COMS4761: FALL 2007

ASSIGNMENT 2.2 SOLUTIONS

Problem 1: Computing MSA

Consider the ezRNA sequences {X1, X2, X3, X4, X5}.
`X1=00101 X2=01010 X3=01100 X4=10110 X5=10101`

A. (25 points) Compute MSA of these sequences using progressive alignment as follows:
(i) Compute pairwise alignments using the scoring matrix of problem 1.

Apply Needleman-Wunsch to all possible pairs. The dynamic-programming matrices are shown below. Elements along an optimal paths are marked in red boldface. The optimal paths can be easily inferred and are thus omitted, except for the first two alignments (X1,X2) and (X1,X3):

Aligning X1 with {X2, X3, X4, X5}

		0	0	1	0	1
	0	**-2**				
0	-2	**2**	**0**			
1		**0**	1	**2**		
0			**2**	0	**4**	
1				**4**	2	**6**
0					**6**	**4**

		0	0	1	0	1
	0	-2				
1	-2	**-1**	0			
0		**0**	**1**			
1			**-1**	**3**	**1**	
1				**1**	**2**	**3**
0					**3**	**1**

		0	0	1	0	1
	0	-2				
0	-2	**2**	0			
1		0	**1**	2		
1				**3**	1	
0				1	**5**	3
0					3	**4**

		0	0	1	0	1
	0	-2				
1	-2	**-1**	-3			
0		0	**1**	-1		
1			-1	**3**	1	
0				1	**5**	3
1					3	**7**

`Aligning X2 with {X3,X4,X5}`

		0	1	0	1	0
	0	-2				
0	-2	**2**	**0**			
1		**0**	**4**	**2**		
1			**2**	**3**	**4**	
0				**4**	**2**	**6**
0					3	**4**

		0	1	0	1	0
	0	**-2**				
1	**-2**	-1	**0**			
0		**0**	-2	**2**		
1			**2**	0	**4**	
0				**4**	2	**6**
1					**6**	**4**

		0	1	0	1	0
	0	**-2**				
1	**-2**	**-1**	**0**			
0		**0**	**-2**	**2**		
1			**2**	**0**	**4**	
1				1	**2**	3
0					0	**4**

Aligning X3 with {X4, X5} and X4 with X5:

		0	1	1	0	0
	0	-2				
1	**-2**	-1	0			
0		**0**	-2	-1		
1			**2**	0	-2	
1				**4**	2	0
0					**6**	**4**

		0	1	1	0	0
	0	**-2**				
1	**-2**	**-1**	**0**			
0		**0**	**-2**	-1		
1			2	**0**	-2	
0				1	**2**	0
1					0	**1**

		1	0	1	1	0
	0	**-2**				
1	**-2**	**2**	**0**			
0		**0**	**4**	**2**		
1			**2**	**6**	**4**	
0				**4**	**5**	**6**
1					**6**	**4**

These dynamic programming matrices admit multiple optimal paths corresponding to multiple pairwise alignments. For example, the following three alignments of (X1,X2) are admissible:

```
X1    00101_    0_0101    00101_
X2    _01010    01010_    0_1010
```

The pairwise alignments are summarized in the matrix below

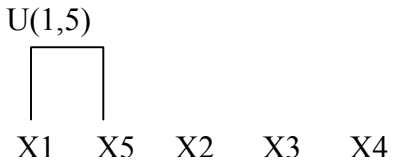
	X1	X2	X3	X4	X5
X1		0_0101 01010_	00101 01100	00101_ 101_10	00101 10101
X2			01010_ 01_100	0101_0 _10110	01010_ _10101
X3				_01100 10110_	01100 10101
X4					101_10 10101_
X5					

(ii) Use the Hamming distance H(X,Y) to compute a UPGMA Guide Tree.

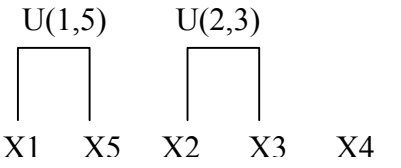
The standard MSA algorithm would use a scoring based distance metric. The Hamming distance was selected to merely simplify the computations.

The left matrix provides the Hamming distances. We apply UPGMA clustering to the nearest pair, marked in yellow. The partial UPGMA trees are depicted on the left.

