Chapter 3(c)

Multiple Alig Sequence Alignment:

A multiple sequence alignment is basically an alignment of more than 2 sequence.

Pairwise Alignment:

A pairwise alignment is basically an alignment between two sequence.

Note: We can derive pairwise alignment from a multiple sequence alignment. For example:-

S1 = AC-GCGG-C

S2 = AC-GC-GAG

S3 = GCCGC-GAG

Induces:

#S1: AC-GCGG-C
S2: AC-GC-GAG

#S2: AC-GC-GAG
S3: GCCGC-GAG

#S1: AC-GCGG-C
S3: GCCGC-GAG

Why we can not always construct a multiple alignment from pairwise alignments?

Ans:

A multiple sequence alignment is an alignment of more than two sequences, whereas a pairwise alignment involves only two sequences. We can derive pairwise alignments from a multiple alignment. For example, given the multiple alignment:

```
X: A C - G C G G - C
Y: A C - G C - G A G
Z: G C C G C - G A G
```

We can extract the following pairwise alignments:

```
X: A C G C G G - C | Y: A C - G C G A G
Y: A C G C - G A G | Z: G C C G C G A G
```

```
X: A C - G C G G - C
Z: G C C G C - G A G
```

However, we cannot always construct a multiple alignment because pairwise alignments may be ~~incosi~~ inconsistent, with each other. From an optimal multiple alignment, we can infer pairwise alignments ~~are~~ not between all pairs of sequences. But these pairwise alignments are not

~~nees~~ necessarily optimal. When aligning two sequences at a time, the algorithm optimizes their alignment independently. However, when aligning all ~~the~~ three sequences together, gaps placed optimally for one pair may not work for another pair. Additionally, even if sequence1 aligns ~~wi~~ well with sequence2 and sequence2 aligns well with sequence3, it does not always mean sequence1 align well with sequence3 in the same way. This inconsistency makes it impossible to always reconstruct a valid multiple alignment from pairwise alignments.

Profile: ~~Pr~~

Profile is usually a probability for each letter to occure in each column.

Multiple Alignment (Greedy Approach):

Steps:

1. At first calculate all possible pairwise alignment of the given multiple sequence.
2. Then we will find two closest sequence among all of them (the pairwise alignment which score is the greatest among all of them.
3. Then we will join these two sequence into one profile.
4. Then will add the new sequence with other sequences.

5. Since, we align a sequence with other sequence profile with other profile, a sequence with other profile, we at the end we will get two one single profile.

Example:

S1: G A T T C A
S2: G T C T G A
S3: G A T A T T
S4: G T C A G C

Match = +1
Indel / Mismatch = -1

Here we will find $^4C_2 = 6$ possible alignment.

1. Before alignment:

S1: G A T T C A
S2: G T C T G A

After alignment:

S1: G A T — T C A
S2: G — T C T G A

Score = 1 - 1 + 1 - 1 + 1 - 1 + 1

= 1

2. Before alignment:

S2: G T C T G A
S3: G A T A T T

2. Before alignment:

S1: G A T T C A

S3: G A T A T T

After alignment:

~~S1: G A T – – C A~~

~~S2: G A T~~

S1: G A T – T C A

S3: G A T A T – T

Score = 1 + 1 + 1 – 1 + 1 – 1 + 1

= 1

3. Before Alignment

S1: G A T T C A

S4: G T C A G C

After Alignment:

S1: G A T T C A – –

S4: G – T – C A G C

Score = 1 – 1 + 1 – 1 + 1 + 1 – 1 – 1

= 0

4. Before Alignment:

S2: G T C T G A
S3: G A T A T T

After Alignment:

S2: G — T C T G A
S3: G A T — A T T

Score = 1 − 1 + 1 − 1 + 1 − 1 − 1
= −1

5. Before alignment:

S2: G T C T G A
S4: G T C A G C

After alignment:

S2: G T C T G A
S4: G T C A G C

Score = 1 + 1 + 1 − 1 + 1 − 1
= 2

6. Before alignment:

S3: G A T A T T
S4: G T C A G C

After alignment:

S3: G A T _ A T T
S4: G — T C A G C

~~Score = 1 - 1 + 1 + 1 - 1 - 1~~

~~= 0~~

Score = 1 - 1 + 1 - 1 + 1 - 1 - 1

= -1.

