Exercise sheet 3: T-Coffee

Exercise 1

You are given the sequences a, b and c

$$a = CACCGGb = ACCAAGc = AACACC$$

The pairwise optimal alignments $A(x, y)$ of the set of sequences S were calculated as:

```
a: CACCG_G          a: __CACCGG     b: ACCAAG
    |||: |              ||||           |:||::
b: _ACCAAG          c: AACACC__     c: AACACC
```

Question 1A Calculate the primary library (L)

Formulae init: $L_{i,j}^{x,y} = 0$

$\forall$ alignments A of sequences x and y of the set S.

$weight(A) = \dfrac{number\ of\ matches}{min(len(x), len(y))} * 100$

$\forall$ aligned positions i, j with $1 \leq i \leq len(x)$ and $1 \leq j \leq len(y)$

$L_{i,j}^{x,y} = L_{i,j}^{x,y} + weight(A)$

Solution $L_{2,1}^{a,b} = L_{3,2}^{a,b} = L_{4,3}^{a,b} = L_{6,6}^{a,b} = 100 * \frac{4}{6} = 67$ and all other $L_{i,j}^{a,b} = 0$

$L_{1,3}^{a,c} = L_{2,4}^{a,c} = L_{3,5}^{a,c} = L_{4,6}^{a,c} = 100 * \frac{4}{6} = 67$ and all other $L_{i,j}^{a,c} = 0$

$L_{1,1}^{b,c} = L_{3,3}^{b,c} = L_{4,4}^{b,c} = 100 * \frac{3}{6} = 50$ and all other $L_{i,j}^{b,c} = 0$

Question 1B Calculate the extended library (EL)

Formulae $EL_{i,j}^{x,y} = L_{i,j}^{x,y} + \sum_{z \in S \setminus \{x,y\}} \sum_{1 \leq k \leq len(z)} min(L_{i,k}^{x,z}, L_{k,j}^{z,y})$

Solution The original Library doesn't change as there are no edges enforcing certain connections. Hence

$EL_{i,j}^{x,y} = L_{i,j}^{x,y} \qquad \forall L_{i,j}^{x,y} \neq 0$

and the following weights are added:

```
a: CACCG_G
    |||: |
b: _ACCAAG
    |:||::
c:  AACACC
    * *
```

$EL^{a,b}_{1,3} = EL^{a,b}_{2,4} = 50$

```
a: __CACCGG
     ||||
c: AACACC__
   |:||::
b: ACCAAG
     **
```

$EL^{a,c}_{2,1} = EL^{a,c}_{4,3} = 50$

```
b:    ACCAAG
      |||: |
a:   CACCG_G
     ||||
c: AACACC
      ***
```

$EL^{b,c}_{1,4} = EL^{b,c}_{2,5} = EL^{b,c}_{3,6} = 67$

Question 1C Realign the sequences b and c using EL for scoring and gap costs and mismatch costs of 0

Formulae

$$
\begin{aligned}
i &\in [1, |x|] \\
j &\in [1, |y|] \\
L(0,0) &= 0 \\
L(i,0) &= w(x_i, -) * i \quad \text{or } = L(i-1,0) + w(x_i, -) \\
L(0,j) &= w(-, y_j) * j \quad \text{or } = L(0, j-1) + w(-, y_j) \\
L(i,j) &= max \begin{cases} L(i-1,j) + w(x_i, -) \\ L(i, j-1) + w(-, y_j) \\ L(i-1, j-1) + w(x_i, y_j) \end{cases} \\
w(x_i, y_j) &= \begin{cases} EL^{x,y}_{i,j} & \text{match-score if } (x_i = y_j) \\ 0 & \text{insert/deletion-score if } (x_i = - \vee y_j = -) \\ 0 & \text{mismatch-score else } (x_i \neq y_j) \end{cases}
\end{aligned}
$$

Solution

-1	-	A1	C2	C3	A4	A5	G6
-	0	0	0	0	0	0	0
A	0	50	50	50	50	50	50

-1	-	A1	C2	C3	A4	A5	G6
A	0	50	50	50	50	50	50
C	0	50	50	100	100	100	100
A	0	67	67	100	150	150	150
C	0	67	133	133	150	150	150
C	0	67	133	200	200	200	0

Question 1D Do the other alignments a-b and a-c change? Provide arguments, without calculating new alignments.

Solution No. The newly added alignment scores in EL represent edges that are incompatible with the current best alignments and can not score higher.

Question 1E Sketch a Guide Tree (either Sketch or Newick format)

Solution Newick: $((a, c), b)$ or $((a, b), c)$

Question 1F Perform a progressive alignment by aligning sequence b to the already existing alignment $A(a, c)$. To score a match between b and $A(a, c)$ use the sum $EL^{a,b} + EL^{b,c}$ with the correct indices. Show the resulting multiple sequence alignment.

Solution

–	–	-A	-A	CC	AA	CC	CC	G-	G-
-	0	0	0	0	0	0	0	0	0
A	0	50	50	50	133	133	133	133	133
C	0	50	50	50	133	267	267	267	267
C	0	50	50	150	150	267	400	400	400
A	0	50	50	150	250	267	400	400	400
A	0	50	50	150	250	267	400	400	400
G	0	50	50	150	250	267	400	400	467