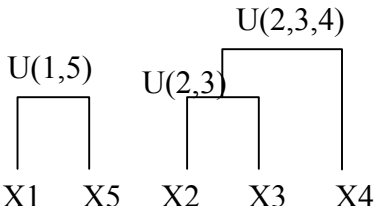
	X1	X2	X3	X4
X1				
X2	2			
X3	2	2		
X4	3	2	2	
X5	1	2	3	2



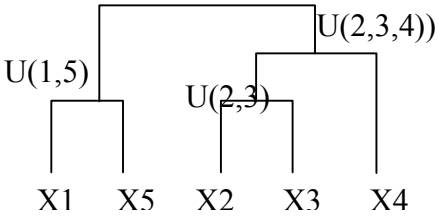
	U(1,5)	X2	X3
U(1,5)			
X2	2		
X3	2.5	2	
X4	2.5	2	2



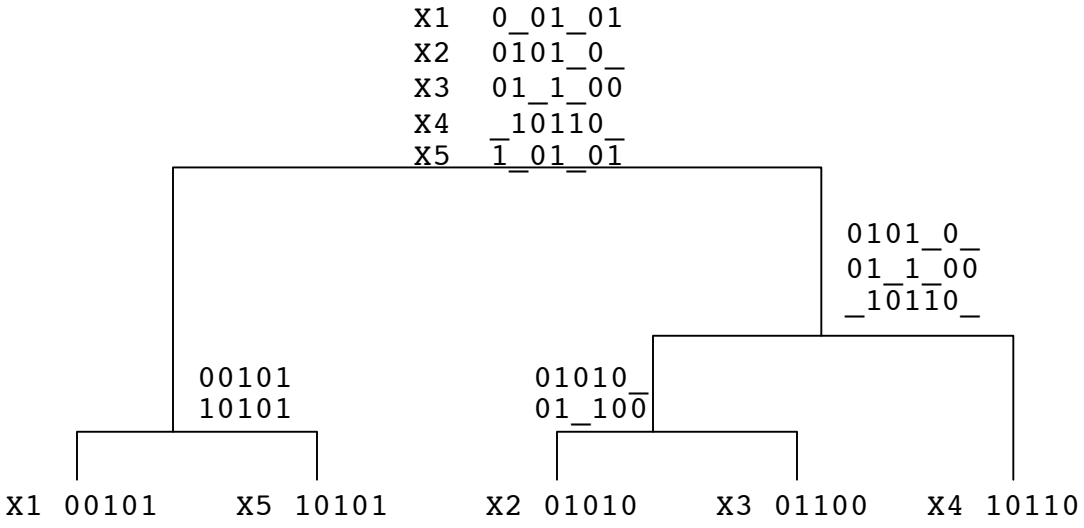
Notice that there are 4 pairs that could be clustered next.
We select X2,X3 for clustering.



	U(1,5)	U(2,3)
U(1,5)		
U(2,3)	2.25	
X4	2.5	2



(iii) Use the Guide Tree to merge the pairwise alignments into an MSA.



B. (15 points) Use the Barton-Sternberg iterative alignment algorithm to improve the MSA of problem A

```
X1  0-01-01
X2  0101-0-
X3  01-1-00
X4  -10110-
X5  1-01-01
P1  0101-0.    (Red means uniform alignment column)
```

Step 1: remove X3 and profile the rest

```
X1  0-01-01
X2  0101-0-
X4  -10110-
X5  1-01-01

P2  0.01-0.
X3  --01100  Align X3,P2

Reinsert X3

X1  0-01-01
X2  0101-0-
X3  --01100
X4  -10110-
X5  1-01-01
P3  .-01-0.

Step 2: remove X1 and profile the rest

X2  0101-0-
X3  --01100
X4  -10110-
X5  1-01-01
P4  ..01-0.
X1  0-01-01  Align X1,P3 (reproduces the alignment of X1 above)

Step 3: reinsert X1 and remove X2

X1  0-01-01
X3  --01100
X4  -10110-
X5  1-01-01
P5  --01-0.
X2  0101-0- Align X2,P5 (reproduces the alignment of X2 above)
```

```
[note that other alignments of X2 may be valid]
Step 4: reinsert X2 and remove X4

X1  0-01-01
X2  0101-0-
X3  --01100
X5  1-01-01
P6  0-01-01
X4  1-0110-  Align X4, P6

Step 5: reinsert X4 and remove X5


X1  0-01-01
X2  0101-0-
X3  --01100
X4  1-0110-
P7  0-01.0-
X5  1-01-01     Align X5, P7 reproduces X5

Step 6: reinsert X5


X1  0-01-01
X2  0101-0-
X3  --01100
X4  1-0110-
X5  1-01-01
P8  .-01-0.
```

Final MSA:

```
X1  0-01-01
X2  0101-0-
X3  --01100
X4  1-0110-
X5  1-01-01
```

Notes:
The iterative refinement modified the initial MSA to best align each sequence with the common profile determined by the rest of the sequences. The initial MSA failed to recognize the possibility that column 3 can be identical for all sequences and should thus be used to guide the MSA. The iterative technique identified this possibility quickly and then used the resulting profile to guide the MSA process. Notice that even if the process does not necessarily stabilize the alignments, it may likely stabilize the parts of the profile that are conserved by the sequences. However, it is possible to have examples where the final profile will depend on the initial MSA and/or on the order of sampling