Multiple Sequence alignment (MSA)

Suppose (3) sequences are :-

S_1 = ATTCGAT
S_2 : TTGAG
S_3 = ATGCT

S_1 = A T T C G A T
S_2 = – T T – G A G
S_3 = A T – – G C T

غالبا مزودش عدد الـ Seq. → mismatch وليس Gap

The Star Algorithm for MSA

Approximation to Sp

↓ depend on distance every pair sequence

let $\sigma(x,y) = 0$ if $x = y$ match
$\sigma(x,y) = 1$ if $x \neq y$ mismatch or Gap

S_1 = G C C A T
S_2 = G – – A T
0 1 1 0 0

distance = 2

S_1 = G C C A T
S_2 = G A – – T
0 1 1 1 0

distance = 3

→ كل ما قلت يبقى احسن

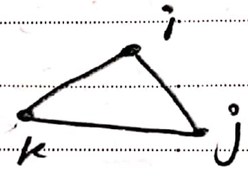
distance included by alignment is defined (as) ⇒ $d(S_i, S_j) = \sum_{t=1}^{\hat{n}} \delta(a_t, b_t)$

match ← أزود
mismatch ← اقلل
? Gap

Properties

① $d(S_i, S_i) = 0$ → بين حاجة ونفسها

② Tringular inequality



$$d(S_i, S_j) + d(S_i, S_k) \geq d(S_j, S_k)$$

Example

S_1 = ATGCT C
S_2 = AGAGC
S_3 = TTCTGT
S_4 ATTGCATGC

① Find similarity
② Miss match : اكمل الباقي ب
او Gap وحاول اقلل
③ احسب الـ distance

① S_1 = A T G C T C
S_2 = A - G A G C
0 1 0 1 1 0
$D(S_1, S_2) = 3$

② S_1 = A T G C T C
S_3 = - T T C T G
1 0 1 0 0 1
$D(S_1, S_3) = 3$

بتاعي

بتاعة الدكتور
S_1 = A T G C T C
S_3 = T T - C T G
1 0 1 0 0 1
$D(S_1, S_3) = 3$

$S_4 =$ A T T G C A T G C

③ $S_1 =$ A T - G C - T - C

0 0 1 0 0 1 0 1 0

$D(S_1, S_4) = 3$

A G A G C

④ $S_2 =$ A G A G C

$S_3=$ T T C T G

1 1 1 1 1

$D(S_2, S_3) = 5$

⑤ $S_4 =$ A T T G C A T G C

$S_2 =$ A - - - G - A - G C

$D(S_2, S_4) = 4$

⑥ $S_4 =$ A T T G C A T G C

$S_3 =$ - T T - C - T G -

$D(S_3, S_4) = 4$

$D(S_1, S_2) + D(S_1, S_3) + D(S_1, S_4) = 9$

$D(S_2, S_1) + D(S_2, S_3) + D(S_2, S_4) = 12$

$D(S_3, S_1) + D(S_3, S_2) + D(S_3, S_4) = 12$

$D(S_4, S_1) + D(S_4, S_2) + D(S_4, S_3) = 1$

S_1 is selected as the Center

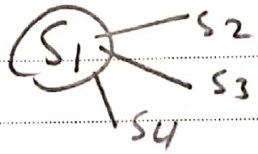
S_1 is the most similar to other

↓

distance أقل

Scanned with CamScanner

Alignment



$S_2 = AGAGC$

$S_3 = TTCTG$

$S_4 = ATTGCATGC$

$S_1 + S_2$

$S_1 = ATGCTC$

$S_2 = A-GAGC$

$S_1 + S_2 + S_3$

$S_1 = ATGCTC$

$S_2 = A-GAGC$

$S_3 = -TTCTG$

$S_1 + S_2 + S_3 + S_4$

$S_1 = AT-GC-T-C$

$S_2 = A--GA-G-C$

$S_3 = -T-TC-T-G$

$S_4 = ATTGCATGC$