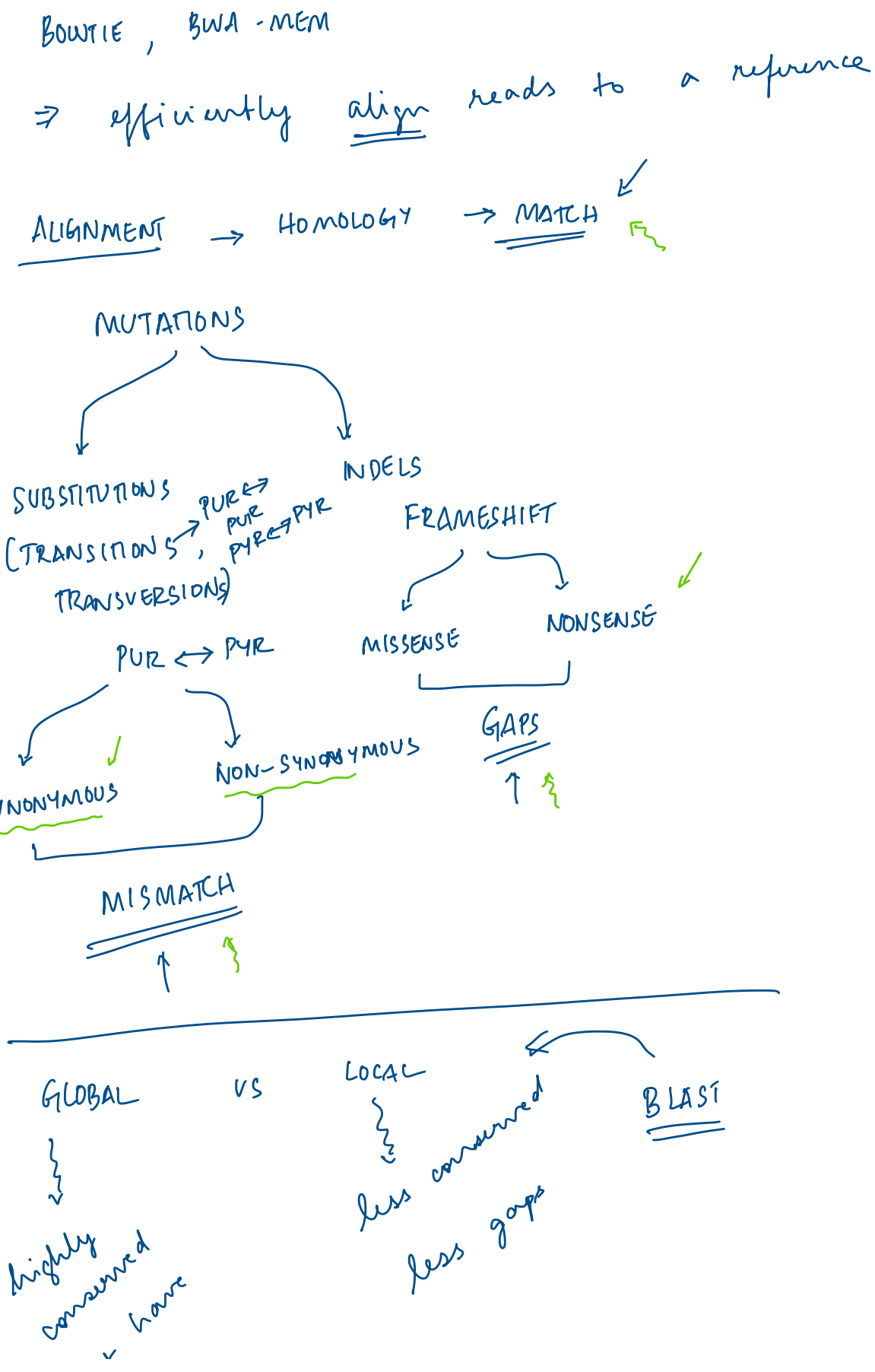
Week 3,4 Guided Assembly

Friday, 4 February 2022 8:08 AM



NEEDLEMAN & WUNSCH

(with gaps)

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:

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

DIAGONAL

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

seq 1 = AGC (m)

seq 2 = AAAC (n)