Now, we can see that the highest score is 2. So, S_2 and S_4 are closest. So now combine them to get a profile.

S2: G T C T G A

S4: G T C A G C

$S_{2,4}$ = G T C t/a G a/c

Or,

G T C t/a G a/c

So, now set of sequence:

S1: G A T T C A

S3: G A T A T T

$S_{2,4}$: G T C t/a G a/c

Continue the above process until we find a single profile.

Multiple alignments:

(i) Entropy (ii) Sum of pairs (SP-Score).

(i) Entropy:

$-\sum P_x \log P_x$; Where, $X = A, T, G, C$; P_x = The probability of each character

Example:

A	A	A
A	A	A
A	A	T
A	T	C

Here,

1st column: $P_A = 1$, $P_T = 0$, $P_C = 0$, $P_G = 0$

2nd column: $P_A = 0.75$, $P_T = ?$, $P_C = 0$, $P_G = 0$

3rd column: $P_A = 0.50$, $P_T = 0.25$, $P_C = 0.25$, $P_G = 0$

Now, For 1st column,

Entropy, $n_1 = -\sum P_x \log P_x$

$= -\{P_A \log P_A + P_T \log P_T + P_C \log P_C + P_G \log P_G\}$

$= -\{1 \log 1 + 0 \log 0 + 0 \log 0 + 0 \log 0\}$

$= 0$

For 2nd column,

Entropy, $n_2 = -\sum P_x \log P_x$

$= -\{P_A \log P_A + P_T \log P_T + P_C \log P_C + P_G \log P_G\}$

$= -\{0.75 \log(0.75) + 0.25 \log(0.25) + 0 \log 0 + 0 \log 0\}$

$= -[\{0.75 \times (-0.12)\} + \{0.25 \times (-0.60)\}]$

$= 0.06$

For 3rd column,

Entropy, $n_3 = -\sum P_x \log P_x$

$= -\{P_A \log P_A + P_T \log P_T + P_C \log P_C + P_G \log P_G\}$

$= -\{0.50 \times \log(0.50) + 0.25 \times \log(0.25) + 0.25 \times \log(0.25) + 0 \log 0\}$

$= -[\{0.50 \times (-0.30)\} + \{0.25 \times (-0.60)\} + \{0.25 \times (-0.60)\}]$

$= -[-0.15 - 0.15 - 0.15]$

$= 0.45.$

So, Alignment entropy $= 0 + 0.606 + 0.45 = 0.51.$

Sum of Pairs (SP-Score):

Sum up the pairwise alignment's score for a multiple alignments.

$$S(a_1 \ldots a_4) = \sum S^*(a_i, a_j) = S^*(a_1, a_2) + S^*(a_1, a_3) + S^*(a_1, a_4) + S^*(a_2, a_3) + S^*(a_2, a_4) + S^*(a_3, a_4).$$

For example:

S1: G A T T C A

S2: G T C T G A

S3: G A T A T T

S4: G T C A G C

$S_{1,2} \rightarrow$ Score = 1

$S_{1,3} \rightarrow$ Score = 1

$S_{1,4} \rightarrow$ Score = 0

$S_{2,3} \rightarrow$ Score = −1

$S_{2,4} \rightarrow$ Score = 2

$S_{3,4} \rightarrow$ Score = −1

S.P pair = 1 + 1 + 0 − 1 + 2 − 1

= 2

Note: SP-Score বেশি হলে ভালো alignment

entropy কম হলে ভালো alignment