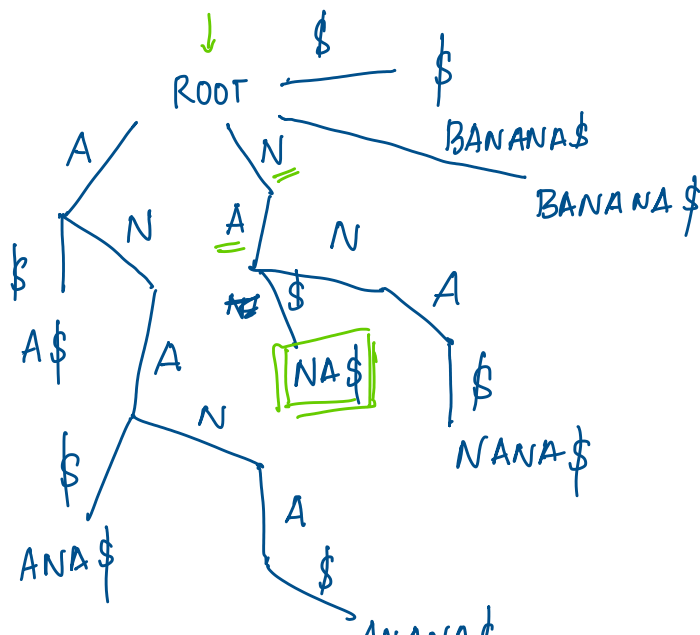
ERGO

we are going to build a dictionary or database

SUFFIX ARRAY / TREE

GENOME: BANANA\$

all possible substrings \rightarrow
BANANA\$, ANANA\$, NANA\$, ANA\$, NA\$, A\$, \$]



ANANA\$

NA\$
n = 1

BURROW - WHEELER TRANSFORM

- ① BANANA\$
- ② A\$BANAN
- ③ NA\$BANA
- ④ ANA\$BAN
- ⑤ NANA\$BA
- ⑥ ANANA\$B

STEP 1

STEP 2: arrange the words alphabetically

- ① \$BANANA
- ② A\$BANAN
- ④ ANA\$BAN
- ⑥ ANANA\$B
- ⑧ BANANA\$
- ③ NA\$BANA
- ⑤ NANA\$BA

SUFFIX
ARRAY

BWT

SA = I column
= \$AABNN

BWT = LAST COLUMN
= ANNBSAA

STEP 3

POSITION (INDEX)	1	2	4	6	0	3	5
BWT	A	N	N	B	\$	A	A
SA	\$	A	A	A	B	N	N

eg. NA (read)

ALGORITHM 1 - BISECTION

⇒ 2 matches → Position ③ & ⑤

BA~~NA~~~~NA~~\$