		A	G	C
	0	−2	−4	−6
A	−2	−1	−1	−3
A	−4	−3	0	−2
A	−6	−5	−2	−1

A	A	A	C
|	X		|
A	G	—	C

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

S & W → if you a -ve score, you push it to 0

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

ERGO

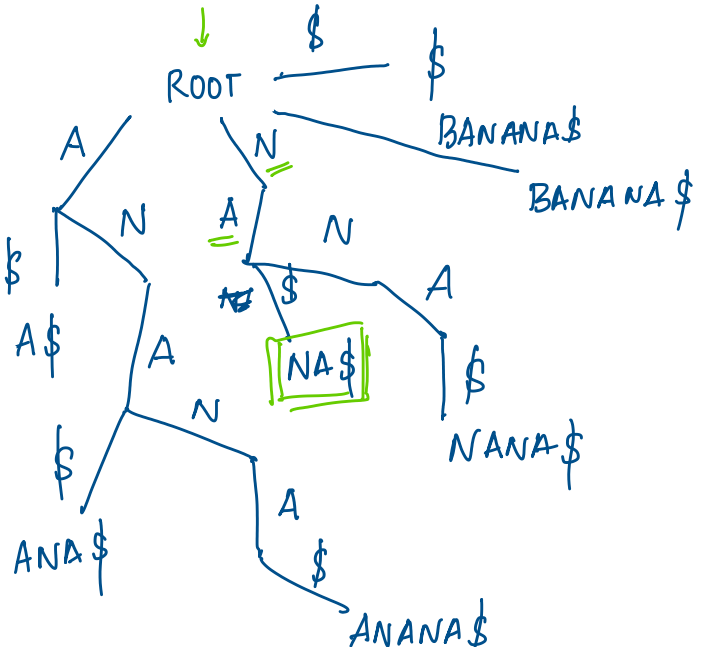
we are going to build a dictionary or database

SUFFIX ARRAY / TREE

GENOME : BANANA $

all possible substrings ↓

BANANA $, ANANA $, NANA $, ANA $, NA $, A $, $



NA$

BURROW - WHEELER TRANSFORM

0 BANANA$

1 $BANANA

2 A$BANAN

3 NA$BANA

4 ANA$BAN

5 NANA$BA

6 ANANA$B

STEP1

STEP2 : arrange the words alphabetically

1 $BANANA

2 A$BANAN

4 ANA$BAN

6 ANANA$B

STEP1

0 BANANA$

3 NA$BANA

5 NANA$BA

STEP2

SUFFIX ARRAY

BWT

=$AAABNN

SA = I COLUMN

BWT = LAST COLUMN

= ANNB$AA

STEP3

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

e.g. NA (read)

ALGORITHM1 - BISECTION

⇒ 2 matches → Position (3) & (5)

BA(N)A(NA)$

Backward First Algorithm

(1) Look for A in SA [2, 4, 6]

(2) Look for all N prefixes in BWT [2, 4]

(3) Whichth occurrence of N's in BWT? [1st, 2nd]

(4) Look for those occurrences of N in SA → positions 3, (5) (backwards) STOP

REF GENOME : ^{0}ATGC^{4}ATGC$

READ : ATG

(0)	ATGCATGC$	(1)	$ATGCATGC
(1)	$ATGCATGC	(5)	ATGC$ATGC
(2)	C$ATGCATG	(0)	ATGCATGC$
(3)	GC$ATGCAT	(2)	C$ATGCATG
(4)	TGC$ATGCA	(6)	CATGC$ATG
(5)	ATGC$ATGC	(3)	GC$ATGCAT
(6)	CATGC$ATG	(7)	GCATGC$AT
(7)	GCATGC$AT	(4)	TGC$ATGCA
(8)	TGCATGC$A	(8)	TGCATGC$A

A ↓ Z

POSITION	1	(5)	(0)	2	6	3	7	4	8
BWT	C	C	$	G	G	T_{1st}	T_{2nd}	A	A
SA	$	A	A	C	C	G	G	T	T

ATG

⇒ INDEX 0 and 5 backwards

i.e. 5th and 9th positions backwards

STEP 1 : Look for G in SA → 3, 7 ;

STEP 2 : which of those are prefixed by T? → both ;
which th T occurrences? 1st, 2nd

STEP 3 : 1st and 2nd occurrences of T in SA → 4, 8

STEP 4 : which of those are prefixed by A? → both
... th A occurrences? 1st, 2nd

STEP 5 : which A occurs

STEP 6 : 1st, 2nd A's in SA → 5, 0 backwards (STOP)

---

2 types of coins in bag → BIASED OR UNBIASED

$P(\text{BIASED}) = P(\text{UNBIASED}) = \frac{1}{2}$

OUTCOMES

BIASED

$P(H \mid \text{BIASED}) = 1$

$P(T \mid \text{BIASED}) = 0$

UNBIASED

$P(H \mid \text{UNBIASED}) = \frac{1}{2}$

$=$

$P(T \mid \text{UNBIASED}) = \frac{1}{2}$

$P(\text{BIASED} \mid H) = ?$

BAYES' THEOREM

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

$P(A \mid B)$: POSTERIOR PROB.

$P(B \mid A)$: LIKELIHOOD

$P(A)$, $P(B)$: PRIOR PROBS.